Multinomiale Modelle in der kognitiven Psychologie

Habilitationsschrift
zur Erlangung der Lehrbefugnis im Fach Psychologie
an der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität in Bonn

vorgelegt von

Edgar Erdfelder
aus Göttingen-Geismar

Bonn, Januar 2000

Vorwort

Gegenstand der vorliegenden Arbeit ist eine Klasse von stochastischen Modellen, die sich vor allem in den letzten zehn Jahren in ganz unterschiedlichen Anwendungskontexten immer wieder als sehr nützlich erwiesen hat. Dies gilt prinzipiell für das Gesamtgebiet der empirischen Psychologie, vor allem aber für die kognitive Psychologie. Multinomiale Modelle sind nach meiner Auffassung besser als andere Ansätze geeignet, eine Verbindung zwischen psychologischen Theorien und psychologischen Daten herzustellen, die einerseits den Kriterien der formalen Präzision und der empirischen Prüfbarkeit genügt, andererseits aber auch einfach, flexibel und robust genug ist, um auf viele unterschiedliche Problemstellungen anwendbar zu sein, auch solche, bei denen auf vorhandene Ansätze formaler Theoriebildung bislang noch nicht zurückgegriffen werden kann. Wie zu zeigen sein wird, erscheint es auf dem Hintergrund vorliegender Befunde zur multinomialen Modellierung immerhin denkbar, daß das „Joch der precision-importance trade-off function" (Schwarz, 1991) in der Psychologie prinzipiell überwindbar ist. Wissenschaftliche Genauigkeit und Eindeutigkeit muß nicht immer mit einem Verlust an praktischer Bedeutsamkeit erkauft werden und umgekehrt muß die Optimierung des Anwendungsaspekts und die Erweiterung des Anwendungsspektrums nicht zwangsläufig einen Verzicht auf Präzision nach sich ziehen.

Die folgenden Kapitel behandeln in erster Linie die Anwendung spezieller multinomialer Modelle in der kognitiven Psychologie, vor allem in der Gedächtnispsychologie. Mathematisch-statistische Aspekte derartiger Modelle stehen also nicht im Vordergrund, müssen aber selbstverständlich gebührend Beachtung finden, weil die korrekte, für substanzwissenschaftliche Zwecke gewinnbringende Anwendung multinomialer Modelle die Kenntnis der statistischen Modellannahmen, der Analysemethoden und der Anwendungsgrenzen voraussetzt. Die Darstellung und Diskussion statistischer Probleme und Lösungsmöglichkeiten erfolgt dabei aber nicht aus der Perspektive der Mathematischen Statistik, sondern grundsätzlich aus der Sicht empirisch arbeitender Psychologinnen und Psychologen, die Interesse an formaler Modellbildung und theoriebezogener Datenanalyse haben. Der Text ist so verfaßt worden, daß solide Grundkenntnisse in psychologischer Methodenlehre, wie sie im Rahmen des psychologischen Grundstudiums vermittelt werden, für das Verständnis der statistischen Teile der Arbeit ausreichen sollten.

Ich habe die Möglichkeiten multinomialer Modellierung im Laufe der letzten Jahre mit verschiedenen Freunden und Fachkollegen ausgiebig diskutiert und von ihrer Kritik und ihren Anregungen profitiert. Diesen Personen möchte ich an dieser Stelle herzlich danken. Ganz besonderer Dank gilt meinem akademischen Lehrer Jürgen Bredenkamp, der mich mit Problemen der Überprüfung allgemein-

psychologischer Theorien und potentiellen Lösungsmöglichkeiten vertraut gemacht und mich immer wieder ermutigt hat, nach weiteren, vielleicht besseren Lösungsansätzen zu suchen. Ohne seine Unterstützung, seine Anstöße und seine konstruktiven Rückmeldungen zu meinen Forschungsarbeiten wäre die vorliegende Arbeit nicht denkbar gewesen. Ebenfalls danke ich den Kolleginnen, Kollegen und Studierenden, die in den vergangenen Jahren die praktische Anwendung multinomialer Modelle mit mir gemeinsam erprobt haben. Besonders hervorheben möchte ich Ute J. Bayen und Axel Buchner, mit denen ich verschiedene Projekte gemeinsam durchgeführt, zahlreiche Diskussionen geführt und ungezählte E-mails ausgetauscht habe. Darüber hinaus habe ich von fachlichen Diskussionen mit William H. Batchelder, Klaus Fiedler, Gerd Gigerenzer, Xiangen Hu, Karl C. Klauer, Thorsten Meiser, Kevin Murnane und Reinhard Niederée stark profitiert.

Der Deutschen Forschungsgemeinschaft (DFG) danke ich für eine seit 1994 gewährte Sachbeihilfe zum Forschungsprojekt „Multinomiale Modelle kognitiver Täuschungen", das im DFG-Schwerpunktprogramm „Informationsverarbeitung im sozialen Kontext" angesiedelt ist. Diese Sachbeihilfe ermöglichte mir die Durchführung von Experimenten und die Optimierung von Modellansätzen, über die u.a. in der vorliegenden Arbeit berichtet wird. Bei Martin Brandt und Arndt Bröder lag die Projektarbeit über viele Jahre hinweg in guten Händen; sie haben mir praktische Unterstützung zukommen lassen, konstruktiv-kritisch an der Optimierung von Modellen und experimentellen Designs mitgewirkt und Feedback zu Vorformen der vorliegenden Arbeit zukommen lassen. Ihre Mitarbeit findet an verschiedenen Stellen der vorliegenden Arbeit ihren Niederschlag. Ebenfalls danken möchte ich den studentischen Hilfskräften Anja Dove, Johannes Hetzenegger, Tanja Hölper, Anna Kirnich, Anette Rüth, Renate Scheib, Momme von Sydow, Franz Urfels und Katja Wiemer, welche die Projektarbeit in den letzten Jahren tatkräftig unterstützt haben. Anna Kirnich hat zusätzlich die meisten Abbildungen in der vorliegenden Arbeit graphisch aufbereitet. Den Kolleginnen und Kollegen des DFG-Schwerpunktprogrammes „Informationsverarbeitung im sozialen Kontext" danke ich für die fruchtbare Zusammenarbeit auf und zwischen den Schwerpunktsymposien und -workshops.

Meine Familie hat mich in der Zeit vor und während der Arbeit am vorliegenden Manuskript nach Kräften unterstützt und viel Verständnis für meinen vollen Terminplan gezeigt, der selten so viel Zeit für das Privatleben ließ, wie es wünschenswert gewesen wäre. Steffi, Felix und nicht zuletzt meinen Eltern sei hierfür ganz herzlich gedankt.

Bonn, im Januar 2000
Edgar Erdfelder

Inhaltsverzeichnis

Vorwort iii

Inhaltsverzeichnis v

1. Gegenstand und Übersicht 1

2. Methodologischer Hintergrund 7

2.1 Introspektion und verbale Protokolle 7

2.2 Operationale Definitionen und Reduktionssätze 10

2.3 Hilfshypothesen und Korrespondenzregeln 11

2.4 Das Bedeutsamkeitsproblem 13

2.5 Das Fehlerproblem 21

2.6 Das Dekomponierungsproblem 32

2.7 Zusammenfassung 38

3. Was sind multinomiale Modelle? 41

3.1 Das einfache multinomiale Modell 45

3.2 Das verbundene multinomiale Modell 48

3.3 Parametrisierte multinomiale Modelle 50

3.4 Allgemeine multinomiale Verarbeitungsbaummodelle 56

3.5 Binäre multinomiale Verarbeitungsbaummodelle 69

3.5.1 Ein Beispiel: Das Enkodierung-Abruf-Modell 72

3.5.2 Weitere Beispiele: Hochschwellen-Modelle 80

3.6 Ereignisbaumrepräsentation von Verarbeitungsbaummodellen 83

3.7 Zusammenfassung 85

4. Analyse und Evaluation multinomialer Modelle 87

4.1 Identifizierbarkeit 88

4.1.1 Globale versus lokale Identifizierbarkeit 89

4.1.2 Notwendige und hinreichende Bedingungen für Identifizierbarkeit 92

4.1.3 Entscheidungen über Identifizierbarkeitsfragen 96

4.2 Parameterschätzung 99

4.2.1 Maximum-Likelihood-Schätzer 101

4.2.2 Minimum-Potenz-Divergenz-Schätzer 113

4.2.3 Andere Schätzer 117

4.2.4 Welchen Schätzer sollte man wählen? 118

4.2.5 Konfidenzintervalle und Konfidenzbereiche 120

4.3 Statistische Tests 128

4.3.1 Modellgeltungstests 129

4.3.2 Tests von Parameterhypothesen und Submodellen 134

4.3.2.1 Konstruktion geeignet restringierter Modelle 135

4.3.2.2 Prüfung restringierter Modelle 144

4.3.3 Teststärkekontrolle 148

4.3.3.1 Teststärke bei verbundenen multinomialen Modellen 151

4.3.3.2 Teststärke als Funktion der Modellparameter unter H1 153

4.3.3.3 Güte der klassischen c2(g)(df)-Teststärkeapproximation 155

4.3.3.4 Alternative Teststärkeapproximationen 156

4.3.3.5 Eine Monte-Carlo-Studie zur Power des EA-Modelltests 159

4.3.3.6 Teststärkeoptimierung 166

4.3.4 Die Prüfung von Omnibus-Nullhypothesen 172

4.3.5 Welche Prüfstatistik sollte man wählen?....178

4.3.6 Informationstheoretische Maße der Anpassungsgüte....182

4.4 Datenäquivalenz von Modellen und verwandte Probleme....186

4.5 Modellvalidierung....193

4.6 Behandlung von Identifizierbarkeitsproblemen....204

4.7 Computerprogramme....214

4.8 Zusammenfassung....217

5. Die Robustheit multinomialer Modellierung....221

5.1 Bisherige Studien zur Robustheit von MVB-Modellanalysen....223

5.2 Robustheit bei korrelierten und unkorrelierten Parametern....228

5.3 Eine Monte-Carlo-Studie zur Robustheit bei EA-Modellanalysen....233

5.4 Alternativen zum Standardvorgehen bei Parametervariabilität....243

5.4.1 Zerlegung von Aggregatstichproben....243

5.4.1.1 Vp-spezifische Analyse: Bootstrap- und Jackknife-Verfahren....244

5.4.1.2 Vp-spezifische Analyse: Bayessche Verfahren....247

5.4.1.3 Analyse von Teilstichproben basierend auf mehreren Vpn....249

5.4.2 Modelle mit variablen Parametern....250

5.4.2.1 Latent-Class-MVB-Modelle....250

5.4.2.2 Beta-MVB-Modelle und Dirichlet-MVB-Modelle....255

5.4.2.3 Rasch-MVB-Modelle....258

5.5 Zusammenfassung....264

6. MVB-Modelle als Rahmenmodelle....267

6.1 Log-lineare Modelle....268

6.2 Genetische Modelle 274

6.2.1 Ein Modell zur Schätzung von Vaterschaftswahrscheinlichkeiten 275

6.2.2 Prüfung des Modells zu Vaterschaftswahrscheinlichkeiten 282

6.3 Latent-Class-Modelle 285

6.3.1 Ein Modell zur Konzeptidentifikationsprozedur von Levine (1966) 286

6.3.2 Prüfung des Modells zur Konzeptidentifikationsprozedur 287

6.4 Binomialmischverteilungsmodelle 289

6.4.1 Ein Modell zum Konjunktionsfehler 293

6.4.2 Prüfung des Modells zum Konjunktionsfehler 297

6.5 Markoff-Ketten-Modelle und latente Markoff-Ketten-Modelle 302

6.5.1 Ein Dual-Code-Markoff-Modell verbalen Lernens 305

6.5.2 Prüfung des Dual-Code-Markoff-Modells 308

6.6 Gemischte und gemischte latente Markoff-Ketten-Modelle 312

6.7 Verallgemeinerte Signalentdeckungsmodelle 313

6.8 Modelle für Antworthäufigkeiten und Antwortzeiten 315

6.8.1 Modelle mit zustandsabhängigen Antwortzeiten 316

6.8.2 Modelle mit additiven Antwortzeiten 323

6.9 Zusammenfassung 330

7. Überblick über bisherige Anwendungen von MVB-Modellen 333

7.1 Gedächtnispsychologische Anwendungen 334

7.1.1 Wiedererkennen und Raten 335

7.1.2 Itemgedächtnis und Quellengedächtnis 336

7.1.3 Explizites und implizites Gedächtnis 346

7.1.4 Speicherung und Abruf 356

7.1.5 Retroaktive und proaktive Hemmung 364

7.1.6 Schemata und Skripts 365

7.1.7 Unmittelbare Reproduktion 368

7.1.8 Scheitern der Rekognition (Recognition Failure) 369

7.1.9 Gedächtnistäuschungen 371

7.2 Lernpsychologische Anwendungen 377

7.2.1 Diskriminationslernen 377

7.2.2 Explizites und implizites Lernen 379

7.2.3 Konzeptwissen versus Beispielwissen 380

7.2.4 Wissenszustände und Wissensräume 382

7.3 Denkpsychologische Anwendungen 386

7.3.1 Schlußfolgerndes Denken 386

7.3.1.1 Die Selektionsaufgabe nach Wason 386

7.3.1.2 Propositionales Schlußfolgern 390

7.3.1.3 Syllogistisches Schlußfolgern 392

7.3.2 Urteilen und Entscheiden 395

7.4 Sprachpsychologische Anwendungen 396

7.5 Wahrnehmungspsychologische Anwendungen 399

7.5.1 Reizerkennung und Raten 399

7.5.2 Parallele versus serielle Reizverarbeitung 400

7.5.3 Illusorische Konjunktionen bei der Objekterkennung 402

7.6 Sozialpsychologische Anwendungen 406

7.6.1 Soziale Kategorisierung 406

7.6.2 Illusorische Korrelationen 409

7.7 Entwicklungspsychologische Anwendungen 411

7.8 Psychometrische Anwendungen 413

7.8.1 Randomisierte Antworttechnik 413

7.8.2 Konsensusanalyse 415

7.8.3 Übereinstimmungsanalyse 417

7.8.4 Finite-State-Theorie für Multiple-Choice-Tests 418

7.9 Zusammenfassung 421

8. Modelle zu unterschiedlichen Quellengedächtnistests 425

8.1 Drei-Alternativen-Tests und Ja-Nein-Tests 426

8.2 Inklusions- und Exklusionstests 432

8.3 Zusammenfassung 442

9. Modelle zu Remember-Know-Prozeduren 445

9.1 Modelle zur einfachen Remember-Know-Prozedur 447

9.1.1 Das eindimensionale Kontinuum-Modell 450

9.1.1.1 Bisherige Befunde zum Kontinuum-Modell 451

9.1.1.2 Eine neue Untersuchung zum Kontinuum-Modell 458

9.1.1.3 Statistische Evaluation des Kontinuum-Modells 462

9.1.2 Das diskrete Zustandsmodell 466

9.2 Modelle zu erweiterten Remember-Know-Prozeduren 471

9.2.1 Das RKG-Kontinuum-Modell 475

9.2.2 Das diskrete RKG-Zustandsmodell 478

9.3 Zusammenfassung....485

10. Modelle zum Falschinformationseffekt....487

10.1 Ein Modell zum 2AFC-Rekognitionstest....493

10.2 Ein Modell zum Quellendiskriminationstest....498

10.2.1 Experiment 1: Gibt es einen Gedächtnisbeeinträchtigungseffekt?....512

10.2.2 Experiment 2: Effekte der Falschinformationswarnung....517

10.2.3 Experiment 3: Effekte der Nachinformationsmodalität....522

10.3 Zusammenfassung....526

11. Multinomiale Modellierung: Gegenwart und Zukunft....529

11.1 Die Kunst des multinomialen Modellierens....531

11.1.1 Heuristiken der Auswahl von Beobachtungskategorien....532

11.1.2 Heuristiken der Modellkonstruktion....534

11.2 Der Nutzen des multinomialen Modellierens....538

11.2.1 Multinomiale Modelle als Meßmodelle....538

11.2.2 Multinomiale Modelle als Theorieformalisierungen....541

11.3 Die Grenzen des multinomialen Modellierens....542

11.4 Die Zukunft des multinomialen Modellierens....546

11.5 Zusammenfassung....548

Literaturverzeichnis....551

Anhang....597

„A theory that is not stated in terms
of the data analysis to be used
cannot be tested."

Louis Guttman (1981, S. 63)

1. Gegenstand und Übersicht

Mit dem Titel der vorliegenden Arbeit wird eine Verbindung zwischen Konzepten aus zwei verschiedenen Wissenschaften — der Psychologie und der Mathematischen Statistik — hergestellt. Das ist einerseits gewagt, weil (a) mathematisch-statistische Modellannahmen im Regelfall nicht reibungsfrei zu den Anwendungsbedingungen in der Psychologie passen und weil (b) psychologische Theorien bzw. Hypothesen typischerweise umgangssprachlich formuliert werden, so daß sie oftmals nicht problemlos und stringent in mathematische Modelle „übersetzbar" sind. Andererseits liegt darin aber auch eine große Chance, weil Formalisierungen in Termini mathematischer Modelle den empirischen Gehalt und das tiefere Verständnis von Theorien verbessern sowie die Analyse der Beziehungen zwischen Theorien erleichtern können (De Leeuw, 1994; Deppe, 1977; Gigerenzer, 1981; Hintzman, 1991; Suppes, 1983). Zudem stellen stochastische Modelle einen geeigneten Rahmen bereit, in dem zentrale methodologische Probleme psychologischer Forschung potentiell lösbar werden. Diese Chancen sollen in der vorliegenden Arbeit für eine wichtige Teilklasse stochastischer Modelle — multinomiale Modelle — und ein in methodologischer Hinsicht besonders anspruchsvolles Teilgebiet der empirischen Psychologie — die kognitive Psychologie — ausgelotet werden, ohne daß dabei die Anwendungprobleme mathematischer Modelle in der Psychologie aus dem Blickfeld geraten.

Multinomiale Modelle sind stochastische Modelle für kategoriale (diskrete) Daten, die sich durch bestimmte Verteilungsannahmen auszeichnen (vgl. Kapitel 3). Anwendungen derartiger Modelle gibt es in der Psychologie schon recht lange. Allerdings waren die Anwendungen in der Vergangenheit im wesentlichen auf fest vorgegebene multinomiale Standardmodelle beschränkt, die losgelöst von konkreten psychologischen Fragestellungen entwickelt wurden und — ähnlich wie varianz- und regressionsanalytische Modelle — routinemäßig als Allzweckwerkzeuge für die Analyse kategorialer Daten Verwendung fanden. Hierzu zählen beispielsweise Log-Lineare Modelle (Agresti, 1990; Andersen, 1990; 1996; Bishop, Fienberg & Holland, 1975; Fienberg, 1977; Langeheine, 1986; Wickens, 1998a) und auch verschiedene psychometrische Modelle, die nicht an bestimmte psychodia-

gnostische Fragestellungen gebunden sind (z.B. Formann, 1984; Langeheine & Rost, 1996; Rost, 1988). Derartige Standardmodelle sind *nicht* Gegenstand der vorliegenden Arbeit. Vielmehr geht es um spezielle, substanzwissenschaftlich begründete multinomiale Modelle. Darunter verstehe ich Modelle, die entweder (a) durch verbal formulierte psychologische Theorien oder Hypothesen motiviert bzw. inspiriert wurden, (b) selbst den Status einer stochastisch formulierten psychologischen Hypothese haben oder (c) zur Beantwortung spezieller psychologischer Fragestellungen konstruiert wurden. In jedem Fall liegt also eine sehr enge Koppelung zwischen dem Modell und einer speziellen psychologischen Fragestellung oder Hypothese vor. Losgelöst von dieser Fragestellung bzw. Hypothese ist das stochastische Modell nicht sinnvoll anwendbar. Typische Anwendungsfälle sind gerade dadurch gekennzeichnet, daß die psychologischen Annahmen, die in ein solches Modell eingehen, empirisch überprüft werden oder daß ein bereits als psychologisch validiert geltendes Modell zur Beantwortung neuer psychologischer Fragen eingesetzt wird.

Wissenschaftshistorisch bewegt sich die skizzierte Modellklasse in der Tradition mathematischer Lernmodelle (Atkinson, Bower & Crothers, 1965; Deppe, 1977; Tack, 1976; Wickens, 1982), die in den fünfziger und frühen sechziger Jahren ihre Blüte hatten.[1] Publikationen zu mathematischen Lernmodellen sind inzwischen rar worden. Ich teile nicht die Auffassung, daß sich hierin eine Krise formaler Modellbildung in der Psychologie ausdrückt oder daß die Psychologie noch nicht „reif" genug für formale Theorieansätze sei (siehe bereits Lewin, 1951/1982). Vielmehr ist die Krise mathematischer Lerntheorien in erster Linie Ausdruck der Krise einer behavioristisch geprägten Lernpsychologie, die sich in den mathematischen Modellen der fünfziger und sechziger Jahre widerspiegelt (vgl. Batchelder & Boyd, 1985, S. 126). Behavioristische Theorien sind inzwischen weitgehend durch kognitionspsychologische Theorien abgelöst worden, weil zu viele interessante empirische Phänomene im Rahmen behavioristischer Theorien nicht befriedigend erklärt werden konnten. Derartige Paradigmenwechsel (Groeben & Scheele, 1977; Kuhn, 1962) konnten nicht ohne Auswirkungen auf die Art der formalen Modellbildung bleiben: Stimulus-Response-Modelle traten in den Hintergrund, wahrnehmungs-, denk- und gedächtnispsychologische Anwendungen stochastischer Modelle dagegen — trotz vereinzelter skeptischer Stimmen[2] — zunehmend

[1] Das gilt für die statistischen Modellklassen, die verwendet wurden, nicht aber unbedingt für die Methoden der Modellevaluation. Die Evaluation mathematischer Lernmodelle erfolgte häufig anhand deskriptiv-statistischer Tests von Modellimplikationen (s. z.B. die Arbeit von Kintsch & Morris, 1965). Die in dieser Arbeit favorisierten Techniken der Modellevaluation erlauben strengere Tests der Modellannahmen (s. dazu Kapitel 4 der vorliegenden Arbeit).

[2] So haben etwa Tulving und Madigan (1970) die Signalentdeckungstheorie als Instrument der modellbasierten Analyse von Rekognitionstestdaten in der Gedächtnispsychologie noch rigoros abgelehnt. Für sie war die Signalentdeckungstheorie ein exotisches Forschungsinstrument: „(It) provides

in den Vordergrund. In einem einflußreichen Sammelreferat zu mathematischen Lernmodellen haben Greeno und Bjork (1973) diesen Übergangsprozeß als Wechsel von der „mental chemistry" zu einer „mental forestry" charakterisiert, „in which our concerns are moving toward an interest in the processes by which (...) cognitive structures are acquired and modified" (a.a.O., S. 82).

Die potentielle Fruchtbarkeit stochastischer Modellbildung für die Kognitionspsychologie wurde von einigen Autoren zwar schon früh erkannt (vgl. das Übersichtsreferat von Greeno & Bjork, 1973); erst in den achtziger Jahren erschienen jedoch einflußreiche Artikel (z.B. Batchelder & Riefer, 1980; Riefer & Batchelder, 1988) und Monographien (vgl. Luce, 1986; Townsend & Ashby, 1983; Wickens, 1982), welche die vielfältigen Möglichkeiten stochastischer Modellbildung in der kognitiven Psychologie einem breiteren Fachpublikum deutlich machten. Von einer Krise kann diesbezüglich derzeit kaum die Rede sein. Im Gegenteil: Betrachtet man die Publikationsjahre von Arbeiten, die ein neueres Sammelreferat zum Thema „multinomiale Modellierung" zusammenfaßt (Batchelder & Riefer, 1999), so ist ein stetiger Anstieg der einschlägigen Veröffentlichungen von 1980 bis 1999 auszumachen. Vielleicht noch eindrucksvoller ist die Tatsache, daß dieses Sammelreferat zum Zeitpunkt seines Erscheinens in Teilen schon wieder veraltet ist (vgl. Kapitel 7).

Die weitaus meisten Modelle, die in der vorliegenden Arbeit diskutiert werden, gehören zur Klasse der sogenannten multinomialen Verarbeitungsbaum-Modelle (multinomial processing tree models, vgl. Batchelder & Riefer, 1999; Hu & Batchelder, 1994; Riefer & Batchelder, 1988)[3]. Verwandte Modellklassen finden ebenfalls Berücksichtigung, aber nur insoweit sie gemeinsame Struktureigenschaften mit multinomialen Verarbeitungsbaum-Modellen (MVB-Modellen) aufweisen, so daß sie zur Lösung ähnlicher substanzwissenschaftlicher Probleme eingesetzt werden können. Obwohl MVB-Modelle anhand rein formaler statistischer Kriterien definierbar sind (vgl. Kapitel 3) und natürlich auch zur Beantwortung nichtpsychologischer Fragestellungen eingesetzt werden können (vgl. z.B. Abschnitt 6.2), soll durch die Bezeichnung deutlich gemacht werden, daß sich die Modellklasse besonders gut als Rahmen zur Formalisierung und Überprüfung von Hypothesen über kognitive Verarbeitungsprozesse eignet.

Die Perspektive, die in dieser Arbeit zugrunde gelegt wird, ist die der experimentellen kognitiven Psychologie. Dennoch muß die mathematisch-statistische Seite multinomialer Modellierung insoweit vertiefend behandelt werden, wie das aus der Perspektive problembewußter, gründlicher Anwender erforderlich ist.

---

good research opportunities for those students who like exotic quantitative techniques" (S. 468). Diese Äußerung blieb jedoch schon Anfang der 70er Jahre nicht unwidersprochen (vgl. Bernbach, 1971).

[3]Der Begriff „multinomial processing tree models" wurde von Batchelder und Riefer (1990) eingeführt. Ich folge hier dem Übersetzungsvorschlag von Klauer (1996b).

Multinomiale Modelle werden in dieser Arbeit primär als mathematisch-statistische Systeme betrachtet, die eine präzise und zugleich flexible formale Sprache bereitstellen, mit deren Hilfe sich kognitionspsychologische Theorien gut explizieren und überprüfen lassen, sofern geeignete empirische Datensätze zur Verfügung stehen.

Die Arbeit verfolgt im wesentlichen drei Ziele. Zunächst einmal soll die methodologische Bedeutung substanzwissenschaftlich motivierter multinomialer Modelle innerhalb der kognitiven Psychologie herausgestellt werden. Wie deutlich werden wird, sind derartige Modelle zwar keineswegs ein Allheilmittel zur Lösung methodologischer Probleme in der kognitiven Psychologie, dennoch aber ein ausgesprochen nützliches Werkzeug, sofern bestimmte Anwendungsbedingungen und Grenzen der Methodik gebührend beachtet werden. Das zweite Ziel ist die Beschreibung und Diskussion der Eigenschaften von MVB-Modellen und verwandten Modellen sowie der Methoden ihrer statistischen Analyse. Hierbei wird besonderes Gewicht auf solche Probleme gelegt, die sich bei Verwendung derartiger Modelle zum Zwecke der Theorie- und Hypothesenprüfung in der kognitiven Psychologie stellen. Pragmatische Lösungsvorschläge werden unterbreitet. Das dritte Ziel ist schließlich die konkrete Anwendung derartiger Modelle auf ausgewählte psychologische Problemstellungen.

In Übereinstimmung mit diesen Zielen gliedert sich die vorliegende Arbeit in drei Teile. Das folgende Kapitel 2 bildet den ersten Teil. Dort wird der methodologische Rahmen abgesteckt, in den substanzwissenschaftlich motivierte multinomiale Modelle eingeordnet werden können. Die nächsten drei Kapitel bilden den zweiten Teil. Sie sind den primär statistischen Aspekten multinomialer Modelle gewidmet, d.h. den Modellannahmen und Modelleigenschaften (Kapitel 3), der Identifizierbarkeitsprüfung, Parameterschätzung, Modellprüfung und Modellvalidierung (Kapitel 4) und schließlich der Robustheit gegenüber Verletzungen der Verteilungsannahmen (Kapitel 5). Die restlichen Kapitel bilden den dritten Teil der Arbeit. Kapitel 6 behandelt die Beziehungen zwischen MVB-Modellen und anderen Modellklassen und erläutert diese Beziehungen anhand von MVB-Reanalysen empirischer Datensätze, die zuvor nicht im MVB-Kontext betrachtet wurden. Kapitel 7 präsentiert einen Überblick über Anwendungen von MVB-Modellen in der kognitiven Psychologie. Die nächsten Kapitel stellen neue Modelle und Befunde zu Quellendiskriminationsaufgaben (Kapitel 8), zu Remember-Know-Gedächtnistests (Kapitel 9), und zum Falschinformationseffekt vor (Kapitel 10). Im abschließenden elften Kapitel werden Möglichkeiten und Grenzen multinomialer Modelle in der kognitiven Psychologie kritisch gewürdigt und mögliche zukünftige Entwicklungen angedeutet. Jedes der Kapitel 2 bis 11 schließt mit einer kurzen Zusammenfassung der wichtigsten Resultate.

Da eine Gesamtdarstellung des Forschungsfeldes angestrebt wurde, die bislang nicht existiert, mußten naturgemäß viele Arbeiten anderer Autorinnen und Autoren sowie eigene, bereits publizierte Arbeiten aufgearbeitet, in den Gesamtzusammenhang eingeordnet und zusammenfassend berichtet werden. Dies gilt insbesondere für die Darstellung der zu behandelnden Modellklassen (Kapitel 3), der statistischen Analysemethoden (Kapitel 4) und der bisherigen Anwendungen von MVB-Modellen in der kognitionspsychologischen Forschung (Kapitel 7). Darüber hinaus werden in der vorliegenden Arbeit jedoch auch viele neue Ansätze und Resultate vorgestellt, und zwar (a) Neuerungen auf dem Gebiet der Grundlagen bzw. der statistischen Methoden, (b) Neuentwicklungen auf dem Gebiet der kognitionspsychologischen Modellbildung und (c) neue empirische Resultate zur Anwendung multinomialer Modelle auf kognitionspsychologisch relevante Fragestellungen und Hypothesen.

Die einzelnen Kapitel der Arbeit wurden so verfaßt, daß sie im Prinzip separat lesbar sein sollten, wenn die Leserin bzw. der Leser mit den Grundlagen der multinomialen Modellierung vertraut ist. Soweit Vorwissen aus anderen Teilen der Arbeit erforderlich ist, wird auf die entsprechenden Kapitel oder Abschnitte verwiesen.

2. Methodologischer Hintergrund

Ein zentrales Problem psychologischer Forschung ist die Klärung der Beziehungen zwischen *theoretischen Konstrukten* und *empirischen Konzepten* (Roskam, 1996a). Psychologische Theorien beinhalten in der Regel Aussagen über theoretische Konstrukte wie z.B. mentale Systeme, Fähigkeiten, Zustände und Prozesse, die der direkten Beobachtung entzogen, d.h. *latent* sind. Versteht man Psychologie als empirische Wissenschaft, so müssen psychologische Theorien aber anhand empirischer Konzepte überprüfbar sein, die prinzipiell durch jede beliebige Person *beobachtbar* sind. Folglich muß eine methodologisch tragfähige Brücke zwischen theoretischen Konstrukten einerseits und beobachtbaren Tatbeständen andererseits hergestellt werden. Wie Guttmann (1981) festgestellt hat, ist eine Theorie, die dies nicht in ganz expliziter Weise leistet, faktisch nicht prüfbar (vgl. das dem ersten Kapitel vorangestellte Zitat). Verschiedentlich ist dieses Problem als „Theorie-Empirie-Überbrückungsproblem“ bezeichnet worden (vgl. z.B. Erdfelder & Bredenkamp, 1994; Herrmann, 1979; Steyer & Eid, 1993).

Man kann die Geschichte der empirischen Psychologie als Folge unterschiedlicher Versuche rekonstruieren, das Theorie-Empirie-Überbrückungsproblem theorieübergreifend in befriedigender Weise zu lösen. In den folgenden Abschnitten werde ich zunächst die wichtigsten historischen Lösungsansätze in chronologischer Folge darstellen und diskutieren, nämlich die Methode der Introspektion bzw. der verbalen Protokolle (Abschnitt 2.1), operationale Definitionen und Reduktionssätze (Abschnitt 2.2) sowie Hilfshypothesen und Korrespondenzregel (Abschnitt 2.3). Der letztgenannte Ansatz vermeidet einige Probleme der beiden erstgenannten Ansätze. Er wird gegenwärtig in der experimentellen Kognitionsforschung favorisiert. Allerdings ist auch dieser Ansatz mit methodologischen Problemen behaftet, die in den Abschnitten 2.4 (Bedeutsamkeitsproblem), 2.5 (Fehlerproblem) und 2.6 (Dekomponierungsproblem) besprochen werden. Es wird herausgearbeitet und an verschiedenen Beispielen illustriert, daß stochastische Modellbildungsansätze — speziell multinomiale Modelle — einen geeigneten Rahmen liefern, um diesen Problemen erfolgreich begegnen zu können. Die wichtigsten Resultate dieses Kapitels werden in Abschnitt 2.7 zusammengefaßt.

2.1 Introspektion und verbale Protokolle

Die wichtigste Methode in der Frühphase der wissenschaftlichen Psychologie war die *Introspektion,* d.h. die Selbstbeobachtung mentaler Prozesse und Zustände. Durch Introspektion wurden Selbstbeobachtungsprotokolle erzeugt, die als Surrogat für die eigentlich interessierenden mentalen Prozesse dienten und sie somit quasi in Beobachtungstatbestände verwandelten. An der Introspektionsmethodik ist jedoch von Anfang an Kritik geübt worden, auch — und sogar besonders heftig

— von maßgeblichen Fachvertretern wie etwa Wilhelm Wundt (zusammenfassend: Erdfelder, 1994, S. 55-62; Pongratz, 1998). Kernpunkte der Kritik waren mögliche Beeinflussungen oder gar Verfälschungen der interessierenden mentalen Prozesse durch den Akt der Introspektion selbst (siehe bereits Kant, 1798/1968, S. 415-417) sowie das Fehlen wichtiger experimentalpsychologischer Kriterien wie z.B. Wiederholbarkeit und planmäßige Variierbarkeit (Wundt, 1907). Durch Verwendung der Retrospektionsmethodik anstelle der Introspektionsmethodik, d.h. durch nachträgliche, gedächtnisbasierte Protokollierung von Bewußtseinsprozessen wie sie u.a. von Karl Bühler (1908) vertreten wurde, können — so die Kritiker — diese Probleme nicht beseitigt werden. Im Gegenteil: Bei diesem Vorgehen seien zusätzlich „... Erinnerungstäuschungen Tür und Tor geöffnet" (Wundt, 1908, S. 452).

Auch 90 Jahre nach der Wundt-Bühler-Debatte ist die Selbstbeobachtungsmethodik noch nicht aus dem Methodenarsenal der kognitiven Psychologie verschwunden (vgl. Funke, 1996). Für einige Autoren bleibt sie prinzipiell unverzichtbar, will man nicht den Gegenstand der Psychologie in unzulässiger Weise einschränken (z.B. Greve, 1996). Durch die einflußreichen Arbeiten von Ericsson und Simon (1980; 1993) ist Selbstbeobachtungsmethodik nun unter den Bezeichnungen „Methode des lauten Denkens", „Methode des verbalen Berichts" oder „protocol analysis" bekannt. Im Gegensatz zu den Introspektionisten aus der Frühphase der experimentellen Psychologie haben Ericsson und Simon jedoch eine Reihe von Adäquatheitsbedingungen formuliert, die erfüllt sein müssen, damit verbale Berichte als sinnvoll interpretierbar gelten können. Ericsson und Simon lassen lediglich phonologisch kodierte Inhalte des Arbeitsgedächtnisses als Grundlage für verbale Berichte zu und schließen Retrospektionen aufgrund des episodischen Langzeitgedächtnisses sowie Berichte über Prozesse, die nicht im phonologischen Arbeitsgedächtnis ablaufen, prinzipiell aus. Als Konsequenz hieraus ist die Selbstbeobachtungsmethodik auch nach Ansicht ihrer Befürworter nur noch auf einen kleinen Teil kognitiver Prozesse anwendbar (z.B. auf Problemlösungs- oder Schlußfolgerungsprozesse), und auch diesbezüglich sind Zweifel an der Validität verbaler Berichte nach wie vor durchaus angebracht (Russo, Johnson & Stephens, 1989).

Es wird zunehmend klarer, daß „kognitive Psychologie" keinesfalls mit „Bewußtseinspsychologie" gleichgesetzt werden darf. Ein großer Teil kognitiver Prozesse läuft offenbar ohne Beteiligung des Bewußtseins ab, nimmt aber dennoch massiv Einfluß auf das Verhalten von Individuen. Nisbett und Wilson (1977) haben in einer klassischen Arbeit u.a. nachgewiesen, daß Personen, die zwischen perzeptuell nicht unterscheidbaren Objekten wählen dürfen, in ihrem Wahlverhalten offenbar durch die räumlichen Positionen der Objekte beeinflußt werden. Genau dies bestreiten die Personen jedoch, wenn man sie nach den Motiven für ihre

Wahl befragt. In ganz ähnlicher Weise zeigte Bredenkamp (1990) eine Inkonsistenz zwischen den verbalen Berichten eines Rechenkünstlers und dessen Antwortzeiten für bestimmte Rechenaufgaben auf. Auch Introspektionen von Experten garantieren also offenkundig keinen unverfälschten Blick auf die mentalen Prozesse, die der Expertise zugrunde liegen. Man muß sogar umgekehrt vermuten, daß die Validität von Selbstbeobachtungen bei Experten eher geringer als bei Novizen ausgeprägt ist: Folgt man z.B. der einflußreichen Theorie von John Anderson (1982a; 1992; 1993), so geht Expertise unter anderem mit einer Entlastung des Arbeitsgedächtnisses durch Transformierung deklarativen Wissens in prozedurales Wissen einher („proceduralization"). Letzteres kann dann automatisiert, d.h. ohne Abruf deklarativen Wissens, angewendet werden und muß daher dem Bewußtsein nicht mehr (oder allenfalls nur noch partiell) zugänglich sein (vgl. auch Gopnik, 1993).

Im Lichte neuerer Forschungsergebnisse zum impliziten Gedächtnis und zum impliziten Lernen sind Dissoziationen zwischen Verhalten und Selbstbeobachtungen durchaus verständlich und auch theoretisch aufklärbar. Es gibt zahlreiche Beispiele aus nahezu allen Bereichen der kognitiven Psychologie, die klar zeigen, daß spezielle Fertigkeiten und Verhaltensweisen durch Erfahrungen in der Vergangenheit beeinflußt werden können, ohne daß diese Beeinflussungen bzw. die erworbenen Fertigkeiten — folgt man dem verbalen Bericht der Versuchsperson (Vp) — bewußt sein müssen (vgl. Bredenkamp, 1998a, S. 31-33; 1998b). Es ist nicht einmal nötig, daß die zugrundeliegenden Erfahrungen erinnerbar sind bzw. ursprünglich bewußten Erfahrungen entsprechen: Auch kaum identifizierbare („subliminale") Stimuli können späteres Verhalten nachweislich gerichtet beeinflussen (Greenwald & Draine, 1997; Kihlstrom, Barnhardt & Tataryn, 1992).

Auch wenn Bunge und Ardila (1987, S. 63) sicher etwas zu weit gehen, wenn sie die Selbstbeobachtung aus der Gruppe der psychologischen Methoden gänzlich ausschließen, so ist doch unbestreitbar, daß das Theorie-Empirie-Überbrückungsproblem durch Selbstbeobachtungsmethodik bestenfalls in sehr begrenzten Anwendungsfeldern lösbar ist. Als genereller Lösungsansatz ist Introspektion schon allein deshalb nicht verwendbar, weil viele verhaltenswirksame kognitive Prozesse offenbar keinen Niederschlag im Arbeitsgedächtnis finden und deshalb schon vom Ansatz her über den verbalen Report nicht kommuniziert werden können (vgl. Erdfelder, 1994). Sofern sich Prozesse im Arbeitsgedächtnis manifestieren, ist zumindest unklar, ob sie durch den verbalen Bericht bzw. durch Erlebnisdeskriptionen erschöpfend und verzerrungsfrei erfaßt werden (Bredenkamp, 1998a; Feger & Graumann, 1983).

2.2 Operationale Definitionen und Reduktionssätze

Die radikale historische Antwort auf die Probleme der Introspektionsmethodik war bekanntlich der *Behaviorismus* und in dessen Gefolge der sog. *Operationalismus,* welcher wissenschaftstheoretisch dem frühen logischen Empirismus zuzuordnen ist (Westermann, 1987b, Abschnitt 2.2). Der klassische Behaviorismus meidet eine Auseinandersetzung mit dem Theorie-Empirie-Überbrückungsproblem, indem er mentalistische Konzepte aus der Psychologie verbannt bzw. — in seiner metaphysischen Variante — die Existenz mentaler Strukturen und Prozesse sogar bestreitet (Watson, 1913). In psychologischen Theorien dürfen demnach nur noch empirische Konzepte auftauchen, so daß sich das Überbrückungsproblem gar nicht erst stellt. Dieser Ansatz hat sich jedoch schnell als zu restriktiv herausgestellt, weil er wichtige aktive Forschungsfelder der Psychologie, beispielsweise die Wahrnehmungspsychologie, nicht befriedigend abdecken konnte.

Der Operationalismus kann als abgeschwächte Form des Behaviorismus gelten, zielt allerdings in seiner ursprünglichen Zielsetzung nicht nur auf die Psychologie, sondern auf alle empirischen Wissenschaften ab (Bridgeman, 1927). Im Gegensatz zum Behaviorismus erlaubt der Operationalismus zwar theoretische Begriffe wie z.B. „Empfindungsintensität", „Gedächtnisspurenstärke" etc. als Theorieelemente und ist damit prinzipiell eine mögliche methodologische Basis für die Wahrnehmungspsychologie und die kognitive Psychologie. Allerdings wird verlangt, daß theoretische Begriffe grundsätzlich vollständig auf beobachtbare Tatbestände reduzierbar sind. Der Sinn derartiger theoretischer Konzepte besteht somit eigentlich nur darin, ökonomische sprachliche Verkürzungen für komplexe empirische Konzepte bereitzustellen (Westermann, 1987b, S. 10).

Anfangs glaubte man, mit sogenannten *operationalen Definitionen* theoretischer Begriffe die Zielsetzung des Operationalismus erreichen zu können. Eine operationale Definition ist ein universeller Allsatz, mit dem ein theoretischer Begriff (Definiendum) ausschließlich anhand von empirischen Konzepten (Definiens) eingeführt wird. Hierbei müssen alle Kriterien einer expliziten, „totalen" Definition des Definiendums durch das Definiens erfüllt sein (Essler, 1982).[4] Eine operationale Definition des quantitativen theoretischen Konstrukts „Empfindungsintensität" könnte umgangssprachlich paraphrasiert folgendermaßen aussehen: „Für alle Personen x und alle Zeitpunkte t: x hat zum Zeitpunkt t die Empfindungsintensität y genau dann, wenn x das Urteil y nennt, falls x zum Zeitpunkt t

[4]Einerseits müssen also Definiendum und Definiens in allen Sätzen einer Theorie austauschbar sein, ohne daß damit eine Bedeutungsveränderung einhergeht (universelle Eliminierbarkeit des Definiendums durch das Definiens); andererseits darf eine Theorie nach Ergänzung um eine operationale Definition nicht mehr Sätze implizieren als zuvor (Nichtkreativität einer Definition; vgl. Essler, 1982).

an einem Größenschätzexperiment für Stimulusintensitäten teilnimmt".[5] Schon Rudolf Carnap (1936; 1937) zeigte allerdings, daß diese Form der Einführung theoretischer Konstrukte die paradoxe Konsequenz hat, auch bei Nichtvorliegen von Beobachtungen zu bestimmten Aussagen gezwungen zu sein, die den theoretischen Begriff beinhalten. So ist beispielsweise für eine Person *x*, die zum Zeitpunkt *t* nicht an einem Größenschätzexperiment teilnimmt, das konditional formulierte Definiens („*x* nennt das Urteil *y*, *falls x* zum Zeit *t* an einem Größenschätzexperiment teilnimmt") automatisch wahr. Wegen der Äquivalenz von Definiens und Definiendum muß Person *x* für diesen Zeitpunkt *t* folglich die Empfindungsintensität *y* zugesprochen werden, auch wenn es keine empirische Grundlage hierfür gibt. Dies ist nicht nur paradox, sondern führt darüber hinaus leicht zu Widersprüchen, weil man natürlich in völlig analoger Weise andere, von *y* verschiedene Empfindungsintensitäten *y´*, *y´´* etc. einführen kann, die einer nicht beobachteten Person dann ebenfalls zugesprochen werden müßten.

Vermeidbar ist dieses Problem, wie ebenfalls Carnap (1936; 1937) zeigte, durch sogenannte *bilaterale Reduktionssätze*. Hierbei handelt ist sich um *bedingte* Definitionen, die eine Äquivalenz zwischen einem theoretischen und einem empirischen Begriff nur für den Fall herstellen, daß zuvor geeignete Beobachtungen durchgeführt wurden (Essler, 1982; Essler & Trapp, 1977). Im oben genannten Fall könnte der bilaterale Reduktionssatz zum Konstrukt „Empfindungsintensität" umgangssprachlich beispielsweise lauten: „Wenn eine Person *x* zum Zeitpunkt *t* an einem Größenschätzexperiment für Stimulusintensitäten teilnimmt, so hat *x* zum Zeitpunkt *t* die Empfindungsintensität *y* genau dann, wenn *x* zum Zeitpunkt *t* das Urteil *y* nennt". Man beachte, daß es sich hierbei nicht mehr um eine explizite Definition handelt, da das Kriterium der universellen Eliminierbarkeit des Definiendums durch das Definiens verletzt ist: Für nicht beobachtete Personen ist das Konstrukt „Empfindungsintensität" nicht definiert. Der enge Rahmen, den der klassische Operationalismus für theoretische Konstrukte absteckt, wird damit also schon verlassen.

2.3 Hilfshypothesen und Korrespondenzregeln

Wenn in der modernen Psychologie von „Operationalisierungen" theoretischer Begriffe oder hypothetischer Konstrukte die Rede ist, so sind damit nicht operationale Definitionen im oben genannten Sinne gemeint, sondern Aussagen, die in

[5]Der Operationalismus wurde in der Psychologie maßgeblich durch den Wahrnehmungspsychologen S.S. Stevens vertreten (vgl. z.B. Stevens, 1935). Stevens ließ nur direkte „Messungen" (Operationalisierungen) von Empfindungen zu und lehnte indirekte Methoden der Empfindungsmessung ab. Diese Auffassung war jedoch in der Psychophysik von Anfang an sehr umstritten (s. bereits Fechner, 1860). Aus heutiger Sicht kann das Argument, nur auf der Grundlage direkter Skalierungen von Empfindungsintensitäten könne die Frage nach dem psychophysikalischen Gesetz empirisch entschieden werden, nicht mehr überzeugen (Falmagne, 1985).

formaler Hinsicht eher bilateralen Reduktionssätzen entsprechen. Es ist nicht intendiert, eine vollständige explizite Definition des theoretischen Begriffs in Termini empirischer Konzepte zu liefern, sondern lediglich, eine Aussage über die *empirische Manifestation* eines mentalen Zustands oder Prozesses bei Vorliegen wohldefinierter Beobachtungsbedingungen (experimentelle Untersuchungsanordnungen, Tests etc.) zu machen. Derartige Manifestationsregeln werden aber nicht als unumstößliche, unveränderbare Festsetzungen aufgefaßt, sondern haben selbst den Charakter von prinzipiell revidierbaren Hypothesen. Es ist also treffender, statt von Reduktionssätzen besser von „Hilfshypothesen" (Bredenkamp, 1980; Gadenne, 1976; 1984; Hager, 1992, Abschnitt 6.2) oder „Korrespondenzregeln" (Herrmann, 1979) zu sprechen, die in Verbindung mit Theorien, welche theoretische Konstrukte beinhalten, die Ableitung empirisch prüfbarer Hypothesen erlauben. Dies ist ausreichend, um die empirische Prüfbarkeit psychologischer Theorien sicherzustellen. Gleichzeitig bleiben die theoretischen Konstrukte „offen" in dem Sinne, daß deren Bedeutung nicht vollständig auf bestimmte empirische Konzepte reduzierbar ist (Herrmann & Stapf, 1971). Weite Teile der gegenwärtigen experimentellen Psychologie scheinen diesem Lösungsansatz zum Theorie-Empirie-Überbrückungsproblem zu folgen (vgl. z.B. Bredenkamp, 1980; Erdfelder & Bredenkamp, 1994; Hager, 1987; 1992).

Allerdings ist dieser Ansatz — im folgenden kurz „Operationalisierungsansatz" genannt — mit Folgeproblemen behaftet, auf die in den nächsten Abschnitten unter den Stichworten „Bedeutsamkeitsproblem", „Fehlerproblem" und „Dekomponierungsproblem" näher eingegangen wird. Ein zentrales Problem ist zunächst die Unklarheit und die prinzipielle Beliebigkeit bei der Zuordnung von empirischen Konzepten zu theoretischen Begriffen. Wenn die zu prüfende Theorie oder Hypothese selbst keine präzise Aussage darüber macht, welcher empirische Indikator für ein theoretisches Konstrukt optimal ist, nach welchen Kriterien ist dann auszuwählen? Es ist immer möglich, zu ein und demselben theoretischen Konstrukt mehrere Hilfshypothesen zu formulieren, derart, daß das Konstrukt an unterschiedliche Test- oder Beobachtungsresultate gekoppelt wird. So sind Empfindungsintensitäten z.B. nicht nur anhand von Größenschätzurteilen, sondern unter anderem auch anhand von Kategorialurteilen oder intermodalen Zuordnungen (cross-modality matchings) operationalisiert worden (Engen, 1972). Zwar ist es prinzipiell möglich, durch konjunktive Verknüpfung derartiger Operationalisierungen ein empirisch gehaltvolles Reduktionssatzsystem, eine prüfbare „Miniaturtheorie" des theoretischen Konstrukts zu erzeugen (Westmeyer, 1972). Allerdings sind die Chancen, daß sich ein solches System empirisch bewährt, im allgemeinen recht gering. Dies gilt insbesondere dann, wenn deterministische Zusammenhänge zwischen theoretischen Konstrukten und empirischen Konzepten unterstellt werden. Typischerweise stehen die empirischen Variablen, die als un-

terschiedliche Operationalisierungen eines Konstrukts in Frage kommen, nicht in perfekt eineindeutiger Beziehung zueinander, wenn man sie simultan an den gleichen Personen erhebt. Folglich können sie nicht ausnahmslos perfekte empirische Indikatoren des gleichen theoretischen Konstrukts sein. Auch die Berücksichtigung von Meßfehlern ändert im allgemeinen nicht viel: Immer wird man zumindest einige Individuen finden, auf die nach einer Manifestationsregel (Test A) eine bestimmte theoretisch postulierte Eigenschaft (z.B. „Intelligenz") sehr wahrscheinlich zutrifft, nach einer anderen Manifestationsregel (Test B) dagegen sehr wahrscheinlich nicht.

Im Grunde genommen können derartige Ergebnisse nicht überraschen: Warum sollten unterschiedliche Test- oder Beobachtungsvariablen, die auf der Grundlage von Plausibilitätsüberlegungen mehr oder minder willkürlich als Indikatoren eines theoretischen Konstrukts ausgewählt werden, konsistente bzw. durch monotone Transformationen ineinander überführbare Ergebnisse liefern? Ein solches Ergebnis kann eigentlich nur zufällig resultieren und sollte entsprechend selten zu beobachten sein. Roskams (1979, S. 124) scharfer Kritik an der Operationalisierungsmethodologie kann also nicht leicht widersprochen werden: „Also, der Weg [der] Operationalisierung führt zu nichts, es sei denn durch reines Glück".

2.4 Das Bedeutsamkeitsproblem

Scheitert ein System von Hilfshypothesen — kann also die Konjunktion mehrerer Manifestationsregeln zum gleichen theoretischen Konstrukt empirisch nicht aufrecht erhalten werden — so impliziert dies, daß nicht alle vorgeschlagenen Regeln das gleiche Konstrukt erfassen. Bewegt man sich im Rahmen einer realistischen Wissenschaftskonzeption, so muß es demnach mindestens eine falsche Hilfshypothese in dem System geben.[6] Es stellt sich folglich das Problem der *Variablenvalidität* (Bredenkamp, 1980; Hager & Westermann, 1983): Wie gut repräsentieren die in den verschiedenen Manifestationsregeln genannten empirischen Konzepte das interessierende theoretische Konstrukt? Welcher Hilfshypothese kann getraut werden, welcher nicht?

Im Rahmen des Operationalisierungsansatzes gibt es auf diese Fragen keine befriedigenden Antworten. Die Konsequenz für die Forschungspraxis ist, daß aus

[6]Im Rahmen einer instrumentalistischen Wissenschaftskonzeption, nach der theoretische Begriffe lediglich „nützliche Fiktionen" (Herrmann, 1983) ohne Realitätsbezug sind, muß diese Konsequenz natürlich nicht zwangsläufig gezogen werden. Man könnte sich z.B. pragmatisch entschließen, ein Subsystem von Hilfshypothesen zu verwenden, das empirisch haltbar ist, und die „störenden" Hilfshypothesen einfach ausschließen. Die Unterscheidung „richtiger" und „falscher" Hilfshypothesen macht aus instrumentalistischer Sicht keinen Sinn. Ein Problem dieses Ansatzes ist jedoch die drohende Gefahr eines degenerativen Forschungsprogramms (Lakatos, 1970) mit sukzessive nachlassendem empirischen Gehalt der aufeinander folgenden Theorien.

der Menge der in der Vergangenheit vorgeschlagenen Hilfshypothesen zu einem bestimmten theoretischen Konstrukt im Regelfall ohne nähere Begründung eine Hilfshypothese ausgewählt und zur Überprüfung von psychologischen Theorien herangezogen wird, in denen das betreffende Konstrukt vorkommt. Verschiedene Forscher, welche die gleichen Theorien prüfen wollen, können dabei durchaus auf unterschiedliche Hilfshypothesen zurückgreifen. Die persönliche Präferenz dürfte durch die Präferenzen von Kolleginnen und Kollegen aus dem eigenen Fachbereich, durch die „Schule", der man sich zugehörig fühlt, durch die Präferenzen der Mehrheit der relevanten „scientific community" und durch die (vermuteten) Präferenzen des Herausgebers der Fachzeitschrift maßgeblich geprägt sein, in der man seine Forschungsresultate veröffentlichen möchte. Liest man einschlägige Forschungspublikationen, so entsteht der Eindruck, daß der bloße Hinweis darauf, daß schon in anderen Publikationen zuvor die gleiche „Operationalisierung" verwendet wurde, als Begründung für die Wahl der betreffenden Hilfshypothese ausreicht. Eine Diskussion denkbarer alternativer Hilfshypothesen und eine Begründung dafür, daß man diese Hilfshypothesen nicht gewählt hat, findet sich in der Fachliteratur nur selten. Dies begünstigt das unverbundene Nebeneinander verschiedener Hilfshypothesen zum gleichen theoretischen Konstrukt, wie es in unterschiedlichen Publikationen oftmals zu beobachten ist. Dieser „Operationalisierungspluralismus" charakterisiert große Teile der aktuellen experimentalpsychologischen Kognitionsforschung.

Wenn verschiedene Forscher unterschiedliche Hilfshypothesen zur „Operationalisierung" des gleichen theoretischen Konstrukts heranziehen, können sich prinzipiell zwei Situationen ergeben:

1) *Konvergierende Evidenz*: Unabhängig davon, welche der verschiedenen Hilfshypothesen ausgewählt wird, ergibt die empirische Prüfung der aus Theorie und Hilfshypothese abgeleiteten empirischen Hypothese immer das gleiche Ergebnis, sei es nun für die zugrundeliegende Theorie durchweg positiv oder durchweg negativ. In diesem Fall kann man sich entschließen, die Theorie als bewährt (bei positivem Ergebnis) oder als nichtbewährt (bei negativem Ergebnis) zu betrachten, ohne eine Antwort auf die Frage nach der Variablenvalidität zu geben. Offenbar hängt ja die empirische Bewertung der psychologischen Theorie nicht von der Wahl der Hilfshypothese ab (Bredenkamp, 1980). Zwei Probleme verbleiben allerdings. Zum einen ist das Vorgehen theoretisch nicht voll befriedigend, weil sich die empirische Bewertung auf den Theoriekern beschränkt und nicht ein zusammenhängendes System aus Theoriekern und verschiedenen Manifestationsregeln umfaßt. Man hat es also mit einer Art Provisorium zu tun, über das man von vornherein weiß, daß es in puncto Hilfshypothesen dringend der Revision bedarf. Oftmals ist es nur eine Frage der Zeit, bis dieses Provisorium auf-

grund inkonsistenter empirischer Resultate zusammenbricht (vgl. den folgenden Punkt 2). Zum zweiten — und dieses Problem erscheint mir noch gravierender — kann nicht ausgeschlossen werden, daß *alle* in Erwägung gezogenen Manifestationsregeln inadäquat sind, so daß sich eine Scheinbestätigung oder Scheinwiderlegung des Theoriekerns ergibt, obwohl alle Ergebnisse in eine einheitliche Richtung weisen.

2) *Widersprüchliche Evidenz*: Die empirische Bewertung der abgeleiteten Hypothesen hängt von der jeweils verwendeten Hilfshypothese ab. Ich habe an anderer Stelle vorgeschlagen, in diesem Fall von einem „Bedeutsamkeitsproblem" zu sprechen (Erdfelder, 1992a). Dies erscheint gerechtfertigt, wenn man sich etwas von der ursprünglichen Bedeutung dieses Begriffs in der Meßtheorie (s. z.B. Suppes & Zinnes, 1963) löst, etwa in Richtung des Konzepts der Bedeutsamkeit dritter Art (type-3 meaningfulness), das Niederée (1994) als Alternative zum Skalenniveaukonzept der Bedeutsamkeit vorgeschlagen hat (vgl. auch Niederée & Mausfeld, 1996). „Wissenschaftlich bedeutsam" oder „wissenschaftlich sinnvoll" wären dann nur solche Aussagen zu nennen, deren Wahrheitswert nicht von willkürlichen methodischen Entscheidungen wie etwa der Wahl der Manifestationsregel aus einer Menge „zulässiger" Manifestationsregeln abhängen. Denn wenn sich Bewährungs- oder Nichtbewährungsurteile über eine psychologische Theorie mit den jeweils verwendeten Hilfshypothesen ändern, sind diese Evaluationsurteile in bezug auf die Theorie offensichtlich nicht bedeutsam, d.h. nicht mehr sinnvoll interpretierbar.

Offenbar können beide Probleme — das Bedeutsamkeitsproblem im Fall widersprüchlicher Evidenz und auch das Problem möglicherweise generell inadäquater Manifestationsregeln im Fall konvergierender Evidenz — nur dann einer befriedigenden Lösung zugeführt werden, wenn empirische Kriterien existieren, die eine *objektive* Unterscheidung inadäquater und adäquater (oder möglicherweise adäquater) Manifestationsregeln erlauben. Solche Kriterien erlauben die *Elimination inadäquater Manifestationsregeln*, schränken damit die Willkür bei der Operationalisierung ein und machen das Bedeutsamkeitsproblem im günstigsten Fall sogar gegenstandslos. Praktisch läuft das Vorgehen darauf hinaus, ein Meßmodell aufzustellen, das festlegt, welche empirisch testbaren Bedingungen ein gültiger (valider) Indikator des betreffenden theoretischen Konstrukts erfüllen muß. Anders als im Falle der oben angesprochenen „Miniaturtheorien" ergeben sich empirisch gehaltvolle Aussagen hierbei durch Einbettung einer Manifestationsregel in ein psychologisch begründetes Meßmodell, welches empirischen Daten Restriktionen auferlegt, und nicht durch „blinde", theorielose Verknüpfung mehr oder minder willkürlich zusammengestellter Hilfshypothesen. Dem entspricht die Forderung nach „Konstruktvalidität" wie sie von Cronbach und Meehl (1955) ur-

sprünglich für psychologische Tests formuliert wurde, jedoch problemlos auch auf andere psychologische Indikatoren anwendbar ist.

Man beachte, daß Meßmodelle bzw. deren Axiome weder explizite totale Definitionen der involvierten theoretischen Konstrukte (wie im Falle der operationalen Definition) noch bedingte Definitionen (wie im Falle eines bilateralen Reduktionssatzes) leisten. Häufig wurde der oftmals David Hilbert zugeschriebene (z.B. von Essler, 1982, S. 67), aber wohl von Moritz Schlick zuerst benutzte Begriff der „impliziten Definition" verwendet, um darauf zu verweisen, daß bestimmte Grundbegriffe von Theorien allein durch ihr Vorkommen in den Axiomen der Theorie „definiert" seien (Stegmüller, 1973a, S. 36). Dieser Fall liegt auch bei psychologischen Konstrukten vor, die über Meßmodelle eingeführt werden (Bredenkamp, 1998a, S. 35 ff). Allerdings hat das Konzept der „impliziten Definition" berechtigte Kritik erfahren; insbesondere ist nicht klar, wodurch sich implizit definierte Begriffe von nicht definierten Begriffen genau unterscheiden (Essler, 1982, S. 67; Stegmüller, 1973a, S. 36).

Eine alternative Einordnung erlaubt der Begriff der T-Theorizität, der in strukturalistischen Konzeptionen psychologischer Theorien gebräuchlich ist (Westermann, 1987a; Westmeyer, 1989; 1992). Grob läßt sich dieser Begriff mit den Worten Stegmüllers (1973a, S. 47) wie folgt umschreiben: „Theoretisch in bezug auf eine Theorie T sind genau diejenigen Größen oder Funktionen, deren Werte sich nicht berechnen lassen, ohne auf diese Theorie T selbst (genauer: auf die erfolgreich angewendete Theorie T) zurückzugreifen". Es erscheint angemessen, psychologische Konstrukte, die über ein Meßmodell T eingeführt werden, als T-theoretische Begriffe aufzufassen. Der Vorteil dieser Charakterisierung ist, daß gar nicht erst der Eindruck erweckt wird, als könne eine Definition geleistet werden. Zugleich wird deutlich, daß eine Verwendung des Konstrukts unabhängig vom zugeordneten Meßmodell sinnlos wäre: Ohne ein Meßmodell gibt es kein Konstrukt und ohne ein empirisch adäquates Meßmodell gibt es keine Aussagen über das fragliche Konstrukt.

Diese relativ abstrakten Ausführungen lassen sich besser verstehen, wenn man ein konkretes Beispiel für ein testbares psychologisches Meßmodell betrachtet. Ein in verschiedener Hinsicht lehrreiches Beispiel ist die inzwischen schon klassische Arbeit von Levine (1966), in der Manifestationsregeln für kognitive Prozesse beim Lernen von Konzepten vorgeschlagen wurden. Ziel von Levine (1966) war es, eine einfache Theorie der Begriffsidentifikation so zu testen, daß auf Introspektionsmethodik bzw. auf verbale Berichte vollständig verzichtet werden kann. Levine (1966) nahm an, daß Personen über einen „Hypothesenraum" verfügen, der alle denkbaren Hypothesen über einen unbekannten, zu identifizierenden Begriff beinhaltet. Sowohl der Hypothesenraum selbst als auch die darauf ablaufen-

den kognitiven Prozesse sind dabei als theoretische Konstrukte aufzufassen, die der Introspektion einer Person nicht notwendigerweise zugänglich sein müssen.

Zur Überprüfung seiner Theorie hat Levine (1966) das sogenannte Rezeptionsparadigma der Konzeptidentifikation verwendet. Hierbei präsentiert ein Versuchsleiter oder eine Versuchsleiterin (Vl) mögliche Exemplare eines gesuchten Begriffs — typischerweise Zeichen oder Wörter —, welche die Vp dann als zum Begriff gehörig (Instanzen) oder nicht zum Begriff gehörig (Nichtinstanzen) klassifizieren muß (vgl. Bredenkamp & Wippich, 1977a, Kapitel 6). Levines Theorie besagt, daß die Vp zunächst zufällig eine Hypothese aus dem Hypothesenraum zieht und die Frage des Vl dann so beantwortet, als würde die betreffende Hypothese dem gesuchten Begriff entsprechen. Bekommt die Vp anschließend vom Vl keine negative Rückmeldung, behält die Vp die Hypothese bei und legt sie dem Urteil auf das nächste vom Vl präsentierte Item ebenfalls zugrunde. Nur dann, wenn die Vp eine negative Rückmeldung bekommt, zieht sie eine neue Hypothese aus dem Hypothesenraum. Diese wird dann bei ihrem folgenden Klassifikationsurteil verwendet, usw. Der Lernprozeß endet, wenn die Vp irgendwann einmal die zutreffende Hypothese aus dem Hypothesenraum zieht und folglich nur noch positive Rückmeldungen bekommt.

Ein scheinbar naheliegendes Vorgehen beim Überprüfen der Theorie könnte darin bestehen, die Vp in jedem Versuchsdurchgang danach zu fragen, welche Hypothese sie gerade testet. Aber dies wollte Levine (1966) aus drei Gründen vermeiden. Erstens gab es empirische Hinweise darauf, daß das beobachtbare Klassifikationsverhalten von Vpn einerseits und deren Verbalisierungen der dem Verhalten zugrundeliegenden Regeln andererseits auseinander klaffen können, so daß sich die Frage der Validität des verbalen Berichts stellt (vgl. etwa Verplanck, 1962; Wilson, 1975). Da der Hypothesenraum und die darauf ablaufenden verhaltenssteuernden Prozesse nach Levines Theorie nicht notwendig bewußt sein müssen, sind derartige Dissoziationen durchaus zu erwarten.[7] Zweitens kann die Aufforderung des Vl, die verhaltenssteuernde Konzepthypothese zu benennen, ein hypothesenzentriertes Vorgehen der Vp bei der Konzeptidenfikation erst künstlich induzieren, während ohne diese Instruktion vielleicht ganz anders geartete Prozesse ab-

[7]Dies widerspricht natürlich der Intuition der Vp, die fast immer der Meinung ist, die verhaltenssteuernde Regel genau benennen zu können. Gopnik (1993) würde dies die „Illusion der direkten Wahrnehmung" nennen. Wie schon erwähnt: Inkonsistenzen zwischen Verhalten und verbalen Berichten sind in der Psychologie keine Seltenheit. Eine nahezu perfekte Analogie zum Antagonismus von verbal berichteten Hypothesen, die dem Bewußtsein zugänglich sind, und verhaltenssteuernden Hypothesen, die unbewußt bleiben, bildet beispielsweise die Unterscheidung von bewußt wahrgenommener Distanz und unbewußt registrierter Distanz in der Psychologie der Raumwahrnehmung. Diese Unterscheidung ist u.a. nötig, um im Rahmen der sogenannten Größen-Distanz-Invarianzhypothese erklären zu können, daß der Mond am Horizont einerseits größer und andererseits näher erscheint als im Zenit (vgl. z.B. Kaufman, 1979, S. 322-332). Zu den neurobiologischen Mechanismen der Dissoziation von erlebter Intentionalität und faktischem Verhalten vgl. z.B. Roth (1997).

gelaufen wären. Tatsächlich gibt es zumindest bei Grundschulkindern Hinweise darauf, daß der Konzeptidentifikationsprozeß beschleunigt wird, wenn man die jeweils verwendeten Hypothese erfragt (Phillips & Levine, 1975). Drittens ist die Befragungstechnik auf sehr kleine Kinder und Tiere nicht anwendbar, aber auch für diese Populationen sollte die Theorie prinzipiell prüfbar sein.

Als Alternative zum verbalen Bericht der Vp konstruierte Levine (1966) deshalb die „blank-trial procedure". Hierbei wird der Hypothesenraum **H** über das zu identifizierende Konzept a priori auf acht Hypothesen h_i, i = 1, ... 8, begrenzt, die der Vp bekannt sind. Die Vp weiß also, daß einer der acht angegebenen Begriffe der gesuchte Begriff ist. Sie bekommt dann im Verlauf des Experiments mehrmals die gleichen vier Itempaare (Buchstabenpaare, siehe Kopfzeile in Tabelle 2.1) vorgelegt, von denen sie jeweils ein Item als zum gesuchten Begriff gehörig auswählen soll (two-alternative forced choice test). Da die Vp für jedes der vier Itempaare zwei Wahlen treffen kann, hat sie insgesamt 2^4 = 16 Antwortmöglichkeiten r_j, j = 1, ... 16. Diese 16 Antwortmöglichkeiten bilden den Verhaltensraum **R** der Vp. Die vier Itempaare sind dabei so gewählt, daß jeder Hypothese $h_i \in \mathbf{H}$ ein bestimmtes Antwortmuster $r_{ij} \in \mathbf{R}$ zugeordnet ist, sofern die Vp bei ihren vier Antworten konsistent die gleiche Hypothese h_i zugrunde legt. Levine (1966) vermied es, den Vpn über die vier blank trials hinweg Rückmeldungen zu ihren Wahlen zu geben. Nach den Annahmen der Theorie müssen die Vpn folglich die anfangs getestete Hypothese über die blank trials hinweg beibehalten.

Tabelle 2.1:
Zuordnung von Antwortmustern zu den acht Hypothesen bei der Blank-Trial-Prozedur von Levine (1966).

Hypothese der Vp	Zu beurteilendes Itempaar			
	X T	x **T**	T X	T x
SCHWARZ	links	rechts	links	rechts
WEISS	rechts	links	rechts	links
LINKS	links	links	links	links
RECHTS	rechts	rechts	rechts	rechts
GROSS	links	rechts	rechts	links
KLEIN	rechts	links	links	rechts
X	links	links	rechts	rechts
T	rechts	rechts	links	links

Tabelle 2.1 veranschaulicht das Levinesche Meßmodell für die acht verschiedenen Hypothesen. Angegeben ist in jeder Zeile, welches Antwortmuster über die vier Itempaare hinweg resultieren muß, wenn die Vp die betreffende Hypothese $h_i \in \mathbf{H}$ (linke Spalte der Tabelle) konsistent verwendet. Die Vp wählt pro

Itempaar (Kopfzeile der Tabelle) jeweils das linke (l) oder das rechte (r) Zeichen als zum Begriff gehörig aus, d.h. jede einzelne Antwort ist ein Element der Menge $\mathbf{A} = \{l, r\}$. Der Verhaltensraum kann daher als Produktmenge $\mathbf{R} = \mathbf{A}^4 = \mathbf{A} \times \mathbf{A} \times \mathbf{A} \times \mathbf{A} = \{ (l, l, l, l), (l, l, l, r), ..., (r, r, r, r)\}$ dargestellt werden.

Formal entspricht das Meßmodell einem Tripel $<\mathbf{H}, \mathbf{R}, f>$ bestehend aus den Mengen **H** und **R** sowie einer Abbildung $f: \mathbf{H} \rightarrow \mathbf{R}$, die jedem $h_i \in \mathbf{H}$ genau ein $r_{ij} \in \mathbf{R}$ zuordnet. Diese Abbildung hat zwei wichtige Eigenschaften. Zunächst einmal ist *f nicht surjektiv*, d.h. es gibt Antwortmuster $r_{ij} \in \mathbf{R}$, auf die keine der acht Hypothesen abgebildet wird. Dies trifft auf genau acht Antwortmuster zu, denn der gesamte Verhaltensraum besteht aus 16 Elementen, während nach Tabelle 2.1 nur acht Antwortmuster resultieren dürfen. Dieser Sachverhalt macht das System empirisch widerlegbar: Sei $f(\mathbf{H}) \subset \mathbf{R}$ die Bildmenge von **H** unter der Abbildung *f*, d.h. die Menge aller zulässigen (d.h. in Tabelle 2.1 aufgeführten) Antwortmuster, und sei $\mathbf{R} \setminus f(\mathbf{H})$ gerade die Komplementmenge zu $f(\mathbf{H})$ bezüglich **R**, so falsifizieren offenbar alle Antwortmuster $r_{ij} \in \mathbf{R} \setminus f(\mathbf{H})$ das Meßmodell. In diesem Fall ist irgendeine der Annahmen, die dem System zugrunde liegen, verletzt. Möglicherweise haben die Vpn zwar Hypothesen getestet, diese aber über die vier blank trials hinweg nicht konstant beibehalten oder Flüchtigkeitsfehler gemacht, möglicherweise haben sie auch gar keine Hypothesen getestet, sondern irgendeine andere Strategie verfolgt.

Man beachte den wesentlichen Unterschied zu bilateralen Reduktionssatzsystemen: Das Meßmodell zur Blank-Trial-Prozedur ist aufgrund der zugrundeliegenden theoretischen Annahmen empirisch gehaltvoll, nicht aufgrund multipler Operationalisierungen des gleichen theoretischen Konstrukts. Deshalb muß ein Scheitern der abgeleiteten Prognosen zwingend auf das Meßmodell selbst bezogen werden. Es kann nicht durch Hinweis auf „einige ungünstig gewählte Operationalisierungen" exhauriert werden, weil die Restriktionen, die das Modell empirischen Daten auferlegt, integraler Bestandteil der Bedeutung der involvierten theoretischen Konstrukte sind. Die Konstrukte selbst, die Theorie zu diesen Konstrukten und deren empirischen Indikatoren bilden also eine untrennbare Einheit; sie sind nicht mehr — wie im traditionellen Operationalisierungsansatz — voneinander separiert. Diese Eigenschaft ist für alle Meßmodelle charakteristisch, die in dieser Arbeit behandelt werden. Sie entschärft das Bedeutsamkeitsproblem ganz wesentlich, weil sie die Beliebigkeit bei der Wahl der Operationalisierungen deutlich einschränkt. Allerdings sei schon jetzt darauf hingewiesen, daß das Bedeutsamkeitsproblem auf diese Weise nicht zwangsläufig komplett beseitigt wird. Es besteht nämlich prinzipiell die Möglichkeit, daß unterschiedliche Meßmodelle immer gleichermaßen gut (oder schlecht) durch die Daten bestätigt werden. Diesem Problem der Datenäquivalenz und möglichen Lösungsansätzen dazu werden wir uns in Abschnitt 4.4 wieder zuwenden.

Die zweite wichtige Eigenschaft der Abbildung *f* ist ihre *Injektivität*, d.h. unterschiedliche Hypothesen aus **H** werden immer auf unterschiedliche Antwortmuster aus **R** abgebildet, so daß $\forall h_i, h_i' \in \mathbf{H}$: $f(h_i) = f(h_i') \Rightarrow h_i = h_i'$. Folglich existiert eine inverse Abbildung f^{-1}: $f(\mathbf{H}) \rightarrow \mathbf{H}$, die jedem zulässigen Antwortmuster aus **R** die zugrundeliegende Begriffshypothese eindeutig zuordnet. Diese Eigenschaft sichert die *Identifizierbarkeit* des Modells, d.h. die Möglichkeit, im Falle der Modellgültigkeit bei jedem Antwortmuster auf die zugrundeliegende Begriffshypothese rückschließen zu können. Dies ist im gegebenen Kontext wichtig, um weitere Überprüfungen der Begriffsidentifikationstheorie von Levine (1966) vornehmen zu können, etwa die Überprüfung der Annahme, daß eine Hypothese nach positiver Rückmeldung immer beibehalten, nach negativer Rückmeldung dagegen durch eine neue Hypothese ersetzt wird.

Levine (1966) fand empirisch folgendes:

1) Insgesamt 92.4% der von den Vpn in den blank trials erzeugten Antwortmuster entsprachen den nach dem Meßmodell zulässigen Antwortmustern aus *f*(**H**).
2) Folgt auf eine Folge von vier blank trials eine positive Rückmeldung des Vl und dann sofort eine erneute Sequenz von vier blank trials, so gleichen sich die Antwortmuster in den beiden Sequenzen in 95% aller Fälle.
3) Ist die Rückmeldung zwischen zwei Blank-Trial-Sequenzen dagegen negativ, so unterscheidet sich das zweite Antwortmuster vom ersten in 98% aller Fälle.
4) Ist die erste Rückmeldung negativ, so folgt in 87.3% der Fälle ein Blank-Trial-Antwortmuster, das einer der vier Hypothesen entspricht, die mit der ersten Rückmeldung prinzipiell vereinbar sind.

Levine (1966) sah in diesem Ergebnis eine Bestätigung seines Meßmodells und darüber hinaus auch der Annahme, daß negatives Feedback zur Aussonderung der widerlegten Hypothese aus dem Hypothesenraum führt. Darüber hinaus scheinen sogar einige nicht getestete Hypothesen eliminiert zu werden, sofern sie mit den bisherigen Rückmeldungen des Vl inkompatibel sind. Dies steht im Gegensatz zu Annahmen anderer Autoren (vgl. z.B. Bower & Trabasso, 1964; Restle, 1962), die davon ausgingen, daß der Hypothesenraum immer gleich groß bleibt. Nach negativem Feedback würde eine Hypothese demnach einfach in den Hypothesenraum „zurückgelegt" und somit möglicherweise später wieder gezogen. Träfe diese Auffassung zu, so hätte negatives Feedback nicht in 98% der Fälle, sondern nur in etwa 87.5% der Fälle (nämlich 100%–12.5%) zu einem veränderten Antwortmuster führen dürfen.

Obwohl klar ist, daß die Ergebnisse besser zu Levines Theorie als zu anderen Theorien der Konzeptidentifikation passen, wird anhand der Befunde doch ein weiteres Problem deutlich: Die bislang besprochenen Ansätze zur Lösung des

Theorie-Empirie-Überbrückungsproblems sind deterministischer Natur und lassen Abweichungen von den exakten Prognosen eigentlich nicht zu. Streng genommen müßte deshalb auch das geschilderte Modell verworfen werden, denn es impliziert, daß (a) 100% der Antwortmuster in die Menge $f(\mathbf{H})$ fallen, (b) 100% der Antwortmuster nach positiver und (c) 0% nach negativer Rückmeldung unverändert bleiben. Levine hat dieses Problem deutlich gesehen und deshalb das deterministische Modell um die Annahme ergänzt, daß Vpn mit einer für alle Urteile und Itempaare gleichen Wahrscheinlichkeit „by accident" Flüchtigkeitsfehler machen: „In practice, this probability is very small (of the order of .02) and this assumption may, with little distortion to the results, be ignored" (Levine, 1966, S. 332). Die bloße Tatsache, daß die beobachteten Abweichungen von den Prognosen prozentual gesehen vernachlässigbar erscheinen, besagt aber nicht, daß sie irrelevant für die Bewertung des Meßmodells sind. Im nächsten Abschnitt wird diskutiert, ob bzw. wie mit diesem Problem in formal befriedigender Weise umgegangen werden kann.

2.5 Das Fehlerproblem

Prinzipiell gibt es zwei Wege, mit mutmaßlich fehlerbelasteten Daten zu verfahren. Einerseits kann man versuchen, die Präzision der empirischen Erhebung zu verbessern, so daß zufällige Fehler eliminiert oder zumindest reduziert werden (Bredenkamp, 1969). Andererseits kann man die Fehlerproblematik mit in das betreffende Meßmodell aufnehmen, so daß die Anerkennung gewisser Abweichungen von den deterministischen Prognosen nicht notwendig zur Modellfalsifikation führen muß. Leider kann die Fehlerproblematik in der Psychologie durch Verfeinerung der Untersuchungs- und Beobachtungstechniken nur sehr eingeschränkt reduziert werden, weil auch unter künstlichen experimentellen Bedingungen immer noch zu viele unkontrollierte Variablen auf menschliches Verhalten Einfluß nehmen. Wie man aus der gegenwärtig dominierenden Forschungspraxis schließen muß, favorisiert deshalb die große Mehrheit empirisch arbeitender Psychologinnen und Psychologen den zweiten Weg. Offenkundig hat man hierbei aber ein neues Problem: das Entscheidbarkeitsproblem (Tack, 1971), d.h. das Problem der Festlegung begründeter empirischer Kriterien für die Annahme oder Verwerfung einer nichtdeterministischen Hypothese.

Auch diesbezüglich kann man grob zwischen zwei Lösungsansätzen unterscheiden, die in der Forschungspraxis Verwendung fanden und immer noch finden. Beiden ist gemeinsam, daß über psychologische Annahmen mittels statistischer Hypothesentests entschieden wird. Ein Weg besteht darin, einem „Zufallsmodell", das durch eine statistische Nullhypothese (H_0) beschrieben wird, eine statistische Alternativhypothese (H_1) gegenüberzustellen, die unter gewissen Zusatz-

annahmen aus der interessierenden substanzwissenschaftlichen Hypothese ableitbar ist. Typisch ist dieses Vorgehen z.B. im Rahmen der von Hildebrand, Laing und Rosenthal (1977) vorgeschlagenen Prädiktionsanalyse. Angewandt auf die Untersuchung von Levine (1966) wäre das Nullhypothesenmodell der Prädiktionsanalyse das Modell der totalen stochastischen Unabhängigkeit für die vierdimensionale 2×2×2×2-Kontingenztafel der möglichen Antwortmuster. Die Reaktion auf ein bestimmtes Itempaar wäre nach diesem Modell also komplett unabhängig davon, welche Antworten man auf die anderen Itempaare gibt. Die Alternativhypothese behauptet demgegenüber, daß die hypothesenkonformen Zellen („Erfolgszellen") der Kontingenztafel — d.h. die Elemente von $f(\mathbf{H})$ — im Vergleich zum Unabhängigkeitsmodell überfrequentiert sind. Man prüft dann anhand der von Rosenthal et al. (1977) vorgeschlagenen „Del"-Statistik (∇), ob die empirische Frequenz der Antwortmuster in den durch Tabelle 2.1 beschriebenen „Erfolgszellen" relativ zu einem bestimmten α-Fehler-Niveau signifikant *größer* ist als nach dem Unabhängigkeitsmodell zu erwarten.[8] Hierbei wird ∇ zumeist in eine unter der Nullhypothese asymptotisch standardnormalverteilte Statistik transformiert, wobei multinomialverteilte Zellhäufigkeiten vorausgesetzt werden. Ist diese transformierte Statistik größer als der kritische z-Wert zum Niveau α, so betrachtet man die zugrundeliegende Hypothese als bewährt. Insignifikante Resultate „belasten" die zugrundeliegende psychologische Hypothese.

Die nicht nur im Rahmen der Prädiktionsanalyse übliche Strategie, psychologische Hypothesen und Modelle mit einer zusammengesetzten H_1 zu verknüpfen, die dann einem bestimmten „Zufallsmodell" als H_0 gegenübergestellt wird, ist nicht ohne Kritik geblieben. So hat bereits Meehl (1967) auf das methodologische Paradox hingewiesen, daß es auf diese Weise psychologischen Hypothesen um so leichter gemacht wird, sich empirisch zu bewähren, je größer der Stichprobenumfang und damit die Genauigkeit der statistischen Aussage wird — ganz im Gegensatz etwa zu den Verhältnissen in der Physik, wo die substanzwissenschaftlichen Hypothesen in der Regel statistischen Nullhypothesen entsprechen. Zwar läßt sich dieses Problem durch simultane Kontrolle der α- und β-Fehlerwahrscheinlichkeiten beim statistischen Hypothesentest entschärfen (Bredenkamp, 1972; 1980), aber es bleibt immer noch das Problem, daß das Ergebnis der statistischen Entscheidung wesentlich von der Wahl der Nullhypothese abhängt, gegen die man die psychologische Hypothese kontrastiert. Über die adäquate Wahl des Nullhypothesenmodells sagt die zu prüfende Hypothese meistens nichts aus, so daß man auf Plausibi-

[8] In die Berechnung von ∇ gehen allerdings nur die hypothesenwidrigen Antwortmuster in den sogenannten Fehlerzellen ein. Liegen keine hypothesenwidrigen Antwortmuster vor, nimmt ∇ den Maximalwert 1 an. ∇ ist genau dann größer als null, wenn in den Fehlerzellen insgesamt weniger Antwortmuster als nach dem Unabhängigkeitsmodell zu erwarten liegen. Dies ist natürlich genau dann der Fall, wenn die Anzahl der Antwortmuster in den Erfolgszellen größer ist als nach dem Unabhängigkeitsmodell zu erwarten.

litätserwägungen angewiesen ist, die denen bei der Wahl von adäquaten Operationalisierungen zu einem theoretischen Konstrukt sehr ähnlich sind. Im Regelfall können mehrere denkbare Nullhypothesenmodelle in Erwägung gezogen werden, die häufig auch noch logisch völlig unverbunden sind. Man hat es somit mit einem weiteren Bedeutsamkeitsproblem dritter Art (Niederée, 1994) zu tun, nämlich dem, daß es an einer stringenten logischen Beziehung zwischen der interessierenden substanzwissenschaftlichen Hypothese und der statistischen Nullhypothese, die letztlich überprüft wird, mangelt. Gerade bei der prädiktionsanalytischen Prüfung psychologischer Hypothesen ist dieses Problem unverkennbar: Es wird nicht deutlich, warum man gerade das von Hildebrandt et al. favorisierte Unabhängigkeitsmodell routinemäßig als maßgebliches Nullhypothesenmodell der Theorieevaluation auszeichnen sollte (Erdfelder, 1991; Goodman & Kruskal, 1974a; 1974b).[9]

Als Alternative empfiehlt es sich, die zu prüfende substanzwissenschaftliche Hypothese nicht mit der Alternativhypothese, sondern mit der Nullhypothese eines statistischen Modellgeltungstests zu verknüpfen. Auf diese Weise kann eine logische Verbindung zwischen substanzwissenschaftlicher Hypothese und Fehlertheorie auf der einen Seite sowie geprüfter statistischer Hypothese auf der anderen Seite im allgemeinen leicht hergestellt werden. Ist — wie im Falle des Ansatzes von Levine (1966) — das psychologische Meßmodell bzw. die zu prüfende Theorie im Kern deterministisch, die Beziehung zwischen theoretischen Konstrukten und ihren (kategorialen) empirischen Indikatoren dagegen mutmaßlich fehlerbelastet, so eignet sich das auf Paul Lazarsfeld (z.B. Lazarsfeld & Henry, 1968) zurückgehende Latent-Class- bzw. Latent-Structure-Modell hervorragend zur Formalisierung und statistischen Überprüfung der zugrundeliegenden Annahmen (vgl. Erdfelder, 1990a; 1991; Rudinger, Chaselon, Zimmermann & Henning, 1985)[10]. Das deterministische Modell beschreibt im Rahmen eines solchen Ansatzes die latente Struktur „hinter" den beobachteten Daten völlig korrekt, d.h. in der latenten 2×2×2×2-Kontingenztafel der „wahren" Antwortmuster sind im Falle der Untersuchung von Levine (1966) wirklich nur die acht Zellen besetzt, die $f(\mathbf{H})$ entsprechen. Durch Fehlerprozesse, die z.B. durch Aufmerksamkeitsschwankungen der Vp, unreliable Meßinstrumente oder diverse Beobachtungs- oder Datenregistrierungsfeh-

[9]Am Rande sei erwähnt, daß sich die Situation in der Konfigurationsfrequenzanalyse (Krauth & Lienert, 1995; Lienert, 1988; von Eye, 1990), welche in der Regel ebenfalls auf dem Nullhypothesenmodell der totalen stochastischen Unabhängigkeit beruht, grundsätzlich anders darstellt: Hier richtet sich das Interesse nicht primär auf Theorieevaluation, sondern typischerweise auf die Beantwortung der Frage „Welche Zellen der Kontingenztafel sind relativ zum Unabhängigkeitsmodell überfrequentiert bzw. unterfrequentiert?" Offensichtlich ist das Nullhypothesenmodell hierbei adäquat, da durch die Fragestellung impliziert. Das Problem ist also nicht das Unabhängigkeitsmodell per se, sondern die Art seiner Verwendung in der Prädiktionsanalyse (Goodman & Kruskal, 1974a).

[10]Wenn die beobachteten Variablen nicht kategorial, sondern kontinuierlich sind, läßt sich das gleiche Ziel mit diversen finiten Mischverteilungsmodellen für stetige Zufallsvariablen (ZVn) oder Zufallsvektoren erreichen (Erdfelder, 1990a; Rost & Erdfelder, 1996; Sixtl, 1993).

ler des Vl hervorgerufen sein können, kommt es jedoch zu Abweichungen zwischen latenter und manifester Kontingenztafel, so daß in der beobachteten Tafel Zellen besetzt sein können, die nicht zu $f(\mathbf{H})$ gehören. Konkret könnte man für den Datensatz von Levine (1966) beispielsweise ein latentes Klassenmodell (LCA-Modell) mit acht latenten Klassen ansetzen, für jede denkbare Begriffshypothese eine Klasse. Die unbekannte Wahrscheinlichkeit, daß die i-te Hypothese h_i, $i = 1, ..., 8$, aus dem Hypothesenraum gezogen wird, sei mit $p(h_i)$ bezeichnet. Da die acht Hypothesen den Hypothesenraum ausschöpfen, muß für die Klassenwahrscheinlichkeiten gelten:

$$\sum_{i=1}^{8} p(h_i) = 1 \,. \tag{2.5-1}$$

Für jede der acht Hypothesen werden die Antworten auf die vier Itempaare der blank trials als unkorreliert angenommen, d.h. die Fehlerprozesse sollen innerhalb der latenten Klassen lokal stochastisch unabhängig sein. Bezeichnet $p(k_j)_{j\,|\,i}$ die Wahrscheinlichkeit der Antwort $k_j \in \mathbf{A}$ beim Itempaar j ($j = 1, ..., 4$) im Falle des Vorliegens der Hypothese h_i, so ergeben sich die Wahrscheinlichkeiten der 2^4 möglichen Antwortmuster zu einer gegeben Hypothese einfach durch Produktbildung der entsprechenden Antwortwahrscheinlichkeiten pro Itempaar. Für ein beliebiges Antwortmuster (k_1, k_2, k_3, k_4), $k_j \in \mathbf{A}$, resultiert bei Vorliegen der i-ten Hypothese daher die Wahrscheinlichkeit

$$p((k_1, k_1, k_3, k_4)|h_i) = p(k_1)_{1|i} \cdot p(k_2)_{2|i} \cdot p(k_3)_{3|i} \cdot p(k_4)_{4|i} = \prod_{j=1}^{4} p(k_j)_{j|i}. \tag{2.5-2}$$

Die unbedingte Wahrscheinlichkeit des Antwortmusters (k_1, k_2, k_3, k_4), ergibt sich dann einfach durch Summierung der mit den Klassenwahrscheinlichkeiten gewichteten Terme gemäß Gleichung (2.5-2):

$$p(k_1, k_2, k_3, k_4) = \sum_{i=1}^{8} p(h_i) \cdot \prod_{j=1}^{4} p(k_j)_{j|i}. \tag{2.5-3}$$

Das durch Gleichung (2.5-3) beschriebene Modell entspricht dem unrestringierten 8-Klassen-LCA-Modell für die 2^4-Kontingenztafel (vgl. Langeheine & Rost, 1996). Damit die bedingten Antwortwahrscheinlichkeiten $p(k_j)_{j\,|\,i}$ im Rahmen eines Fehlermodells sinnvoll interpretierbar sind, müssen sie geeignet restringiert werden. Naheliegend ist z.B. die Annahme, daß beim j-ten Itempaar mit Wahrscheinlichkeit a_j ein Antwortfehler gemacht wird, wenn die wahre Antwort „links" lautet, und mit Wahrscheinlichkeit b_j ein Fehler resultiert, wenn die wahre Antwort „rechts" lautet. (Unter einer „wahren Antwort" ist hierbei jeweils die theoriekonforme Antwort zu verstehen wie sie sich in der latenten Kontingenztafel nieder-

schlagen sollte.) Die aus dieser Annahme resultierenden Restriktionen für die Wahrscheinlichkeiten von „Rechts"-Antworten, $p(\{r\})_{j\,|\,i}$, sind in Tabelle 2.2 zusammengestellt. Die komplementären Wahrscheinlichkeiten für „Links"-Antworten ergeben sich einfach aus der Gleichung $p(\{l\})_{j\,|\,i} = 1 - p(\{r\})_{j\,|\,i}$. Damit die a_j- und b_j-Parameter sinnvoll als Antwortfehlerwahrscheinlichkeit interpretierbar sind, wird man nicht nur verlangen, daß das Modell gemäß Tabelle 2.2 zu vorliegenden Stichprobendaten paßt, sondern darüber hinaus, daß die a_j- und b_j-Parameterschätzungen zu „kleinen" Werten führen (z.B. $\forall\, j\colon \hat{a}_j \leq .10 \wedge \hat{b}_j \leq .10$).

Tabelle 2.2:
Parameter eines 8-Klassen-Latent-Class-Modells zur Blank-Trial-Prozedur von Levine (1966).

Hypothese der Vp	Klassen-größe	Wahrscheinlichkeit der Wahl des rechten Items			
		X T	x **T**	T $\mathbb{X}$	$\mathbb{T}$ x
SCHWARZ	$p(h_1)$	a_1	$(1 - b_2)$	a_3	$(1 - b_4)$
WEISS	$p(h_2)$	$(1 - b_1)$	a_2	$(1 - b_3)$	a_4
LINKS	$p(h_3)$	a_1	a_2	a_3	a_4
RECHTS	$p(h_4)$	$(1 - b_1)$	$(1 - b_2)$	$(1 - b_3)$	$(1 - b_4)$
GROSS	$p(h_5)$	a_1	$(1 - b_2)$	$(1 - b_3)$	a_4
KLEIN	$p(h_6)$	$(1 - b_1)$	a_2	a_3	$(1 - b_4)$
X	$p(h_7)$	a_1	a_2	$(1 - b_3)$	$(1 - b_4)$
T	$p(h_8)$	$(1 - b_1)$	$(1 - b_2)$	a_3	a_4

Das vorgeschlagene probabilistische Meßmodell umfaßt insgesamt 15 freie Parameter, nämlich vier *a*-Parameter, vier *b*-Parameter und sieben Klassenwahrscheinlichkeiten $p(h_i)$; die achte Klassenwahrscheinlichkeit ist kein freier Parameter, da sich die Wahrscheinlichkeiten gemäß Gleichung (2.5-1) zu 1 aufaddieren müssen. Selbstverständlich kann man daran denken, das Modell noch weiter zu restringieren, indem man z.B. die *a*- und *b*-Parameter pro Itempaar ($\forall\, j\colon a_j = b_j$) oder über alle Itempaare hinweg gleichsetzt ($a_1 = a_2 = a_3 = a_4 = a$ und $b_1 = b_2 = b_3 = b_4 = b$). Nimmt man beide Gleichsetzungsrestriktionen simultan vor, so entspricht das dem von Levine (1966) selbst vorgeschlagenen Fehlermodell mit nur noch einem Fehlerparameter $a = b = x$. Durch die extremste denkbare Restriktion,

$$\forall\, j\colon a_j = b_j = x = 0, \tag{2.5-4}$$

geht das Modell schließlich in seinen Grenzfall, das deterministische Originalmodell von Levine (1966; vgl. letzter Abschnitt), über.[11]

[11]Diese Aussage ist mathematisch unsauber, da ein probabilistisches Modell auch nach Fixierung der Wahrscheinlichkeiten auf Extremwerte ein probabilistisches Modell bleibt. Es erlaubt keine „sicheren", sondern nur „fast sichere" Aussagen und kann damit nicht mit einem deterministischen

Levine (1966) selbst hat keine statistische Analyse des probabilistischen Meßmodells vorgenommen. Er hat vielmehr unter Voraussetzung der Modellgültigkeit und unter der weitergehenden Annahme generell gleicher Fehlerwahrscheinlichkeiten $a_j = b_j = x$ geschätzt, wie groß x im Falle seiner Daten in etwa anzusetzen ist. Durch Summieren der vom Modell vorhergesagten Wahrscheinlichkeiten für die acht Fehlerzellen des Antwortraumes über alle latenten Klassen hinweg ergibt sich, daß die Wahrscheinlichkeit eines nicht hypothesengemäßen Antwortmusters

$$p(\mathrm{R} \setminus f(\mathrm{H})) = 4 \cdot (x^3 \cdot (1-x) + x \cdot (1-x)^3) \qquad (2.5\text{-}5)$$

betragen muß (vgl. dazu die Modellgleichungen in Abschnitt 5 des Anhangs). Setzt man die empirisch beobachtete relative Fehlerhäufigkeit 1–.924 = .076 (s. S. 20, Punkt 1) für $p(\mathbf{R} \setminus f(\mathbf{H}))$ in Gleichung (2.5-5) ein und löst dann nach x auf, so erhält man die Schätzung $\hat{x}$ = .0202. Dieses Ergebnis wertete Levine als Bestätigung seiner These, daß Fehlerprozesse weitgehend vernachlässigbar sind.

Unüberprüft blieb hierbei allerdings, ob alle in den Schätzwert eingehenden Annahmen zutreffen. Tatsächlich könnte man eine solche Analyse leicht nachholen, wenn die Häufigkeiten der 16 möglichen Blank-Trial-Antwortmuster für die Studie von Levine (1966) verfügbar wären. Auf dem Hintergrund einer Multinomialverteilungsannahme für die Stichprobenhäufigkeiten der Antwortmuster erweist sich das LCA-Modell gemäß Gleichung (2.5-3) und Tabelle 2.2 nämlich als Spezialfall multinomialer Verarbeitungsbaummodelle (MVB-Modelle), die in den folgenden Kapiteln ausführlich diskutiert werden. Wie in Abschnitt 6.3 explizit gezeigt wird, kann das Modell im Prinzip mit den gleichen statistischen Techniken analysiert und evaluiert werden, die üblicherweise für MVB-Modelle Verwendung finden. Im gegebenen Falle scheitert das nur daran, daß Levine (1966) die Häufigkeiten der verschiedenen Antwortmuster nicht veröffentlicht hat. In Abschnitt 6.3 wird das Modell daher auf Daten einer Replikationsstudie zum Levine-Experiment angewandt, deren Ergebnisse — so viel sei jetzt schon vorweggenommen — im wesentlichen für die Adäquatheit des restriktiven Fehlermodells von Levine (1966) sprechen, auch wenn der Schätzwert von ca. $\hat{x}$ = .02 mit einem Fragezeichen versehen werden muß.

Zwei Einwände gegen die Erweiterung deterministischer psychologischer Meßmodelle zu probabilistischen Modellen, die mit der Nullhypothese eines statistischen Modellgeltungstests verknüpft und entsprechend statistisch evaluiert werden, scheinen auf der Hand zu liegen. Der erste Einwand lautet, daß eine logische Verbindung zwischen den deterministischen psychologischen Kernannah-

---

Modell gleichgesetzt werden (Tack, 1971). In bezug auf empirisch testbare Konsequenzen spielt diese Unterscheidung allerdings keine wesentliche Rolle.

men und der letztlich geprüften statistischen Hypothese nur um den Preis schlecht begründeter Zusatzannahmen möglich wird (z.B. Unabhängigkeit der Fehlerprozesse, Gleichheit von Fehlerwahrscheinlichkeiten). Der zweite Einwand betrifft das Entscheidbarkeitsproblem bei probabilistischen Modellen, das — so könnte man argumentieren — wegen der prinzipiellen Nichtfalsifizierbarkeit probabilistischer Aussagen weiterhin als ungelöst betrachtet werden muß (vgl. z.B. Tack, 1971).

Ich halte beide Einwände für nicht sehr gravierend. Auf den ersten Kritikpunkt kann man mit Ward Edwards (1954, S. 382) antworten: „The most useful thing to do with a theory is not to criticize its assumptions but rather to test its theorems". Solange die statistischen Annahmen zu den Fehlerprozessen den faktischen Verhältnissen wenigstens annähernd entsprechen, kann das Modell als Approximation an die Realität eine sinnvolle, ordnungsstiftende Funktion erfüllen. Es kommt in der kognitiven Psychologie im Regelfall nicht darauf an, eine punktgenaue Beschreibung empirischer Datenstrukturen zu leisten. Probabilistische Modelle haben vielmehr die Funktion, eine möglichst einfache und dennoch gute Annäherung an die empirische Datenstruktur zu liefern; sie müssen nicht zwingend in jedem technischen Detail auf vollständig korrekten Annahmen basieren: „Most substantive models are only approximations to reality and therefore are technically false" (Riefer & Batchelder, 1988, S. 325). Noch eindeutiger äußerte sich Batchelder (1991): „All statistical models are wrong in some details when applied to worldly settings. Of course some models are more accurate than others, and many models are sufficiently accurate for the purposes they were intended". Natürlich muß die Möglichkeit fehlspezifizierter Zusatzannahmen bei der Modellprüfung insofern im Blick behalten werden, als Prüfmethoden vermieden werden sollten, die sehr sensibel auf Verletzungen der statistischen Zusatzannahmen reagieren. Aber hierfür gibt es Mittel und Wege (vgl. z.B. Kapitel 5 der vorliegenden Arbeit).

Auf dem Hintergrund dieser Auffassung besteht das Entscheidbarkeitsproblem bei der empirischen Evaluation probabilistischer Modelle nur noch in dem Sinne fort, daß eine Falsifikation derartiger Modelle über klassische Basissätze im Popperschen (1934/1982) Sinne tatsächlich nicht möglich ist, weil probabilistische Modelle ihrem Wesen nach kein Stichprobenergebnis mit absoluter Sicherheit ausschließen. Allerdings kann diese Schwierigkeit im Rahmen eines methodologischen Falsifikationismus (Lakatos, 1970) durch Festlegung rational begründeter Kriterien bei der statistischen Entscheidung („rejection rules", vgl. Lakatos, 1970, S. 169) umgangen werden (Bredenkamp, 1972; 1980; Erdfelder & Bredenkamp, 1994; Gadenne, 1976; 1984; Hager, 1987; Westermann & Hager, 1986). Grundlage hierfür ist die Neyman-Pearson-Theorie statistischen Testens (einführend: Willmes, 1998). Hierbei kommt der statistischen Teststärkekontrolle (power analysis) besondere Bedeutung zu: Es ist einerseits wichtig, ein Modell „streng" zu testen, so daß substantielle, praktisch bedeutsame Verletzungen der Modellprognosen auch mit ho-

her Wahrscheinlichkeit statistische Entscheidungen gegen das Modell zur Folge haben. Andererseits muß das Modell aber auch „fair" getestet werden, so daß im Falle der (approximativen) Modellgültigkeit mit hoher Wahrscheinlichkeit Entscheidungen für das Modell gefällt werden. Techniken der Teststärkekontrolle, die eine in diesem Sinne strenge und zugleich faire Evaluation probabilistischer Modelle erlauben, liegen aber bereits weitgehend ausgearbeitet vor (Bredenkamp, 1980; Cohen, 1988; Erdfelder, Faul & Buchner, 1996; Hager, 1987). Sie sind anwendbar, sofern in Form eines sogenannten Effektstärkeparameters — bei multinomialen Modellen beispielsweise Cohens (1988) *w* — angegeben wird, wo man die Grenze zwischen „approximativer Gültigkeit eines Modells" und „substantieller Verletzung der Modellprognosen" genau ziehen will. Wir werden im vierten Kapitel hierauf zurückkommen.

Wenn durch Teststärkekontrollen sichergestellt ist, daß die Fehlerwahrscheinlichkeiten bei statistischen Entscheidungen gering sind, rechtfertigt das allerdings noch keinen Falsifikationsautomatismus: Ein Ergebnis, das für die Verwerfung einer Modellgeltungshypothese spricht, ist im Rahmen eines methodologischen Falsifikationismus lediglich eine notwendige, nicht jedoch eine hinreichende Bedingung für die Falsifikation der vorgeordneten psychologischen Hypothese (vgl. Bredenkamp, 1980, sowie Erdfelder & Bredenkamp, 1994, S. 609).

Da die hier nur kurz skizzierte Methodologie der Prüfung psychologischer Hypothesen über Tests statistischer Hypothesen jüngst kritisiert wurde (Westmeyer, 1998), möchte ich diesen Abschnitt nicht ohne eine Antwort auf diese Kritik abschließen. Westmeyers Kritik betrifft im wesentlichen vier Punkte. Zunächst sieht er in der hier vertretenen Methodologie „ein Amalgam aus wenigstens drei Ansätzen" (Neyman-Pearson-Theorie, Poppers Falsifikationstheorie und der Methodologie wissenschaftlicher Forschungsprogramme nach Lakatos), deren Kompatibilität „zumindest bezweifelt werden kann" (Westmeyer, 1998, S. 258). Diese Kritik steht klar im Widerspruch zur Position von Lakatos, der in dem „Amalgam" eine zwingende Konsequenz des methodologischen Falsifikationismus sieht:

> „Probabilistic theories may qualify now as 'scientific': although they are not falsifiable they can be easily made 'falsifiable' (...) by specifying certain rejection rules which may make statistically interpreted evidence 'inconsistent' with the probabilistic theory. (...) Indeed, this methodological falsificationism is the philosophical basis of some of the most interesting developments in modern statistics. The Neyman-Pearson approach rests completely on methodological falsificationism" (Lakatos, 1970, S. 109).

Die behauptete Inkompatibilität zwischen Lakatos sowie Neyman und Pearson ist auf dem Hintergrund dieses Zitats schwer nachzuvollziehen. Wenn eine Inkonsistenz wirklich vorliegen sollte, dann kann es sich nur um ein immanentes Pro-

blem des methodologischen Falsifikationismus handeln, nicht dagegen um ein Problem, das durch die Verbindung der Ansätze erst entsteht. Auch Popper (1934/1982, Abschnitt 68) hat — aller Polemik zwischen ihm und Lakatos zum Trotz — Ideen geäußert, die als Vorwegnahme der oben zitierten Position von Lakatos zu deuten sind:

> „Wir schließen deshalb die unbeschränkte Verwendung der Wahrscheinlichkeitsaussagen aus — durch den methodologischen Beschluß, Effekte (...) niemals auf gehäufte Zufälle zurückzuführen" (Popper, 1934/1982, S. 153).

Diese und weitere Textpassagen, die man nennen könnte, machen deutlich, daß die Verknüpfung einiger grundlegender Ideen von Popper, Lakatos sowie Neyman und Pearson — soweit sie sich auf die Verwendung von statistischen Hypothesen und die Entscheidung über statistische Hypothesen beziehen — nicht nur möglich ist, sondern auch naheliegend erscheint. Daß man sicher auch Aussagen finden kann, die auf Disharmonie speziell zwischen Popper und Lakatos schließen lassen, ändert an der prinzipiellen Vereinbarkeit der genannten Ideen nichts.

Westmeyer (1998, S. 261f) bezweifelt zweitens die Herstellbarkeit von Implikationsbeziehungen zwischen psychologischen Hypothesen (PHn) und statistischen Hypothesen (SHn). Es ist ausreichend, auf diesen Kritikpunkt mit einer Existenzbehauptung zu reagieren: Weiter oben in diesem Abschnitt wurde bereits ein implikativ verknüpftes PH-SH-Paar vorgestellt, nämlich das Latent-Class-Modell (SH) zur Konzeptidentifikationstheorie von Levine (PH). Die folgenden Kapitel nennen eine Reihe weiterer Beispiele, wobei die SH grundsätzlich die Modellgeltungshypothese eines stochastischen Modells ist. Diese Beispiele — neben vielen anderen — widerlegen die prinzipielle Unmöglichkeit einer logischen Brücke zwischen PH und SH. Richtig ist zweifellos, daß eine logische Verknüpfung nicht in jedem Fall und vor allem nicht immer problemlos gelingt[12]. Das schmälert jedoch nicht den Wert der Methodologie für die große (und immer größere werdende) Klasse von PHn, bei denen der skizzierte Weg beschritten werden kann.

Ein dritter Kritikpunkt Westmeyers (1998) betrifft die Ebene der Hypothesenprüfung. Überprüfungen von allgemeinpsychologischen Hypothesen anhand von SHn, die für Aggregate formuliert werden, hält er für „sinnlos" (Westmeyer, 1998, S. 261). Zu entgegnen ist, daß die hier vertretene Methodologie eine Hypothesenprüfung auf Aggregatebene keineswegs zwingend fordert. Im Gegenteil: Wo Überprüfungen allgemeinpsychologischer Hypothesen an einzelnen Individuen

[12]Iseler (1999) trifft die nützliche Unterscheidung zwischen vorstatistischen psychologischen Hypothesen und statistisch explizierten psychologischen Hypothesen (vgl. auch die Unterscheidung von psychologischen Hypothesen, psychologischen Vorhersagen und statistischen Vorhersagen bei Hager, 1992). Die Behauptung einer Implikationsbeziehung zwischen PH und SH ist selbstverständlich nur dann sinnvoll, wenn die Abkürzung „PH" für das Explikat (die statistisch explizierte psychologische Hypothese) und nicht für das Explikandum (die vorstatistische psychologische Hypothese) reserviert wird.

praktikabel sind und wo sie die Validität der Untersuchung und der statistischen Analyse nicht gefährden, sollten sie unbedingt durchgeführt werden (vgl. u.a. Bredenkamp, 1972, 1980, 1990; s. auch Kapitel 5.4). Wenn die individuelle Prüfung dagegen unter Praktikabilitäts- oder Validitätsgesichtspunkten nicht möglich ist — was u.a in der Kognitionspsychologie leider häufig der Fall ist (vgl. Abschnitt 5.4.1) — *muß* wohl oder übel auf die Aggregatebene übergegangen werden. Dieser Übergang ist dann nicht sinnlos, wenn die allgemeinpsychologische Hypothese eine Aggregathypothese logisch impliziert: Die Verwerfung der letzteren impliziert dann nämlich die Verwerfung der ersteren. Insofern genügt der Ansatz der grundlegenden Forderung, daß eine empirische Kritik psychologischer Hypothesen prinzipiell möglich sein muß (Herrmann, 1979). Daß ein analoger Schluß vom Aggregat auf das einzelne Individuum bei Verifikationsurteilen nicht gilt, ist unbestritten. Das Verifikationsproblem ist bekanntlich ein fundamentales Problem bei der Überprüfung allgemeiner Aussagen und für die hier vertretene Methodologie keineswegs spezifisch.

Der explizite Verzicht auf einen Falsifikationsautomatismus bei Ergebnissen, die gegen die aus einer PH abgeleitete SH sprechen, bildet den vierten Kritikpunkt Westmeyers (1998). Seiner Meinung nach ist ein solcher Verzicht auf dem Hintergrund einer Implikationsbeziehung zwischen vorgeordneter PH und geprüfter SH (hier: Modellgeltungshypothese) logisch unhaltbar:

> „Plötzlich soll auch die Verneinung von SH nur eine notwendige Bedingung für die Verneinung (Falsifikation) von PH sein. (...) Diese Bescheidenheit hat aber Konsequenzen. Zunächst einmal steht sie im Widerspruch zu ‚Wenn PH, dann SH'. Durch Kontraposition folgt daraus nämlich ‚Wenn nicht SH, dann nicht PH', d.h. die Verneinung von SH *ist* eine hinreichende Bedingung für die Verneinung von PH" (Westmeyer, 1998, S. 259; Hervorhebung im Original).

Diese Analyse und weitere Schlußfolgerungen, die Westmeyer daraus ableitet, beruhen auf dem Mißverständnis, daß „ein Ergebnis, das für die Verneinung der implizierten SH spricht" (Erdfelder & Bredenkamp, 1994, S. 609) mit der Negation der SH logisch äquivalent ist. Wäre dies der Fall, hätte Westmeyer Recht: Legt man den Wahrheitswert „falsch" für die SH fest, erzwingt dies via modus tollens die Falsifikation der vorgeordneten PH. Tatsächlich ist jedoch „ein Ergebnis, das für die Verneinung der implizierten SH spricht" immer ein bestimmter Wert einer Teststatistik, bei Modellgeltungstests z.B. ein (hoher) G^2-Wert (vgl. Kapitel 4.3). Wie auch immer der Wert dieser Teststatistik ausfällt: Eine Entscheidung über den Wahrheitswert der geprüften SH ist in *keinem* Fall logisch impliziert. Weder die Anerkennung noch die Negation der SH kann durch Daten erzwungen werden, schon gar nicht durch Daten einer einzelnen empirischen Untersuchung. Ob man in Anbetracht eines Prüfstatistik-Wertes, der in den Rejektionsbereich einer SH

fällt, die geprüfte SH für falsch hält, ist daher vernünftigerweise von Randbedingungen abhängig zu machen, beispielsweise von der Stabilität des entsprechenden Befundes in Replikationsstudien, von etwaigen weiteren „belastenden" Befunden zur gleichen PH, von der Verfügbarkeit alternativer psychologischer Hypothesen, welche die gleichen Befunde besser erklären können usw. (vgl. Bredenkamp, 1980). Aus diesem Grunde ist „ein Ergebnis, das für die Verneinung der implizierten SH spricht", im Rahmen eines methodologischen Falsifikationismus nicht hinreichend für die Falsifikation der vorgeordneten PH. Notwendig für eine Falsifikation ist ein solcher Befund, weil wir Hypothesen niemals „grundlos" verwerfen wollen, d.h. niemals ohne mindestens ein belastendes empirisches Resultat. Argumente wie beispielsweise „Unplausibilität der theoretischen Annahmen" oder „unzeitgemäße bzw. überholte Theoriebildung", die in Review-Verfahren wissenschaftlicher Veröffentlichungen gelegentlich als hinreichende Gründe für die Zurückweisung eines Manuskripts angesehen werden – können im Rahmen eines methodologischen Falsifikationismus niemals als alleinige Gründe für die Verwerfung von Hypothesen oder Theorien akzeptiert werden.

Man mag einwenden, daß die Unmöglichkeit, Hypothesenfalsifikationen allein durch Daten zu erzwingen, der Preis ist, den man für die Einführung statistischer Hypothesen in den Hypothesenprüfungsprozeß zu bezahlen hat. Dieses Argument beruht jedoch auf einer unkritischen Sicht des Falsifikationsmechanismus bei rein deterministischen Theorien. Auch dort setzt die Anerkennung falsifizierender „Protokollsätze" oder „Basissätze" gewisse konventionalistische Elemente voraus. Niemals ist die Übertragung von Beobachtungen in „Protokollsätze" oder „Basissätze" ein Automatismus, der ohne methodologische Konventionen auskommt. Schon Popper (1934/1982, S. 71) sagte ganz klar: „Die Basissätze werden durch Beschluß, durch Konvention anerkannt, sie sind *Festsetzungen*". Lakatos (1970, S. 99) stellte noch allgemeiner fest: „*No factual proposition can ever be proved from an experiment*. Propositions can only be derived from propositions, they cannot be derived from facts". An anderer Stelle stellt er konsequenterweise fest: „Methodological falsificationism is a brand of conventionalism" (a. a. O., S. 104). Der Unterschied zu anderen konventionalistischen Schulen (z.B. Dingler, 1931) besteht im wesentlichen darin, daß Falsifikationisten vereinbaren, nur „beobachtungsnahe" Aussagen (Basissätze oder auch statistische Hypothesen) per Beschluß anzuerkennen, nicht jedoch theoretische Kernannahmen selbst. Kernannahmen bleiben somit kritisierbar; sie sind nicht – wie beispielsweise im Dinglerschen Konventionalismus – gegenüber empirischer Kritik immun.

Als Fazit bleibt festzuhalten, daß man sich mit einer deduktivistischen Methodologie der Hypothesenprüfung, die wesentlich auf statistischen Hypothesentests aufbaut, keine prinzipiell neuen Probleme einhandelt, die im Rahmen rein deterministischer Theorien vermeidbar wären.

2.6 Das Dekomponierungsproblem

Mit dem Fehlerproblem und dessen Lösung über probabilistische Meßmodelle hängt ein weiteres Problem des Operationalisierungsansatzes eng zusammen: Empirische Konzepte sind typischerweise niemals „prozeßreine" (Jacoby, 1991) oder „systemreine" Indikatoren eines ganz bestimmten theoretischen Konstrukts, sondern reflektieren im Regelfall auch noch andere psychologische Prozesse, die mit dem fraglichen Konstrukt gar nichts zu tun haben. Wenn der empirische Indikator von einer Situation zur anderen variiert, so heißt das also nicht zwingend, daß auch das interessierende theoretische Konstrukt variiert, denn andere (Stör-) Einflüsse können hierfür verantwortlich sein. Man kann dies als „Separierungsproblem" oder „Dekomponierungsproblem" bezeichnen (Erdfelder, 1992a; Erdfelder & Bredenkamp, 1994). Eine adäquate Verwendung der Indikatoren impliziert offensichtlich, daß die verschiedenen psychologischen Einflüsse auf den betreffenden Indikator voneinander separiert werden, was praktisch bedeutet, daß man den empirischen Indikator in adäquate Maße der darauf einflußnehmenden Prozesse dekomponieren muß.

Das Dekomponierungsproblem hat zunächst einmal die praktische Konsequenz, daß man keine injektive Abbildung eines „psychologischen Raumes" auf einen „Verhaltensraum" oder „Antwortraum" konstruieren kann, wie das im Falle des deterministischen Meßmodells von Levine (1966) möglich war (vgl. Abschnitt 2.4). Folglich muß man auch die Idee aufgeben, im Einzelfall von einem bestimmten Verhaltens- oder Antwortmuster der Vp auf den verursachenden psychologischen Prozeß treffsicher rückschließen zu können. Schon im Falle des in Abschnitt 2.5 geschilderten LCA-Modells zum Levine-Paradigma ist ein derartiger Rückschluß nicht mehr möglich, denn ein bestimmtes beobachtetes Antwortmuster kann aufgrund von Fehlerprozessen aus ganz unterschiedlichen latenten Antwortmustern resultieren. Das Identifizierbarkeitsproblem stellt sich bei probabilistischen Modellen somit grundsätzlich anders dar als bei deterministischen Modellen: Nicht um die Identifizierbarkeit des verursachenden psychologischen Prozesses kann es bei probabilistischen Modellen gehen, sondern nur um die Identifizierbarkeit der *Modellparameter*, mit Hilfe derer die beobachtbare Verteilung der Antwortmuster beschrieben wird. Wir kommen im vierten Kapitel hierauf zurück.

Probabilistische Modelle ermöglichen es immerhin, sogenannte „recruitment probabilities" für beobachtbare Antwortmuster anzugeben, d.h. Angaben über die bedingten Wahrscheinlichkeiten potentiell zugrundeliegender psychologischer Prozesse bei gegebenem Antwortmuster (vgl. z.B. Rost & Erdfelder, 1996, S. 337). Vorauszusetzen ist hierbei natürlich, daß alle relevanten psychologischen

Zustände oder Prozesse im Modell explizit berücksichtigt wurden. Für das vorgestellte LCA-Modell zum Levine-Paradigma ergeben sich die „recruitment probabilities" zu einem beliebigen Antwortmuster (k_1, k_2, k_3, k_4) unter Verwendung des Bayesschen Theorems sowie der Gleichungen (2.5-2) und (2.5-3) wie folgt:

$$p(h_l|(k_1,k_2,k_3,k_4)) = \frac{p(h_l)\cdot p((k_1,k_2,k_3,k_4)|h_l)}{p(k_1,k_2,k_3,k_4)} = \frac{p(h_l)\cdot\prod_{j=1}^{4} p(k_j)_{j|l}}{\sum_{i=1}^{8} p(h_i)\cdot\prod_{j=1}^{4} p(k_j)_{j|i}} \quad . \qquad (2.6\text{-}1)$$

Sind die Parameter des LCA-Modells bekannt bzw. aus Stichprobendaten geschätzt, kann man unter Verwendung von Gleichung (2.6-1) die Wahrscheinlichkeit errechnen bzw. schätzen, daß dem beobachteten Antwortmuster (k_1, k_2, k_3, k_4) die latente Begriffshypothese h_l, l = 1, ... 8, zugrunde liegt.

Solange die Fehlerwahrscheinlichkeiten a_j und b_j gemäß Tabelle 2.2 klein sind, werden die „recruitment probabilities" oftmals eine bestimmte Begriffshypothese als die eindeutig wahrscheinlichste auszeichnen. Auch wenn man den verhaltenssteuernden psychologischen Prozeß nicht exakt bestimmen kann, so kann man ihn in diesem Fall doch wenigstens einigermaßen genau abschätzen. Leider liegen die Verhältnisse aber nicht immer so günstig wie hier angenommen. Das Dekomponierungsproblem resultiert nämlich häufig nicht ausschließlich aus Fehlerprozessen mit geringen Fehlerwahrscheinlichkeiten, sondern typischerweise aus verschiedenen, annähernd gleich wahrscheinlichen psychologischen Prozessen, die in dieselbe beobachtbare Verhaltenskategorie einmünden.

Ein einfaches Beispiel sind Ja-Nein-Rekognitionstests in der Gedächtnispsychologie: Ein Proband bekommt eine Folge von Items vorgelegt, von denen einige zuvor bereits dargeboten wurden (alte Items), andere dagegen nicht (neue Items oder Distraktoren). Für jedes Item soll ein „Alt"- oder ein „Neu"-Urteil abgeben werden. Offensichtlich können „Alt"-Antworten durch Vertrautheit bzw. bewußtes Wiedererkennen eines Items auf der einen Seite und durch Antworttendenzen („Alt"-Sage-Tendenzen) auf der anderen Seite beeinflußt sein. Der Einfluß der Antworttendenzen kann unter bestimmten Kontextbedingungen so stark sein, daß eine beobachtete „Alt"-Antwort keinen der denkbaren verursachenden Prozesse als eindeutig wahrscheinlichsten auszeichnet. Dies ist z.B. dann der Fall, wenn hohen Belohnungen (payoffs) für korrekte „Alt"-Antworten vergleichsweise geringe Bestrafungen (negative payoffs) im Falle falscher „Alt"-Antworten gegenüberstehen. Die Vpn werden in solchen Fällen geneigt sein, sehr häufig „alt" zu sagen, auch wenn keine bewußte Rekognition und nur ein schwaches Vertrautheitsgefühl vorliegt. Das Dekomponierungsproblem besteht dann darin, gedächtnisbasierte Einflüsse (Vertrautheit, bewußte Rekognition) und motivationale Einflüsse (Antworttendenzen) auf beobachtbare „Alt"-Antworten voneinander zu se-

parieren. Da keiner der beiden Einflußfaktoren vernachlässigbar ist, kann das Dekomponierungsproblem nicht mehr unter Hinweis auf die mutmaßlich geringen Fehlerwahrscheinlichkeiten beim Schluß vom Verhalten auf den verursachenden psychologischen Prozeß ignoriert werden.

Welche Lösungsansätze stehen in solchen Fällen zur Verfügung? Man kann mit Brainerd (1985) zwei Lösungsansätze unterscheiden, nämlich den designbasierten (design based) und den modellbasierten (model based) Ansatz. Der designbasierte Ansatz ist der in der experimentellen kognitiven Psychologie bislang am häufigsten gewählte Weg der Überwindung des Dekomponierungsproblems. Im Kern besteht der Ansatz darin, die empirischen Indikatoren für die interessierenden theoretischen Konstrukte auf der einen Seite und die Kontextbedingungen der Untersuchungssituation auf der anderen Seite so aufeinander abzustimmen, daß Störeinflüsse konkurrierender psychologischer Prozesse minimiert werden.

Ein anschauliches Beispiel für dieses Vorgehen ist der sogenannte „modifizierte Rekognitionstest" von McCloskey und Zaragoza (1985). McCloskey und Zaragoza (1985) haben die von Elizabeth Loftus (z.B. Loftus, Miller & Burns, 1978) vertretene Theorie zum Effekt nachträglicher Falschinformation auf das Zeugengedächtnis kritisiert (vgl. Kapitel 10). In einem typischen Falschinformationsexperiment bekommt eine Experimentalgruppe von Vpn nach visueller Darbietung eines kritischen Details X (z.B. ein Stop-Schild auf einem Dia) nachträglich eine falsche Information Y zu diesem kritischen Detail (z.B. das Dia habe ein Vorfahrt-Achten-Schild gezeigt), während die Kontrollgruppe eine neutrale Nachinformation erhält (z.B. das Dia habe ein Verkehrsschild gezeigt). In einem überraschenden two-alternative forced-choice (2AFC) Rekognitionstest muß später angegeben werden, ob ursprünglich das Detail X oder das Detail Y visuell dargeboten wurde. Dieses experimentelle Design soll im folgenden kurz „Loftus-Paradigma" genannt werden. Die Tatsache, daß Vpn der Experimentalgruppe im Loftus-Paradigma häufiger die Falschinformation Y wählen als Vpn der Kontrollgruppe, muß nicht zwingend bedeuten, daß das Gedächtnis für die Originalinformation durch die spätere Falschinformation beeinträchtigt oder dessen Abrufbarkeit herabgesetzt wurde. Denkbar sei, so McCloskey und Zaragoza (1985), daß z.B. aufgrund von mangelnder Aufmerksamkeitszuwendung zum kritischen Detail X nur ein bestimmter Anteil der Vpn-Stichprobe dieses Detail während des Gedächtnistests abrufen kann, ein Anteil, der in Experimental- und Kontrollgruppe gleich groß ist und folglich nichts mit möglichen Gedächtnisbeeinträchtigungen durch die Nachinformation zu tun haben kann. Dennoch kann ein systematischer Experimental-Kontrollgruppen-Unterschied im anschließenden Rekognitionstest resultieren, weil die Vpn, die sich nicht an die Originalinformation erinnern, in der Experimentalgruppe wahrscheinlich eher die falsche Nachinformation wählen werden als in der Kontrollbedingung, in der diese Nachinformation ja gar nicht prä-

sentiert wurde. Der Gruppenunterschied käme somit nicht aufgrund eines Gedächtnisbeeinträchtigungseffekts zustande, sondern aufgrund einer Beeinflussung von Antworttendenzen durch nachträgliche Falschinformationen.

Um diese Hypothese zu prüfen, ersetzten McCloskey und Zaragoza (1985) den von Loftus et al. (1978) verwendeten Rekognitionstest durch eine modifizierten 2AFC-Test, bei dem nicht die Originalinformation X gegen die Falschinformation Y kontrastiert wird, sondern die Originalinformation X gegen eine neue, für beide Versuchsgruppen neutrale Information Z. Liegt wirklich eine Gedächtnisbeeinträchtigung vor, so sollte auch in diesem modifizierten Test ein Experimentalgruppennachteil relativ zur Kontrollgruppe feststellbar sein. Nach der Antworttendenzhypothese sollte der Unterschied dagegen verschwinden, denn die Alternative Z ist nun für beide Gruppen des Experiments neu, so daß es keinen Grund für differentielle Rateeffekte gibt. Es stellte sich in den Experimenten von McCloskey und Zaragoza (1985) tatsächlich kein Effekt im modifizierten Rekognitionstest ein, was die Autoren als Beleg für die Antworttendenzhypothese interpretierten.

Die Arbeit von McCloskey und Zaragoza (1985) ist ein besonders gelungenes und oft zitiertes Beispiel für den designbasierten Ansatz zum Dekomponierungsproblem (weitere Beispiele finden sich u.a. bei Brainerd, 1985). Dennoch werden auch in dieser Untersuchung die Nachteile des Ansatzes deutlich. Zunächst einmal ist es nicht leicht, die Alternative Z im modifizierten Test so auszuwählen, daß sie wirklich für beide Experimentalbedingungen gleichermaßen neutral ist. Wenn das nicht gelingt, ist die Kontaminierung des Testergebnisses mit Antworttendenzen im modifizierten Test bloß reduziert und nicht komplett aufgehoben. Darüber hinaus liegt die Kritik nahe, daß der modifizierte Rekognitionstests nicht nur weniger sensitiv für Antworttendenzen, sondern auch weniger sensitiv für mögliche Gedächtniseffekte ist. Folglich lassen sich Gedächtnisbeeinträchtigungen mit dem modifizierten Test nicht mehr so leicht nachweisen, wenn sie tatsächlich vorliegen (Ayers & Reder, 1998; Belli, 1989; Loftus, Schooler & Wagenaar, 1985). Tatsächlich konnte man inzwischen zeigen, daß bei günstig gewählten Randbedingungen ein kleiner Experimentalgruppen-Kontrollgruppen-Unterschied auch mit dem modifierten Test feststellbar ist (Belli, Windschitl, McCarthy & Winfrey, 1992; Payne, Toglia & Anastasi, 1994; Windschitl, 1996). Schreiber und Servan (1998) haben jüngst sogar starke Effekte im modifizierten Rekognitionstest beobachtet; ihnen gelang das dadurch, daß sie vor dem modifizierten Rekognitionstest den Abruf der Falschinformation in einem anderen Kontext erzwungen haben („repeated retrieval"). Derartige Ergebnisse lassen sich mit einer Gedächtnisbeeinträchtigungstheorie, wie sie von Elizabeth Loftus vertreten wird, wieder vereinbaren. Der modifizierte Rekognitionstest hat also bislang nicht zu einer endgültigen Klärung der Frage nach den Ursachen des Falschinformationseffekts geführt.

Die aufgeführten Nachteile des designbasierten Ansatzes machen den modellbasierten Ansatz zum Dekomponierungsproblem attraktiv. Kernidee ist hierbei, alle relevanten Prozesse, die auf die Daten Einfluß nehmen, explizit mitzumodellieren, seien dies nun Fehlerprozesse wie im Falle des LCA-Modells zur Blank-Trial-Prozedur von Levine oder Antworttendenz- und Rateprozesse, die man bei Rekognitionstests annehmen muß. Die verschiedenen Prozesse werden dabei durch unterschiedliche Modellparameter repräsentiert, die auf der Grundlage von Stichprobendaten schätzbar sind.

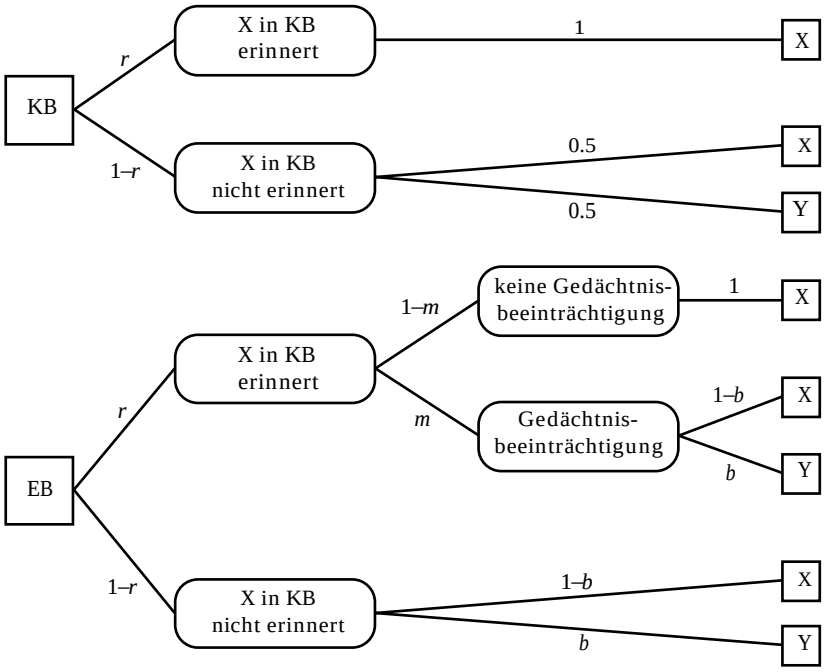

Abbildung 2.1:
Ein multinomiales Modell zum Paradigma von Loftus et al. (1978). EB steht für die Experimentalbedingung mit nachträglicher Falschinformation Y zum kritischen Detail X der Originalinformation, KB für Kontrollbedingung ohne falsche Nachinformation. Versuchsbedingungen sind in eckigen Kästchen links, Antworten der Vpn in den eckigen Kästchen rechts angegeben. Kästchen mit abgerundenten Ecken repräsentieren latente Prozesse. Die Parameter werden im Text erläutert.

Im Falle des von Loftus et al. (1978) verwendeten 2AFC-Rekognitionstests könnte ein entsprechendes Modell z.B. so wie in Abbildung 2.1 dargestellt aussehen. Beobachtbare Ereignisse werden dort in eckigen Kästchen, latente psychologische Zustände und Prozesse in Kästchen mit abgerundeten Ecken dargestellt. Die Parameter und Zahlen an den Verbindungslinien stellen Übergangswahrscheinlichkeiten dar, die jeweils von links nach rechts zu lesen sind: Links steht immer der Ausgangszustand, rechts der Zustand, dessen bedingte Wahrscheinlichkeit bei gegebenem Ausgangszustand repräsentiert werden soll. Dies kann — muß aber nicht — der Zustand sein, der auf den Ausgangszustand zeitlich folgt. Die in Abbildung 2.1 verwendete „Syntax” wird in dieser Arbeit bei allen graphischen Veranschaulichungen von sogenannten Verarbeitungsbaummodellen zugrunde ge-

legt. Im nächsten Kapitel wird genauer ausgeführt, was unter einem „Verarbeitungsbaummodell" zu verstehen ist.

Der Parameter *r* des vorgeschlagenen Modells ist die unbekannte Wahrscheinlichkeit, daß in der Kontrollbedingung das kritische Detail X der Originalinformation erinnert und im Rekognitionstest folglich auch gewählt wird. Wenn der Abruf nicht gelingt, wird in der Kontrollbedingung mit einer Wahrscheinlichkeit von jeweils 0.5 zugunsten von X bzw. Y geraten. Diese Annahme ist plausibel, wenn — wie im Falle der Untersuchung von Loftus et al. (1978) — die Itemmaterialien über Vpn-Gruppen hinweg ausbalanciert wurden, so daß für die Hälfte der Vpn ein bestimmtes Detail A als X und ein davon abweichendes Detail B als Y fungiert, während für die andere Hälfte der Probanden umgekehrt B als X und A als Y Verwendung findet. In diesem Falle sind etwaige Antworttendenzen zugunsten von A oder B nicht mehr mit X bzw. Y konfundiert, so daß in der Kontrollbedingung von gleichen Ratewahrscheinlichkeiten ausgegangen werden kann.

In der Experimentalbedingung sind die Ratewahrscheinlichkeiten demgegenüber nicht notwendig gleich, weil möglicherweise die falsche Nachinformation Y abrufbar ist und das Antwortverhalten beeinflussen kann. Es wird deshalb angenommen, daß dort mit einer bedingten Wahrscheinlichkeit $b \geq 0.5$ zugunsten der falschen Nachinformation Y geraten wird, wenn X nicht mehr abrufbar ist.

Von zentraler Bedeutung ist der Parameter *m*. Er steht für den Anteil der kritischen Items, die aufgrund einer Gedächtnisbeeinträchtigung durch die nachträgliche Falschinformation in der Experimentalbedingung nicht mehr erinnerbar sind, in der Kontrollbedingung jedoch wegen des Fehlens der Falschinformation abgerufen werden können. Nach der Gedächtnisbeeinträchtigungstheorie von Elizabeth Loftus (und auch nach anderen Gedächtnis- oder Abrufbeeinträchtigungstheorien, s. Kapitel 10) muß dieser Parameter größer als null sein, nach der Antworttendenzhypothese von McCloskey und Zaragoza dagegen gleich null. Sofern das Modell als adäquat betrachtet werden kann und der Parameter *m* hinreichend genau schätzbar ist, kann daher mit statistischen Mitteln zwischen den genannten Theorieklassen zum Falschinformationseffekt entschieden werden.

Das Modell gemäß Abbildung 2.1 hat allerdings ein Problem, das deutlich wird, wenn man die Modellgleichungen betrachtet. In den Modellgleichungen werden die Wahrscheinlichkeiten für die Wahl einer bestimmten Alternative beim 2AFC-Rekognitionstests als Funktion der Modellparameter *r*, *b* und *m* ausgedrückt. Da sich die Wahrscheinlichkeiten für die Wahl der X- und der Y-Alternative zu eins aufaddieren müssen, gibt es für das ursprüngliche Loftus-Miller-Burns-Experiment nur zwei unabhängige Modellgleichungen, nämlich die für die Wahrscheinlichkeiten richtiger Antworten in der Experimental- und in der Kontrollbedingung, $p(\{X\})_{EB}$ bzw. $p(\{X\})_{KB}$. Um diese Wahrscheinlichkeiten zu erhalten, berechnet man zunächst die Wahrscheinlichkeiten der verschiedenen „Äste"

aus Abbildung 2.1, indem man einfach die (bedingten) Wahrscheinlichkeiten der involvierten Teilprozesse multipliziert. Anschließend summiert man über alle Äste, die in die korrekte Antwort X einmünden. In der Experimentalbedingung sind dies drei Äste, in der Kontrollbedingung zwei. Es ergibt sich:

$$p(\{\mathrm{X}\})_{\mathrm{EB}} = r \cdot (1-m) + r \cdot m \cdot (1-b) + (1-r) \cdot (1-b) \qquad (2.6\text{-}2)$$

$$p(\{\mathrm{X}\})_{\mathrm{KB}} = r + 0.5 \cdot (1-r) \qquad (2.6\text{-}3)$$

Offenkundig kann eine Bestimmung von drei unbekannten Parameter auf der Grundlage von nur zwei Modellgleichungen nicht gelingen. Wie man mit diesem sogenannten Identifizierbarkeitsproblem umgehen kann, soll zunächst ausgeklammert bleiben. Wir kommen in Kapitel 4 hierauf zurück. Nur soviel sei schon gesagt: Es gibt fast immer eine befriedigende Lösung des Identifizierbarkeitsproblems, wenn man gewillt ist, die empirische Basis des Modells geeignet zu erweitern (vgl. insbesondere Abschnitt 10.1).

Hier kommt es darauf an, den wesentlichen Unterschied zwischen designbasiertem und modellbasierten Lösungsansatz zum Dekomponierungsproblem deutlich zu machen: Während beim designbasierten Ansatz versucht wird, Störeinflüsse auszuschalten bzw. konstant zu halten, nimmt man sie beim modellbasierten Ansatz mit in das Modell auf, um auf diese Weise unkontaminierte, von Störeinflüssen unverfälschte Maße der eigentlich interessierenden kognitiven Prozesse zu erhalten, im o.g. Beispiel z.B. den Parameter *m*. Dies ist ein großer Vorteil im Vergleich zum designbasierten Ansatz, weil man an den experimentellen Kontextbedingungen und abhängigen Variablen festhalten kann, welche die zu erklärenden Effekte am klarsten zeigen. Was dieser Ansatz potentiell leisten kann, wird in den folgenden Kapiteln demonstriert und kritisch reflektiert.

2.7 Zusammenfassung

Dieses Kapitel hatte das Theorie-Empirie-Überbrückungsproblem in der empirischen Psychologie zum Gegenstand. Herausgearbeitet wurde, daß der klassische Lösungsansatz der Selbstbeobachtung (Introspektionismus bzw. Methode des verbalen Reports) allenfalls in Teilbereichen der kognitiven Psychologie geeignet ist, eine Verbindung zwischen theoretischen Konstrukten und empirischen Konzepten herzustellen. Viele kognitive Prozesse sind unbewußt und damit der Selbstbeobachtung nicht zugänglich, andere werden durch den Akt der Selbstbeobachtung möglicherweise sogar verfälscht.

Als Alternative zur Selbstbeobachtung kommen nur behaviorale Indikatoren für theoretisch angenomme Prozesse in Frage, allerdings nicht im Sinne des klassischen Operationalismus, sondern im Sinne eines Systems von Hilfshypothesen, das theoretische Konstrukte und empirische Konzepte zueinander in Bezie-

hung setzt. Dieser Ansatz, der gegenwärtig von den meisten empirisch arbeitenden Psychologen favorisiert wird, stößt allerdings auf drei Probleme, die mit den Bezeichnungen „Bedeutsamkeitsproblem", „Fehlerproblem" und „Dekomponierungsproblem" belegt wurden. Das Bedeutsamkeitsproblem ist das Problem der Beliebigkeit bei der Auswahl von empirischen Indikatoren. Dies hat zur Folge, daß Bewährungs- und Nichtbewährungsurteile über Theorien möglicherweise vom jeweils verwendeten Indikator abhängen können. Einen Ausweg aus diesem Problem bieten nur Meßmodelle, welche Bedingungen festlegen, die ein valides Maß des betreffenden Konstrukts erfüllen muß, ehe es zur Theorieevaluation herangezogen werden kann.

Das Fehlerproblem betrifft die genuine Variabilität psychologischer Daten, die aufgrund der vielfältigen Einflüsse auf menschliches Verhalten auch unter kontrollierten experimentellen Bedingungen kaum auszuschalten ist. Das Dekomponierungsproblem hängt eng mit dem Fehlerproblem zusammen: Empirische Indikatoren reflektieren oftmals nicht nur den primär interessierenden psychologischen Prozeß, sondern eine Vielzahl von Prozessen, seien dies nun Fehlerprozesse oder kognitive Prozesse anderen Ursprungs. Zu beiden Problemen wurden in der Literatur unterschiedliche Lösungsansätze ausgearbeitet. Gegenwärtig scheint ein modellbasierter Ansatz, bei dem die verschiedenen einflußnehmenden Prozesse durch Parameter in stochastischen Modellen repräsentiert werden, am vielversprechendsten zu sein (vgl. auch Steyer & Eid, 1993). Derartige Modelle können mit der Nullhypothese eines statistischen Modellgeltungstests identifiziert und entsprechend statistisch evaluiert werden. Dies ermöglicht einen formal adäquaten Umgang mit dem Bedeutsamkeitsproblem, dem Fehlerproblem und dem Dekomponierungsproblem.

Die folgenden Kapitel beschäftigen sich mit einer wichtigen Teilklasse von stochastischen Modellen, sogenannten multinomialen Modellen. Wie zu zeigen sein wird, sind multinomiale Modelle wegen ihrer Anforderungen an die Datenqualität, ihrer Flexibilität und ihrer relativen Robustheit gegenüber Verletzungen von statistischen Oberhypothesen i. S. von Stegmüller (1973b) ein besonders geeigneter stochastischer Rahmen für Meßmodelle in der kognitiven Psychologie.

3. Was sind multinomiale Modelle?

Multinomiale Modelle sind Modelle für sogenannte qualitative Daten, d.h. Daten, die kategorialer oder diskreter Natur sind. Sie setzen zunächst einmal ein wohldefiniertes System von endlich vielen Beobachtungskategorien voraus, wobei unter einer Beobachtungskategorie eine Äquivalenzklasse von Beobachtungen zu verstehen ist, die untereinander in bestimmter Hinsicht als qualitativ gleichwertig aufgefaßt werden. Ferner setzen sie voraus, daß jede mögliche Beobachtung, auf die sich das Modell bezieht, genau einer dieser diskreten Beobachtungskategorien eindeutig zugeordnet werden kann, so daß die Beobachtungskategorien nicht nur untereinander disjunkt, sondern zusammengenommen auch exhaustiv sind. Die Zuordnung einer Beobachtungskategorie zu einer Beobachtung bezeichnet man auch als Beobachtungsresultat. „Endlich viele Beobachtungskategorien" bedeutet in der praktischen Anwendung häufig „wenige Kategorien". Die Menge möglicher Beobachtungsresultate — man sagt häufig „Menge der Elementarereignisse" oder „Stichprobenraum (sample space) **S**" des Modells — steht also a priori fest und enthält typischerweise nur wenige Elemente. Hat eine Vp in einem 2AFC-Rekognitionstest zwei Wahlalternativen A und B (z.B. die Originalinformation und die falsche Nachinformation im Experiment von Loftus et al., 1978), so ist $\mathbf{S} = \{A, B\}$. Hat sie — wie im Falle der Untersuchung von Levine (1966) — die Möglichkeit, das linke (l) oder das rechte (r) Item beim *j*-ten Itempaar zu wählen, so ist $\mathbf{S}_j = \{l, r\}$; gibt es insgesamt vier solche Itempaare, so entspricht die Menge der Elementarereignisse dem in Abschnitt 2.5 eingeführten „Verhaltensraum", der alle 16 möglichen Antwortmuster über die vier Itempaare umfaßt. In diesem Fall ist also $\mathbf{S} = \mathbf{S}_1 \times \mathbf{S}_2 \times \mathbf{S}_3 \times \mathbf{S}_4 = \{(l, l, l, l), (l, l, l, r), ..., (r, r, r, r)\}$. Kategoriale Daten erhält man nun einfach dadurch, daß man den Beobachtungsvorgang zu einem bestimmten Stichprobenraum **S** mehrfach wiederholt und die Häufigkeiten der verschiedenen Beobachtungsresultate auszählt.

Die Einschränkung auf kategoriale Daten scheint auf den ersten Blick gravierend zu sein, denn die meisten Standardanalyseverfahren in der empirischen Psychologie — insbesondere Verfahren im Rahmen des sogenannten Allgemeinen Linearen Modells (ALM, vgl. z.B. Andres, 1996) — beziehen sich auf stetig verteilte Zufallsvariablen (ZVn) mit überabzählbar vielen Elementen in **S**. Betrachtet man die Anwendungen des ALM bzw. der darauf beruhenden Varianzanalyse in der Psychologie jedoch genauer, so wird deutlich, daß die typische abhängige Variable in der Psychologie im Grunde genommen kategorialer Natur ist und nur wie eine stetig verteilte ZV *behandelt* wird. Dies gilt etwa für Ratingskalen, Fehlerhäufigkeiten, die Anzahl gelöster Aufgaben in einem bestimmten Test, Häufigkeiten bestimmter Verhaltensweisen usw. Ungewöhnlich und aus statistischer Sicht problematisch ist also eher die Anwendung von Analyseverfahren für stetig verteilte

ZVn, weniger die Anwendung von Verfahren für kategoriale Daten. Im Grunde genommen lassen sich ALM-Anwendungen auf derartige Daten nur mit dem Argument rechtfertigen, daß die Varianzanalyse gegenüber Verletzungen der Verteilungsannahmen relativ robust ist, sofern die Zellenbesetzungen in einem Versuchsplan nicht zu unterschiedlich ausfallen und die Varianzen in verschiedenen Versuchsbedingungen annähernd homogen sind (Box, 1954). Allerdings ist fraglich, ob das Robustheitsargument in der praktischen Anwendung wirklich stichhaltig ist (Bradley, 1978; Oldenbürger, 1996).

Auch wenn Analyseverfahren für kategoriale Daten in weiten Bereichen der psychologischen Forschung angemessener erscheinen als Standard-ALM-Anwendungen, so gibt es doch keinen Zweifel daran, daß zumindest in einigen kognitionspsychologischen Forschungsfeldern Modelle für stetig verteilte ZVn durchaus ihre Berechtigung haben. Zu denken wäre hier etwa an Antwort- bzw. Reaktionszeitexperimente. Ist in solchen Fällen die Anwendung multinomialer Modelle prinzipiell ausgeschlossen? Dies ist schon allein deshalb nicht der Fall, weil man kontinuierliche Daten häufig zwanglos und fragestellungsbezogen „diskretisieren" kann, ohne daß aus substanzwissenschaftlicher Perspektive damit ein bedeutsamer Informationsverlust verbunden wäre. Zu denken ist hier etwa an Dichotomisierungen einer Variablen am Stichprobenmedian oder an die Betrachtung von Rangreihen kontinuierlicher Größen anstelle der Originalwerte. In vielen Kontexten wird es etwa von Interesse sein, ob die Antwortzeit eines Probanden unter einer experimentellen Bedingung A (t_A) größer oder kleiner ausfällt als unter einer Bedingung B (t_B). Wenn detailliertere quantitative Prognosen nicht intendiert oder nicht möglich sind, kann man ein Paar von positiv-reellwertigen Antwortzeiten, $(t_A, t_B) \in \mathbb{R}^+ \times \mathbb{R}^+$, problemlos auf eines der beiden diskreten Beobachtungsresultate aus $\mathbf{S} = \{(t_A > t_B), (t_A \leq t_B)\}$ abbilden, so daß multinomiale Modelle potentiell anwendbar werden[13]. Ein ähnliches Prinzip wurde beispielsweise bei der Konstruktion eines multinomialen Modells zum Hindsight-Bias-Effekt eingesetzt (Erdfelder, 1992a; Erdfelder & Buchner, 1998a), auf das ich in Abschnitt 7.1.9 eingehen werde.

Es wäre verfehlt, derartige Überlegungen als Empfehlung zu deuten, das Standardanalysemodell in der kognitiven Psychologie einfach zu wechseln, gleichsam im Sinne des Slogans „Das ALM ist tot, es lebe die multinomiale Modellierung!". Diese Lesart wäre schon deshalb sehr schädlich, weil es einige Modelle für stetig verteilte ZVn gibt, die sich speziell für die Überprüfung kognitiver Prozeßmodelle anhand von Antwortzeiten bzw. Reaktionszeiten eignen und teilweise besser zwischen Theorien trennen können als multinomiale Modelle dies erlau-

[13]Das Symbol $\mathbb{R}$ steht in dieser Arbeit für die Menge der reellen Zahlen, $\mathbb{R}^+$ entsprechend für die Menge der positiven reellen Zahlen.

ben (für einen Überblick siehe Luce, 1986; Townsend & Ashby, 1983). Weiterhin gibt es die Möglichkeit, multinomiale Modelle mit Modellen für kontinuierliche Reaktionszeiten zu verknüpfen (vgl. Kapitel 6.8). Derartige Ansätze sollten natürlich genutzt werden, wann immer es die substanzwissenschaftliche Fragestellung nahelegt. Hier soll lediglich argumentiert werden, daß die gegenwärtige Dominanz der Varianzanalyse und anderer ALM-Verfahren bei der Analyse kognitionspsychologischer Daten kein Grund ist, Modellen für kategoriale Daten einen untergeordneten Stellenwert zuzusprechen. Im Gegenteil: Wie zu zeigen sein wird, lassen sich kognitionspsychologische Annahmen leichter und zwangloser im Rahmen eines multinomialen Modellierungsansatzes als über ALM-basierte Analysen überprüfen. In diesem Sinne können multinomiale Modelle eine Lücke schließen zwischen (a) umfassenden kognitionspsychologischen Theorien wie z.B. globalen Gedächtnismodellen (vgl. Clark & Gronlund, 1996) und konnektionistischen Modellen (Anderson, 1995), die kognitive Prozesse gleichsam auf einer mikroskopischen Ebene beschreiben, und (b) den traditionell verwendeten Standardanalysemodellen, die oftmals keine oder nur eine sehr vage Beziehung zu den zugrundeliegenden substanzwissenschaftlichen Annahmen aufweisen (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988).

Die obigen Ausführungen sollten deutlich gemacht haben, daß der Terminus „multinomiale Modelle" in dieser Arbeit nicht als Sammelbezeichnung für Log-Lineare-Modelle, Logit-Modelle, Latent-Class-Modelle, Probit-Modelle und andere Analysemodelle benutzt wird, die auf der Multinomialverteilungsannahme basieren und sich bei der Analyse von multivariaten kategorialen Daten (Kontingenztafeln) inzwischen eingebürgert haben (Andersen, 1990; 1996; Fienberg, 1977; Haberman, 1978; 1979; Langeheine & Rost, 1996). Das typische multinomiale Modell bezieht sich in der kognitiven Psychologie nicht auf *mehrere*, sondern auf *eine* kategoriale Variable, so daß die oben genannten Modelle gar nicht anwendbar sind. Da jedoch die möglichen Ausprägungskombinationen mehrerer kategorialer Variablen (d.h. die Zellen einer Kontingenztafel) immer auch als Ausprägungen einer Variablen aufgefaßt werden können, lassen sich umgekehrt die Modelle, von denen in dieser Arbeit die Rede ist, durchaus auch auf multivariate Datensätze anwenden. Ein Beispiel hierfür wurde bereits in Abschnitt 2.5 besprochen, nämlich das LCA-Modell zum Levine-Paradigma, welches — wie deutlich werden wird — zugleich auch ein MVB-Modell ist.

Die typische Anwendung multinomialer Modelle in der kognitionspsychologischen Forschung weist noch eine weitere Besonderheit auf: Man hat es selten mit *einem* multinomialen Modell für *einen* Datensatz, sondern häufig mit *mehreren* verbundenen Modellen für *mehrere* Datensätze zu tun, wobei es sich typischerweise um Datensätze für verschiedene experimentelle Bedingungen oder Probandengruppen handelt. Beim LCA-Modell zum Levine-Paradigma war das

zwar nicht der Fall, wohl aber beim in Abschnitt 2.6 vorgestellten Modell zum 2AFC-Rekognitionstest von Loftus et al. (1978): Dort wurden zwei verbundene multinomiale Modelle für (a) die Experimentalbedingung mit falscher Nachinformation und (b) die Kontrollbedingung ohne falsche Nachinformation formuliert. Oftmals ist gerade der Vergleich der Parameter aus verschiedenen experimentellen Bedingungen das eigentliche Ziel der multinomialen Modellierung. Liegen derartige Mehr-Gruppen-Modelle vor, spricht man von *verbundenen multinomialen Modellen* (joint multinomial models). Wir werden den Terminus im folgenden noch genauer definieren.

Allgemeine Zielsetzung von substanzwissenschaftlich motivierten (verbundenen) multinomialen Modellen ist die *Erklärung* der Wahrscheinlichkeiten von Beobachtungskategorien bzw. der Unterschiede in den Kategoriewahrscheinlichkeiten zwischen verschiedenen experimentellen Bedingungen oder Probandengruppen. „Erklären" bedeutet in diesem Zusammenhang, die empirisch schätzbaren Kategoriewahrscheinlichkeiten auf zugrundeliegende latente Parameter, für die bestimmte psychologische Interpretationen existieren, zurückführen zu können. Damit gehört die auf multinomialen Modellen basierende Datenanalyse ebenso wie z.B. die Strukturgleichungsmodellierung (z.B. Rietz, Rudinger & Andres, 1996; Rudinger, Andres & Rietz, 1994) zur Klasse der *konfirmatorischen Analyseverfahren*: Sie setzt voraus, daß a priori ein Modell formuliert wird, das angibt, wie die Wahrscheinlichkeiten der Beobachtungskategorien von den (psychologisch interpretierten) Parametern des Modells abhängen. Ohne ein solches Ausgangsmodell ist die Anwendung multinomialer Modellierung unmöglich.

Trotz dieser Tatsache ist — wie an vielen Beispielen in dieser Arbeit deutlich werden wird — auch eine exploratorische Nutzung multinomialer Modelle möglich, sofern sich die exploratorische Fragestellung auf die Werte bestimmter Parameter oder Parameterunterschiede bezieht. So kann man etwa die exploratorische Frage „Gibt es einen Gedächtnisbeeinträchtigungseffekt *m* im Experiment von Loftus et al. (1978)?" unter Verwendung des Modells aus Abschnitt 2.6 durch Anpassung auf geeignete Daten beantworten (vgl. Abschnitt 10.1). Hierbei ist jedoch immer vorauszusetzen, daß das Ausgangsmodell, in dem der betreffende Parameter vorkommt, adäquat ist. Primär geht es also darum, vorliegende Theorien oder Hypothesen formal zu präzisieren und in dieser präzisierten Form zu prüfen bzw. gegeneinander zu testen. Von sekundärem Interesse ist dabei, ob durch die Datenanalyse gänzlich neue Hypothesen oder Theorien heuristisch inspiriert werden.

Nach dieser informellen Einführung in den Gegenstand und die Zielsetzung kognitionspsychologisch motivierter multinomialer Modelle sollen die restlichen Abschnitte dieses Kapitels nun eine etwas genauere Beschreibung jener Modellklasse liefern, die in der vorliegenden Arbeit zugrunde gelegt wird. Ich beginne mit einer Charakterisierung des Modells der Multinomialverteilung, zunächst für

den Fall einer Versuchsbedingung oder Probandengruppe (Abschnitt 3.1), dann für den Fall mehrerer Gruppen (Abschnitt 3.2). Eine weitere Verallgemeinerung ist das Konzept parametrisierter multinomialer Modelle, das in Abschnitt 3.3 vorgestellt und an Beispielen veranschaulicht wird. Unter dieses Konzept lassen sich alle aus der Literatur bekannten multinomialen Modelle subsumieren, auch die bereits erwähnten Standardmodelle (Log-Lineare Modelle etc.). In Abschnitt 3.4 folgt eine Charakterisierung jener Teilklasse parametrisierter multinomialer Modelle, die in der vorliegenden Arbeit von primärem Interesse ist, nämlich allgemeine multinomiale Verarbeitungsbaummodelle (MVB-Modelle). Ein wichtiger und in Anwendungskontexten besonders nützlicher Spezialfall — binäre MVB-Modelle — wird in Abschnitt 3.5 separat behandelt und an Beispielen veranschaulicht. Im Anschluß an eine Diskussion der Repräsentierbarkeit von MVB-Modellen durch sogenannte Ereignisbaum- oder Verarbeitungsbaumdiagramme (Abschnitt 3.6) schließt eine Zusammenfassung das Kapitel ab (Abschnitt 3.7).

3.1 Das einfache multinomiale Modell

Betrachten wir zunächst ein einfaches multinomiales Modell für eine Probandengruppe bzw. eine Versuchsbedingung. Die Menge möglicher Beobachtungsresultate umfasse $J \in \mathbb{N}$ Beobachtungskategorien C_j, $j = 1, ..., J$, so daß für den Stichprobenraum $\mathbf{S} = \{C_1, C_2, ..., C_J\}$ gilt.[14] Im Falle des Modells zum Experiment von Levine (1966) ist z.B. $J = 16$ und die verschiedenen Kategorien C_j entsprechen den 16 möglichen Antwortmustern der Probanden über die vier blank trials hinweg. Zu $\mathbf{S}$ sei nun ein Wahrscheinlichkeitsmaß P: $\text{Pot}(\mathbf{S}) \rightarrow [0, 1]$ definiert, und die (unbekannte) Wahrscheinlichkeit der j-ten Kategorie sei $\pi_j := P(\{C_j\})$. Wegen der Disjunktheit und Exhaustivität der verschiedenen $\{C_j\}$ muß natürlich

$$\sum_{j=1}^{J} \pi_j = 1 \qquad (3.1\text{-}1)$$

gelten. Nur J–1 Kategoriewahrscheinlichkeiten sind also funktional unabhängig[15]; π_J ergibt sich immer aus der Differenz zwischen 1 und der Summe $\pi_1 + ... + \pi_{(J-1)}$.

Multinomiale Modelle basieren auf folgender Annahme: Es werden $N \in \mathbb{N}$ *unabhängige* Beobachtungen zum *gleichen* Wahrscheinlichkeitsraum $\langle \mathbf{S}, P \rangle$ ge-

[14] Das Symbol $\mathbb{N}$ bezeichnet die Menge der natürlichen Zahlen. Das Symbol $\text{Pot}(\mathbf{S})$ steht für die Menge aller Teilmengen von $\mathbf{S}$, d.h. die Potenzmenge von $\mathbf{S}$. [0, 1] ist das abgeschlossene reelle Intervall $\{x \in \mathbb{R} | 0 \leq x \leq 1\}$. Das Symbol „:=“ schließlich bedeutet „definitionsgemäß gleich”. Ansonsten lehne ich mich in dieser Arbeit weitgehend an die von Batchelder und Riefer (1990; 1999), Hu und Batchelder (1994) sowie Riefer und Batchelder (1988) verwendete Notation an, um die Orientierung in der Grundlagenliteratur zu erleichtern.

[15] J Variablen $x_1, x_2, ..., x_J$ heißen funktional abhängig, wenn es eine Funktion g gibt, so daß für alle Argumente $g(x_1, x_2, ..., x_J) = 0$ gilt. Gibt es eine solche Funktion nicht, so heißen die Variablen funktional unabhängig (vgl. z.B. Elandt-Johnson, 1971, Gleichung (5.45), S. 97).

macht, d.h. die Wahrscheinlichkeit eines beliebigen Beobachtungsresultats $\{C_{j(n)}\} \in \mathbf{S}$ bei der n-ten Beobachtung, $n = 1, ..., N$, hängt nur von der Kategorie, nicht aber von n oder den Resultaten anderer Beobachtungen ab. Durch diese Annahmen wird ein Wahrscheinlichkeitsraum zum erweiterten Stichprobenraum $\mathbf{S}^N$ induziert, der alle J^N möglichen Kombinationen von Beobachtungsresultaten bei den N Beobachtungen umfaßt. Dieser Stichprobenraum ist also eine Menge von N-Tupeln, d.h. $\mathbf{S}^N = \{(C_{j(1)}, ..., C_{j(n)}, ..., C_{j(N)}) | j(n) = 1, ..., J; n = 1, ..., N\}$. Der zugehörige Wahrscheinlichkeitsraum ist der Produktraum $<\mathbf{S}^N, P^N>$, wobei $P^N(\{(C_{j(1)}, ..., C_{j(n)}, ..., C_{j(N)})\}) = P(\{C_{j(1)}\}) \cdot ... \cdot P(\{C_{j(N)}\})$ für alle Elemente aus $\mathbf{S}^N$ gelten muß (vgl. z.B. Krengel, 1991, S. 27f). Hierauf definieren wir nun J Zufallsvariablen $Y_1, Y_2, ..., Y_J$, deren Werte $y_j \in \mathbb{N}, j = 1, ..., J$, jeweils die Häufigkeit der Beobachtungskategorie C_j bei den insgesamt N Beobachtungen angeben, wobei natürlich $\Sigma_j Y_j = N$ gelten muß. Ebenso wie im Falle der Kategoriewahrscheinlichkeiten π_j sind also nur $J-1$ der Zufallsvariablen Y_j funktional unabhängig. Man kann nun zeigen, daß unter den genannten Annahmen die Wahrscheinlichkeit, genau y_1 mal die Kategorie C_1, y_2 mal die Kategorie C_2 etc. und schließlich genau y_J mal die Kategorie C_J zu beobachten,

$$P^N(y_1, y_2, ..., y_J; \pi_1, \pi_2, ..., \pi_J) = \frac{N!}{y_1! \cdot y_2! \cdot ... y_J!} \pi_1^{y_1} \cdot \pi_2^{y_2} \cdot ... \cdot \pi_J^{y_J} = \\ = N! \prod_{j=1}^{J} \frac{\pi_j^{y_j}}{y_j!} \tag{3.1-2}$$

betragen muß.

Gleichung (3.1-2) ist die Wahrscheinlichkeitsfunktion für J multinomialverteilte Zufallsvariablen $Y_1, Y_2, ..., Y_J$ bei N-facher Beobachtung und Kategoriewahrscheinlichkeiten $\boldsymbol{\pi} := (\pi_1, \pi_2, ..., \pi_J)$ oder kurz die $(J, N, \boldsymbol{\pi})$-Multinomialverteilung. Manche Autoren benutzen auch den Terminus „Polynomialverteilung" (z.B. Bosch, 1986; Lienert, 1973). In der vorliegenden Arbeit soll dennoch der verbreitetere Terminus „Multinomialverteilung" benutzt werden.

Der Beweis von Gleichung (3.1-2) ist einfach (vgl. z.B. Bosch, 1986; Krengel, 1991, §1 und §2). Wir betrachten zunächst die Wahrscheinlichkeit des Ereignisses, daß die ersten y_1 Beobachtungen der Beobachtungsserie in die Kategorie C_1, die folgenden y_2 Beobachtungen in die Kategorie C_2 etc. und schließlich die letzten y_J Beobachtungen in die Kategorie C_J fallen. Da angenommen wird, daß sich (a) die Wahrscheinlichkeiten der Beobachtungskategorien über die N Beobachtungen hinweg nicht ändern und (b) die Beobachtungen selbst unabhängig sind, muß diese Wahrscheinlichkeit aufgrund der Definition der stochastischen Unabhängigkeit gerade das Produkt der Wahrscheinlichkeiten einzelner Beobachtungsresultate sein, d.h.

$$\pi_1^{y_1} \cdot \pi_2^{y_2} \cdot \ \ldots \ \cdot \pi_J^{y_J} . \tag{3.1-3}$$

Nun wird aber in Gleichung (3.1-2) nicht nach der Wahrscheinlichkeit einer *bestimmten* Sequenz der Beobachtungsresultate gefragt, sondern vielmehr nach der Wahrscheinlichkeit einer *beliebigen* Sequenz zu einem bestimmten Vektor $\boldsymbol{y} := (y_1, y_2, ..., y_J)$ von Kategoriehäufigkeiten. Da jede Sequenz von Beobachtungsresultaten mit dem gleichen Kategoriehäufigkeitsvektor $\boldsymbol{y}$ auch die gleiche Wahrscheinlichkeit hat, muß man den Term (3.1-3) einfach mit der Anzahl der Möglichkeiten multiplizieren, y_1 Beobachtungen der Kategorie C_1, y_2 Beobachtungen der Kategorie C_2 usw. in unterschiedlicher Weise sequentiell anzuordnen. Aus der Kombinatorik ist bekannt, daß diese Anzahl durch den sogenannten Multinomialkoeffizienten

$$\binom{N}{y_1, y_2, \ldots, y_J} := \frac{N!}{y_1! \cdot y_2! \cdot \ldots \cdot y_J!} \tag{3.1-4}$$

gegeben ist (Krengel, 1991, S. 14 f). Durch Multiplikation von (3.1-3) und (3.1-4) erhält man dann die Wahrscheinlichkeitsfunktion (3.1-2).

Ein anschauliches Beispiel für einen stochastischen Prozeß, der multinomialverteilte Kategoriehäufigkeiten generiert, ist das N-malige blinde Ziehen einer Kugel aus einer Urne, in der Kugeln mit J verschiedenen Farben C_1, C_2, ..., C_J in den Anteilen $\pi_1, \pi_2, ..., \pi_J$ vorhanden sind. Nach jedem Ziehen einer Kugel wird deren Farbe notiert; die Kugel wird anschließend in die Urne zurückgelegt und die Urne wird geschüttelt, so daß die Anteile der Farben bei jeder Ziehung gleich und die Ziehungsergebnisse unabhängig sind.

Ein wichtiger Spezialfall — speziell im Zusammenhang mit verbundenen Modellen (s. nächster Abschnitt) — ergibt sich, wenn nur zwei Beobachtungskategorien gegeben sind ($J = 2$). In diesem Fall heißt der stochastische Prozeß Bernoulli- oder Binomialprozeß, und die Wahrscheinlichkeitsfunktion (3.1-2) reduziert sich, wie man leicht nachvollziehen kann, auf die (N, π_1)-Binomialverteilung

$$P^N(y_1; \pi_1) = \frac{N!}{y_1! \cdot (N - y_1)!} \cdot \pi_1^{y_1} \cdot (1 - \pi_1)^{(N - y_1)} . \tag{3.1-5}$$

Die geschilderten Verteilungsannahmen entsprechen einer „statistischen Oberhypothese" im Sinne von Stegmüller (1973b). Typischerweise richtet sich das Interesse nicht auf die Überprüfung derartiger Oberhypothesen, sondern vielmehr auf die Schätzung der Parameter π_j und die Überprüfung bestimmter statistischer Hypothesen zu diesen Parametern (sog. Parameterhypothesen). Dennoch sind die Verteilungsannahmen wichtig, nämlich um (a) bestimmte Formen der Parameterschätzung (z.B. die Maximum-Likelihood-Schätzung) und (b) die Prüfung von Parameterhypothesen mittels statistischer Tests mathematisch begründen zu können

(vgl. Kapitel 4). Sind die Verteilungsannahmen bedeutsam verletzt, weil z.B. die Kategorienwahrscheinlichkeiten über die N Beobachtungen hinweg nicht gleich oder die Beobachtungen abhängig sind, so sind nominelle Maximum-Likelihood-Schätzer (ML-Schätzer) nicht wirkliche ML-Schätzer und haben folglich auch nicht ohne weiteres deren Eigenschaften (vgl. Abschnitt 4.2). Insbesondere können Angaben über die Größe von Konfidenzintervallen irreführend sein. Vielleicht noch gravierender ist, daß das nominelle Typ-1- (α) und Typ-2- (β) Fehlerrisiko beim statistischen Test einer Parameterhypothese von den faktischen Fehlerrisiken mehr oder minder stark abweichen kann. In diesem und im nächsten Kapitel soll zunächst unterstellt werden, daß die geschilderten Verteilungsannahmen erfüllt (oder zumindest annähernd erfüllt) sind. Praktisch entspricht das der Annahme, daß (a) die Kategoriewahrscheinlichkeiten verschiedener Vpn und/oder Items, über die zwecks Gewinnung von Stichprobenhäufigkeiten aggregiert wird, hinreichend ähnlich und (b) die Beobachtungen unabhängig oder wenigstens nahezu unabhängig von Beobachtungsergebnissen bei anderen Vp-Item-Kombinationen sind. Im Kapitel zur „Robustheit multinomialer Modellierung" (Kapitel 5) werde ich dann die Konsequenzen möglicher Verletzungen der Verteilungsannahmen näher untersuchen.

3.2 Das verbundene multinomiale Modell

Betrachten wir nun den Fall mehrerer Versuchsbedingungen oder Vpn-Gruppen wie er z.B. bei dem Modell zum Paradigma von Loftus et al. (1978) vorliegt, das in Abschnitt 2.6 vorgestellt wurde. Eine naheliegende Verallgemeinerung des multinomialen Modells (3.1-2) aus dem letzten Abschnitt basiert auf der Annahme, daß ein solches einfaches Modell innerhalb jeder von $K \in \mathbb{N}$ verschiedenen Versuchsbedingungen gilt. Die Anzahl der Beobachtungskategorien $C_{k,j}$ in Bedingung k, k = 1, ..., K, sei mit J_k bezeichnet (d.h. j = 1, ..., J_k), die beobachtete Kategoriehäufigkeit in Kategorie j der Bedingung k mit $y_{k,j}$ und die korrespondierende Kategoriewahrscheinlichkeit entsprechend mit $\pi_{k,j}$. Das komplette Kategoriehäufigkeitsmuster über alle K Versuchsbedingungen hinweg wird durch den Vektor

$$\boldsymbol{y} := (y_{1,1}, y_{1,2}, \ldots, y_{1,J_1}, y_{2,1}, y_{2,2}, \ldots, y_{2,J_2}, \ldots, \ldots, \ldots, y_{K,1}, y_{K,2}, \ldots, y_{K,J_K}) \quad (3.2\text{-}1)$$

beschrieben, welcher insgesamt $J_1 + J_2 + \ldots + J_K$ Kategoriehäufigkeiten umfaßt. Analog bezeichnet

$$\boldsymbol{\pi} := (\pi_{1,1}, \pi_{1,2}, \ldots, \pi_{1,J_1}, \pi_{2,1}, \pi_{2,2}, \ldots, \pi_{2,J_2}, \ldots, \ldots, \ldots, \pi_{K,1}, \pi_{K,2}, \ldots, \pi_{K,J_K}) \quad (3.2\text{-}2)$$

den Vektor der zugrundeliegenden Kategoriewahrscheinlichkeiten über alle K Versuchsbedingungen hinweg.

Der Fall, daß die Anzahl der Beobachtungskategorien in allen Versuchsbedingungen gleich ist ($\forall\, k\colon J_k = J$), ist als Spezialfall zugelassen, wird aber nicht vorausgesetzt. Ein solcher Spezialfall liegt bei dem Modell zum Loftus-Paradigma aus Abschnitt 2.6 vor. Wir haben es hierbei mit $K = 2$ experimentellen Bedingungen zu tun — neutrale und irreführende nachträgliche Information —, und die Anzahl der Beobachtungskategorien (gleichbedeutend mit der Anzahl der Wahlalternativen beim 2AFC-Rekognitionstest) ist in beiden Bedingungen $J_1 = J_2 = 2$.

Ebenso wie die Kategorienanzahl kann auch der Stichprobenumfang N_k in den Bedingungen gleich oder ungleich sein. Es wird also nicht vorausgesetzt, daß $N_1 = N_2 = ... = N_K$ gilt. Nimmt man weiterhin an, daß die Beobachtungshäufigkeiten in verschiedenen Versuchsbedingungen *unabhängig* multinomialverteilt sind, so induziert dies ein Wahrscheinlichkeitsmaß $Pr = P^{N1}\times P^{N2}\times ...\times P^{NK}$ auf dem Produkt der Wahrscheinlichkeitsräume $\langle \mathbf{S}^{N1}, P^{N1}\rangle$, $\langle \mathbf{S}^{N2}, P^{N2}\rangle$, ..., $\langle \mathbf{S}^{NK}, P^{NK}\rangle$. Hieraus ergibt sich die Wahrscheinlichkeit $Pr(\boldsymbol{y}; \boldsymbol{\pi})$ eines bestimmten Mehr-Gruppen-Stichprobenresultats $\boldsymbol{y}$ bei zugrundeliegenden Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ einfach durch Multiplikation der entsprechenden Wahrscheinlichkeiten für die Resultate einzelner Gruppen, d.h.

$$
\begin{aligned}
Pr(\boldsymbol{y};\boldsymbol{\pi}) &= \prod_{k=1}^{K} P^{Nk}(y_{k,1}, y_{k,2}, \ldots, y_{k,J_k}; \pi_{k,1}, \pi_{k,2}, \ldots, \pi_{k,J_k}) = \\
&= \prod_{k=1}^{K} \frac{N_k!}{y_{k,1}!\cdot y_{k,2}!\cdot\ldots\cdot y_{k,J_k}!} \pi_{k,1}{}^{y_{k,1}} \cdot \pi_{k,2}{}^{y_{k,2}} \cdot\ldots\cdot \pi_{k,J_k}{}^{y_{k,J_k}} = \qquad (3.2\text{-}3) \\
&= \prod_{k=1}^{K} N_k!\cdot \prod_{j=1}^{J_k} \frac{\pi_{k,j}{}^{y_{k,j}}}{y_{k,j}!} \ .
\end{aligned}
$$

Gleichung (3.2-3) ist die Wahrscheinlichkeitsfunktion des verbundenen multinomialen Modells (joint multinomial model) für K Gruppen mit Kategorienanzahl pro Gruppe gemäß $\boldsymbol{J} := (J_1, J_2, ..., J_K)$, Stichprobenumfängen pro Gruppe gemäß $\boldsymbol{N} := (N_1, N_2, ..., N_K)$ und Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$. Ich spreche auch kurz von der $(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi})$-Multinomialverteilung und kürze das Modell im folgenden mit $M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi})$ ab.

Natürlich ist die einfache $(J, N, \boldsymbol{\pi})$-Multinomialverteilung aus dem letzten Abschnitt ein Spezialfall des verbundenen multinomialen Modells, nämlich $M(1, J, N, \boldsymbol{\pi})$. Wenn es der Kontext erlaubt, beziehe ich mich daher im folgenden bei der Herleitung bzw. Beschreibung statistischer Resultate immer auf das verbundene multinomiale Modell; die Ergebnisse lassen sich problemlos auf das einfache multinomiale Modell übertragen. Ist $J_1 = J_2 = ... = J_K = J$, so heißt das verbundene multinomiale Modell $M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi})$ auch produktmultinomiales Modell. Ist zusätzlich noch die Kategorienzahl in allen Versuchsbedingungen $J = 2$, so kann man $M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi})$ produktbinomiales Modell nennen.

3.3 Parametrisierte multinomiale Modelle

Bislang wurde lediglich das statistische Rahmenmodell für den Ein- und den Mehr-Gruppen-Fall beschrieben. Wie einleitend in diesem Kapitel bereits betont wurde, besteht das Ziel multinomialer Modellierung in der kognitiven Psychologie wie auch in anderen Wissenschaften häufig nicht darin, die Wahrscheinlichkeiten der Beobachtungskategorien zu schätzen bzw. einfache Gleichheits- bzw. Ungleichheitshypothesen über diese Wahrscheinlichkeiten zu testen. Dies mag zwar in Einzelfällen von Interesse sein, und hierfür reicht der bislang beschriebene Modellrahmen auch aus. Häufiger ist jedoch der Fall, daß die Kategoriewahrscheinlichkeiten in Termini grundlegenderer Parameter erklärt werden sollen, für die bestimmte substanzwissenschaftliche Interpretationen existieren:

> „In many fields the objective of the investigator´s inquisitiveness is not just a ´population´ in the sense of a distribution of observable variables, but a physical structure projected behind this distribution, by which the latter is thought to be generated. The word ´physical´ is used merely to convey that the structure concept is based on the investigator´s ideas as to the ´explanation´ or ´formation´ of the phenomena studied, briefly, on his theory of these phenomena, whether they are classified as physical in the literal sense, biological, psychological, sociological, economic or otherwise" (Koopmans & Reiersøl, 1950).

Für kategoriale Daten kann dieses Ziel mit parametrisierten multinomialen Modellen erreicht werden, die in diesem Abschnitt näher erläutert werden sollen.[16]

Betrachten wir zunächst einige Beispiele. Im Modell zum Loftus-Paradigma aus Abschnitt 2.6 wurde versucht, die Kategoriewahrscheinlichkeiten — in diesem Fall die Antwortwahrscheinlichkeiten der Vpn — anhand von drei Parametern zu erklären, nämlich den Wahrscheinlichkeiten für eine erfolgreiche Erinnerung der Originalinformation in der Kontrollbedingung (r), für eine Gedächtnis- oder Abrufbeeinträchtigung in der Experimentalbedingung (m) und für das Erraten der Falschinformation bei nicht abrufbarer Originalinformation (b). Derartige Parameter nennt man auch *latente Parameter* oder *Parameter zweiter Ordnung*, um sie von den Parametern erster Ordnung — den Kategoriewahrscheinlichkeiten $\pi_{k,j}$ — abzugrenzen. Die aus bestimmten latenten Parametern vorhergesagte Wahrscheinlichkeit für eine Kategorie $C_{k,j}$ notieren wir mit $p_{k,j}$; sie muß mit der fakti-

[16] Manche Autoren ziehen den Begriff „parametrisches multinomiales Modell" (parametric multinomial model) vor (z.B. Andersen, 1990). Ich wähle die Bezeichnung „parametrisiertes multinomiales Modell", weil sie deutlicher macht, daß die Standardparameter $\boldsymbol{\pi}$ der Multinomialverteilung hierbei anhand *zusätzlicher* Parameter erklärt werden sollen. Die Bezeichnung „parametrisches multinomiales Modell" ist demgegenüber insofern mißverständlich, als jedes einfache oder verbundene multinomiale Modelle Parameter beinhaltet.

schen Kategoriewahrscheinlichkeit $\pi_{k,j}$ nicht zwangsläufig übereinstimmen. Jede Wahrscheinlichkeit $\pi_{k,j}$, ist eine im konkreten Anwendungsfall feststehende, wenn auch zumeist unbekannte Konstante, während $p_{k,j}$ mit den Werten der latenten Parameter variiert. Eine vollständigere Notation ist $p_{k,i}(r, m, b)$; sie macht die Abhängigkeit der $p_{k,j}$ von den latenten Parametern r, m und b deutlich.

Im allgemeinen Fall bezieht sich ein parametrisiertes multinomiales Modell — wie ein verbundenes multinomiales Modell — auf K Versuchsbedingungen mit J_k Beobachtungskategorien in Bedingung k, $k = 1, ..., K$, wobei K eine beliebige natürliche Zahl und J_k eine natürliche Zahl größer als 1 ist. Die insgesamt $J_1 + J_2 + ... + J_K$ Kategoriewahrscheinlichkeiten $\pi_{k,j}$ werden durch $S \in \mathbb{N}$ reellwertige latente Parameter „erklärt", die mit $\theta_1, \theta_2, ..., \theta_S$ notiert werden sollen. Eine notwendige Bedingung dafür, daß man von einer erfolgreichen Erklärung der faktischen Kategoriewahrscheinlichkeiten $\pi_{k,j}$ in Termini der latenten Parameter sprechen kann, ist offenbar, daß es latente Parameterwerte $\theta_1, \theta_2, ..., \theta_S$ gibt, derart, daß für die Parameter erster Ordnung $p_{k,j}(\theta_1, \theta_2, ..., \theta_S) = \pi_{k,j}$ für alle Kategorien $C_{k,j}$ gilt. Die faktisch vorliegenden Kategoriewahrscheinlichkeiten $\pi_{k,j}$ müssen also aus den latenten Parametern rekonstruierbar sein, wobei deren Werte innerhalb der vom Modell vorgegebenen Grenzen variieren können. Auf die Prüfung dieser kritischen Annahme komme ich in Abschnitt 4.3 zurück.

Grundlegend für parametrisierte multinomiale Modelle sind die Modellgleichungen, welche die vorhergesagten Wahrscheinlichkeiten $p_{k,j}(\theta_1, \theta_2, ..., \theta_S)$ als Funktion der latenten Parameter $\theta_1, \theta_2, ..., \theta_S$ ausdrücken. Im Falle des Modells zum Loftus-Paradigma sind das die Gleichungen (2.6-2) und (2.6-3). Diese Modellgleichungen machen deutlich, wie die Kategoriewahrscheinlichkeiten $\pi_{k,j}$ ($k = 1, 2$, $j = 1, 2$) bei Gültigkeit des Modells und gegebenen Werten für die latenten Parameter zustande kommen; sie „erklären" mithin die Kategoriewahrscheinlichkeiten in Termini der latenten Parameter.

Die latenten Parameter $\theta_1, \theta_2, ..., \theta_S$ können selbst Wahrscheinlichkeiten sein bzw. als Wahrscheinlichkeiten interpretiert werden. Sie sind dann natürlich zwangsläufig Elemente des reellen Intervalls [0, 1]. Dies ist z.B. bei dem soeben besprochenen Modell und auch bei dem Modell zum Levine-Paradigma aus Abschnitt 2.5 der Fall. Latente Parameter können jedoch auch beliebige andere reelle Werte annehmen.

Das in der kognitiven Psychologie wohl bekannteste Beispiel für ein solches Modell ist die Signalentdeckungstheorie (signal detection theory, vgl. Green & Swets, 1966; Macmillan & Creelman, 1991; Swets, 1998). Angewandt auf einen einfachen Ja-Nein-Rekognitionstest für zuvor dargebotene Items (alte Items) und zuvor nicht dargebotene Items (neue Items oder Distraktoren) basiert die einfachste Variante der Signalentdeckungstheorie auf der Annahme, daß Ja- bzw. Alt-Antworten resultieren, wenn die „Vertrautheit" eines Items — eine latente, nicht di-

rekt beobachtbare Variable — einen bestimmten kritischen Wert überschreitet, nämlich das Antwortkriterium $c \in \mathbb{R}$. Wird das Antwortkriterium nicht überschritten, so resultiert eine Nein- bzw. Neu-Antwort. Die Vertrautheit von Distraktoritems wird dabei als stetig verteilte reellwertige Zufallsvariable X_d aufgefaßt, die mit Erwartungswert 0 und Streuung 1 normalverteilt ist. Bezeichnet man die Verteilungsfunktion einer derartigen standardnormalverteilten Zufallsvariablen X mit $\Phi(X = x)$, so ist die Wahrscheinlichkeit p_d eines „falschen Alarms", d.h. einer Ja-Antwort für Distraktoritems, folglich

$$p_d = Pr(\{X_d \geq c\}) = 1 - \Phi(c) = \Phi(-c) \qquad (3.3\text{-}1)$$

und die komplementäre Wahrscheinlichkeit einer „korrekten Zurückweisung" (Nein-Antwort bei Distraktoren) entsprechend $1-p_d = \Phi(c)$.[17]

Für die Vertrautheit von alten Items, X_a, wird ebenfalls eine Normalverteilung mit Streuung 1 angenommen, deren Mittelwert d' jedoch wegen des möglichen Vertrautheitsanstieges, der durch die vorherige Präsentation der Items hervorgerufen wird, typischerweise größer als null ist. Um das Modell möglichst allgemein zu halten, lassen wir beliebige Vertrautheitsmittelwerte ($d' \in \mathbb{R}$) zu. Die Wahrscheinlichkeit p_a einer Ja-Antwort für alte Items (Treffer oder hit) ist folglich

$$p_a = Pr(\{X_a \geq c\}) = 1 - \Phi(c - d') = \Phi(d' - c) \qquad (3.3\text{-}2)$$

und die komplementäre Wahrscheinlichkeit eines „Verpassers" (miss) ist entsprechend $1-p_a = \Phi(c - d')$.

Die Gleichungen (3.3-1) und (3.3-2) sind die beiden Modellgleichungen der Standardversion der Signalentdeckungstheorie.[18] Sie erklären die Wahrscheinlichkeiten für falsche Alarme und Treffer in Termini des Alt-Neu-Diskriminationsparameters d' und des Antwortkriteriums c, wobei letzteres gelegentlich noch in andere, leichter interpretierbare Maße des Antwortbias transformiert wird (Macmillan & Creelman, 1990; Snodgrass & Corwin, 1988). Anders als die Parameter der Modelle zum Loftus- und zum Levine-Paradigma aus Kapitel 3 sind d' und c zwar latente Parameter, aber nicht latente Wahrscheinlichkeiten. Dennoch handelt es sich auch hierbei um ein parametrisiertes verbundenes multinomiales Modell, sofern die in Abschnitt 3.1 und 3.2 geschilderten Verteilungsannahmen verbundener multinomialer Modelle vorausgesetzt werden. Um dies schlüssig zu

[17] Die Verteilungsfunktion $\Phi(x)$ ist in nahezu allen Statistikbüchern tabelliert. Auch für die direkte Berechnung von $\Phi(x)$ für gegebenes x existieren recht einfache, aber dennoch sehr genaue Approximationsformeln (vgl. z.B. Abramowitz & Stegun, 1964; Johnson & Kotz, 1970a). Umgesetzt werden diese z.B. in dem FORTRAN-Unterprogramm PROBZ von O´Grady (1981). Da die zugehörige Dichtefunktion bzgl. $x = 0$ symmetrisch ist, gilt für Verteilungsfunktion $1-\Phi(x) = \Phi(-x)$ für alle $x \in \mathbb{R}$.

[18] Auch die beiden Gleichungen für die Wahrscheinlichkeiten korrekter Zurückweisungen und Verpasser sind selbstverständlich Modellgleichungen. Sie sind jedoch redundant, da sie sich aus den beiden Gleichungen (3.3-1) und (3.3-2) unmittelbar ableiten lassen.

zeigen, müssen wir die Klasse parametrisierter multinomialer Modelle zunächst in allgemeiner Form einführen.

Ein parametrisiertes multinomiales Modell kann wie folgt definiert werden:

(a) Sei $M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi})$ ein einfaches oder verbundenes multinomiales Modell.

(b) Sei $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_S)$ ein Vektor von $S \in \mathbb{N}$ latenten Parametern, die sämtlich Elemente bestimmter (u.U. parameterspezifischer) Intervalle $I_s \subseteq \mathbb{R}$ sind, $s = 1, ..., S$. Ein solches Intervall I_s kann also beispielsweise das durch 0 und 1 begrenzte geschlossene Intervall [0, 1], ein durch zwei beliebige reelle Zahlen begrenztes offenes, halboffenes oder geschlossenes Intervall oder auch die Menge der reellen Zahlen selbst sein. Die Menge aller möglichen Parametervektoren $\boldsymbol{\theta}$, der sogenannte Parameterraum

$$\boldsymbol{\Omega} = \{\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_S) \mid \theta_s \in I_s, s = 1, ..., S\} , \tag{3.3-3}$$

ist folglich eine (nicht notwendigerweise echte) Teilmenge des $\mathbb{R}^S$.

(c) Sei

$$\boldsymbol{\Gamma} = \left\{ \begin{array}{c} \boldsymbol{p} = (p_{1,1}, p_{1,2}, ..., p_{1,J_1}, p_{2,1}, p_{2,2}, ..., p_{2,J_2}, ..., p_{K,1}, p_{K,2}, ..., p_{K,J_K}) \\ \mid \forall k, j\colon p_{k,j} \in [0,1];\ \forall k\colon \sum_{j=1}^{J_k} p_{k,j} = 1 \end{array} \right\} \tag{3.3-4}$$

die Menge der möglichen Kategoriewahrscheinlichkeiten und $f\colon \boldsymbol{\Omega} \rightarrow \boldsymbol{\Gamma}$ eine Abbildung des Parameterraumes $\boldsymbol{\Omega}$ in $\boldsymbol{\Gamma}$.

Wenn die Bedingungen (a) bis (c) erfüllt sind, nenne ich

$$\langle M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}),\ \boldsymbol{\Omega},\ \boldsymbol{\Gamma},\ f \rangle \tag{3.3-5}$$

ein „parametrisiertes multinomiales Modell".

Führt man den Begriff eines parametrisierten multinomialen Modells so ein, dann kann ein solches Modell prinzipiell gültig oder auch ungültig sein. Gültig ist es, wenn es einen Parametervektor $\boldsymbol{\theta^0} \in \boldsymbol{\Omega}$ gibt, der über die Modellgleichungen auf die Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ abgebildet wird, d.h. $\boldsymbol{\pi} = f(\boldsymbol{\theta^0})$. Gibt es einen solchen Parametervektor nicht, so ist das parametrisierte multinomiale Modell ungültig (vgl. Abschnitt 4.3).

Wie man leicht sehen kann, erfüllen alle bislang besprochenen Modelle die Bedingungen parametrisierter multinomialer Modelle. Bei dem Modell zum Loftus-Paradigma aus Abschnitt 2.6 handelt es sich z.B. um ein Modell mit dem Parameterraum $\boldsymbol{\Omega} = [0, 1]^3$, weil jede der drei latenten Wahrscheinlichkeiten r, m und b zwischen 0 und 1 variieren kann. Ferner ist die Menge möglicher Kategoriewahrscheinlichkeiten gerade das Kreuzprodukt des Intervalls [0, 1] mit sich selbst — d.h. $\boldsymbol{\Gamma} = [0, 1]^2$ —, weil pro Versuchsbedingung nur eine der beiden Wahlwahrscheinlichkeiten frei zwischen 0 und 1 variieren kann. Im Falle des signa-

lentdeckungstheoretischen Modells liegt die gleiche Menge $\mathbf{\Gamma}$ vor, in diesem Fall allerdings zu interpretieren als Menge möglicher Wahrscheinlichkeiten von Treffern und falschen Alarmen. Demgegenüber haben wir es mit einem anderen Parameterraum zu tun, nämlich $\mathbf{\Omega} = \mathbb{R}^2$, da sowohl d' als auch c beliebige reelle Werte annehmen können. Die Modellgleichungen der beiden Modelle — (2.6-2) und (2.6-3) bzw. (3.3-1) und (3.3-2) — definieren jeweils die Abbildungen von $\mathbf{\Omega}$ in $\mathbf{\Gamma}$. Nimmt man noch die im letzten Abschnitt erläuterten Verteilungsannahmen verbundener multinomialer Modelle für $K = 2$ Versuchsbedingungen mit $J_1 = J_2 = 2$ Beobachtungskategorien hinzu, so ist die Zugehörigkeit zur Klasse der parametrisierten multinomialen Modelle (in diesem Falle: der parametrisierten produktbinomialen Modelle) offensichtlich.

Unglücklicherweise werden in manchen Veröffentlichungen Signalentdeckungsmodelle und multinomiale Modelle als einander ausschließende oder sogar antagonistische Modellklassen aufgefaßt (z.B. von Yonelinas & Jacoby, 1996). Der Terminus „multinomiale Modelle" wird von diesen Autoren für eine Teilklasse von parametrisierten multinomialen Modellen reserviert, auf die in den nächsten Abschnitten ausführlich eingegangen wird. Diese Einschränkung des Begriffs „multinomiale Modelle" ist unglücklich, weil der falsche Eindruck erweckt wird, als würden Signalentdeckungsmodelle nicht auf denn Annahmen parametrisierter multinomialer Modelle basieren. Tatsächlich ist jedoch der Unterschied zwischen Signalentdeckungsmodellen und anderen parametrisierten multinomialen Modellen nicht sehr groß. Wie soeben gezeigt, unterscheiden sich lediglich der Parameterraum $\mathbf{\Omega}$ und die Modellgleichungen. Auf die Methoden der statistischen Analyse, die damit verbundenen Probleme und die Interpretation der Ergebnisse hat das aber nur marginal Einfluß (Riefer & Batchelder, 1988; Wallsten, Bender & Li, 1999).

Nicht nur die bereits besprochenen, sondern auch alle anderen Modelle, die in dieser Arbeit vorgestellt werden, sind parametrisierte multinomiale Modelle im Sinne der obigen Definition. Dies gilt im übrigen für alle Klassen multinomialer Modelle, die in der einschlägigen Literatur als Grundlage für die Analyse von Kontingenztafeln und Häufigkeitsdaten vorgeschlagen wurden (Log-Lineare-Modelle, Logit-Modelle, Probit-Modelle etc.). Das Konzept eines parametrisierten multinomialen Modells ist also offenbar ein sehr allgemeines Konzept. Selbstverständlich enthält es auch einfache und verbundene multinomiale Modelle ohne latente Parameter als Spezialfälle; um dies zu zeigen, wählt man einfach $\mathbf{\Omega} = \mathbf{\Gamma}$ und f als Identitätsfunktion.

Bevor wir uns einer für die kognitive Psychologie besonders wichtigen Teilklasse parametrisierter multinomialer Modelle zuwenden — den bereits mehrfach erwähnten multinomialen Verarbeitungsbaummodellen — sollen noch drei für die folgenden Kapitel nützliche Begriffe eingeführt werden: die Begriffe „voll pa-

rametrisiertes Modell", „überparametrisiertes Modell" und „unterparametrisiertes Modell". Die Begriffe basieren auf der Dimensionalität der beiden Mengen $\mathbf{\Omega}$ und $\mathbf{\Gamma}$, d.h. auf der Anzahl funktional unabhängiger Parameter und der Anzahl funktional unabhängiger Kategoriewahrscheinlichkeiten (zur Definition der funktionalen Unabhängigkeit vgl. Fußnote 15). Wenn das Modell S funktional unabhängige latente Parameter θ_s umfaßt, ist der Parameterraum $\mathbf{\Omega}$ S-dimensional. Wenn K Gruppen mit jeweils J_k Beobachtungskategorien vorliegen, gibt es

$$M = \sum_{k=1}^{K} (J_k - 1) \tag{3.3-6}$$

funktional unabhängige Kategoriewahrscheinlichkeiten, da die letzte Kategoriewahrscheinlichkeit pro Versuchsbedingung wegen der Summationsregel (3.1-1) eindeutig festgelegt ist. Der Raum möglicher Kategoriewahrscheinlichkeitsvektoren $\mathbf{\Gamma}$ ist entsprechend M-dimensional.

Ein Modell heißt nun *voll parametrisiert,* wenn der Parameterraum und der Raum der Kategoriewahrscheinlichkeitsvektoren von gleicher Dimension sind, so daß $S = M$ gilt. Ist dagegen $S > M$, so ist das Modell *überparametrisiert,* ist $S < M$, so nennt man das Modell *unterparametrisiert*. Im letztgenannten Fall ist mit dem Modell offenbar eine empirisch potentiell widerlegbare Behauptung verbunden, die Behauptung nämlich, daß nicht die volle Dimensionalität des Raumes der Kategoriewahrscheinlichkeiten ausgeschöpft wird, so daß bestimmte Konstellationen von Kategoriewahrscheinlichkeiten empirisch nicht vorkommen. Ein unterparametrisiertes Modell ist also immer auch ein testbares, d.h. ein falsifizierbares Modell[19]. Voll parametrisierte und überparametrisierte Modelle sind oft — aber nicht immer — unwiderlegbar (vgl. Bamber & van Santen, 1985). Voll parametrisierte Modelle, die nicht testbar sind, d.h. jede denkbare Konstellation von Kategoriewahrscheinlichkeiten aus $\mathbf{\Gamma}$ als Funktion der Modellparameter darstellen können, nennt man auch *saturierte Modelle.*

Die Beziehungen der Begriffe „saturiert", „voll parametrisiert", „überparametrisiert" und „unterparametrisiert" zu den Begriffen der Testbarkeit und Identifizierbarkeit von parametrisierten multinomialen Modellen sind komplizierter als gemeinhin angenommen wird (Bamber & van Santen, 1985). Ich werde im folgen-

[19]Bamber und van Santen (1985) unterscheiden zwischen schwacher und starker quantitativer Testbarkeit. Die Behauptung, daß unterparametrisierte Modelle immer testbar sind, gilt nur in bezug auf schwache Testbarkeit. Sie geben ein konkretes Beispiel für ein unterparametrisiertes Modell, das zwar prinzipiell falsch sein kann, im Falle von Stichprobenvariabilität aber niemals als falsch ausgewiesen werden kann, auch wenn der Stichprobenumfang beliebig groß gewählt wird (Bamber & van Santen, 1985, S. 449, Beispiel 4). In einem solchen Fall liegt schwache, aber nicht starke quantitative Testbarkeit vor. Für die vorliegende Arbeit ist diese Unterscheidung jedoch unwichtig, da alle hier betrachteten schwach testbaren Modelle auch stark quantitativ testbar sind.

den Kapitel ausführlich hierauf zurückkommen. Vorerst reicht es, sich die Bedeutung der erstgenannten Begriffe an Beispielen anschaulich klarzumachen.

Die Signalentdeckungstheorie — angewandt auf *eine* Trefferrate und *eine* Rate falschen Alarms — ist ein Beispiel für ein voll parametrisiertes, saturiertes Modell, denn hierbei ist $S = M = 2$, und jede Konstellation von Trefferrate und Rate falschen Alarms kann als Funktion der Modellparameter dargestellt werden. Das Modell zum Loftus-Paradigma aus Abschnitt 2.6 ist dagegen ein Beispiel für ein überparametrisiertes Modell, denn hierbei gilt $S = 3$ und $M = 2$, sofern nur eine Bedingung ohne und eine Bedingung mit falscher Nachinformation in die Analyse eingehen. Schließlich ist das Modell zum Levine-Paradigma aus Abschnitt 2.5 ein Beispiel für ein saturiertes Modell mit $S = M = 15$, sofern die Fehlerwahrscheinlichkeiten a_j und b_j, $j = 1, ..., 4$, unrestringiert bleiben. Folgt man dagegen dem Vorschlag von Levine (1966), nur eine für alle Items und Hypothesen gleiche Fehlerwahrscheinlichkeit x anzusetzen, hätte das Modell $S = 8$ Parameter (neben x noch sieben freie Klassenwahrscheinlichkeiten) und wäre bei $M = 15$ freien Kategoriewahrscheinlichkeiten folglich stark unterparametrisiert.

3.4 Allgemeine multinomiale Verarbeitungsbaummodelle

Ein wichtiger Unterschied zwischen dem Modell zum Loftus-Paradigma und dem signalentdeckungstheoretischen Modell resultiert aus der Tatsache, daß die Parameter des ersteren wegen ihrer Beschränkung auf das Intervall [0, 1] potentiell als latente Wahrscheinlichkeiten interpretierbar sind, die Parameter des letzteren jedoch nicht. Die Beschränkung der Parameter auf das Intervall [0, 1] ist ein Charakteristikum der schon mehrfach angesprochenen multinomialen Verarbeitungsbaummodelle (multinomial processing tree models, vgl. Batchelder & Riefer, 1990; Batchelder & Riefer, 1999; Hu & Batchelder, 1994). Multinomiale Verarbeitungsbaummodelle (MVB-Modelle) bilden eine Teilklasse parametrisierter multinomialer Modelle, die sich zunächst einmal dadurch auszeichnet, daß der Parameterraum $\boldsymbol{\Omega}$ immer eine (nicht notwendigerweise echte) Teilmenge von $[0, 1]^S$ ist, wobei S die Gesamtzahl der Modellparameter bezeichnet.

MVB-Modelle basieren darüber hinaus auf einer Annahme, die schon auf den ersten Blick sehr gut auf das Dekomponierungsproblem in der kognitionspsychologischen Forschung zugeschnitten zu sein scheint (vgl. Abschnitt 2.6). Informell ausgedrückt besagt diese Annahme, daß Beobachtungsresultate prinzipiell immer *mehrere unterschiedliche* Ursachen haben können, und zwar derart, daß diese Ursachen häufig selbst wieder als Konjunktionen latenter, nichtbeobachtbarer Ereignisse darstellbar sind. Stellt man sich nun einen erweiterten Wahrscheinlichkeitsraum vor, dessen Stichprobenraum nicht nur die beobachtbaren Elementarereignisse, sondern alle Kombinationen von beobachtbaren und latenten Ereig-

nissen umfaßt, so kann man auch den latenten Ereignissen Wahrscheinlichkeiten zuordnen (vgl. Fußnote 20). Diese Wahrscheinlichkeiten (oder bedingten Wahrscheinlichkeiten) latenter Ereignisse bilden gerade die Parameter von MVB-Modellen.

Betrachten wir zunächst ein einfaches Beispiel, nämlich das Modell zum Loftus-Paradigma. Im Rahmen dieses Modells kann eine korrekte Wahlentscheidung für ein Originalitem im 2AFC-Rekognitionstest der Experimentalbedingung drei Ursachen haben:

(1) Das Originalitem wird korrekt klassifiziert, weil es auch in der Kontrollbedingung erinnert worden wäre (mit Wahrscheinlichkeit r) und keine Gedächtnisbeeinträchtigung durch die falsche Nachinformation hervorgerufen wurde (mit Wahrscheinlichkeit $1-m$).

(2) Das Originalitem wäre in der Kontrollgruppenbedingung erinnert worden (mit Wahrscheinlichkeit r), es liegt eine Gedächtnisbeeinträchtigung vor (mit Wahrscheinlichkeit m), aber es wird infolge korrekten Ratens mit Wahrscheinlichkeit $(1-b)$ trotzdem die richtige Wahl getroffen.

(3) Das Originalitem wäre in der Kontrollbedingung zwar nicht erinnert worden (mit Wahrscheinlichkeit $1-r$), aber es wird infolge korrekten Ratens mit Wahrscheinlichkeit $(1-b)$ dennoch die richtige Wahlantwort gegeben.

Diese drei Ursachen entsprechen disjunkten latenten Ereignissen $A_{k,j,i}$, nämlich auf Weg i, $i = 1, ..., 3$, in Versuchsbedingung k das Beobachtungsresultat $C_{k,j}$ (hier: $k = 1$ und $j = 1$) zu generieren. Da diese Wege exhaustiv sind, d.h. alle möglichen Wege zum Beobachtungsresultat $C_{1,1}$ ausschöpfen, gilt für die prognostizierte Wahrscheinlichkeit dieses Beobachtungsresultats $p_{1,1}(r, m, b) = Pr^*(A_{1,1,1}) + Pr^*(A_{1,1,2}) + Pr^*(A_{1,1,3})$.[20]

Jede Ursache, d.h. jeder Weg $A_{k,j,i}$, entspricht nun selbst wieder einer Verknüpfung (formal: einem Durchschnitt) von latenten Ereignissen, die auf unterschiedlichen Ebenen einer hierarchischen Ordnung angesiedelt sind. Man kann sich einen Weg $A_{k,j,i}$, also auch als „Ast" (branch) eines „Baumes" (tree) vorstel-

[20]Ich benutze hier die Bezeichnung Pr^* für das Wahrscheinlichkeitsmaß, um deutlich zu machen, daß ich mich an dieser Stelle nicht auf den Wahrscheinlichkeitsraum $<\mathbf{S}^{N_1} \times ... \times \mathbf{S}^{N_J}, Pr>$ für verbundene multinomiale Modelle beziehe, der in Abschnitt 3.2 eingeführt wurde. Jener ist auf beobachtbare Ereignisse beschränkt, während hier zu Erläuterungszwecken ein erweiterter Wahrscheinlichkeitsraum $<\mathbf{S}^{*N_1} \times ... \times \mathbf{S}^{*N_J}, Pr^*>$, Pr^*: Pot($\mathbf{S}^{*N_1} \times ... \times \mathbf{S}^{*N_J}$) $\rightarrow$ [0, 1], auf der Basis eines (nur informell eingeführten) Stichprobenraumes $\mathbf{S}^*$ zugrunde gelegt wird. Die Elemente von $\mathbf{S}^*$ umfassen alle Kombinationen von beobachtbaren und latenten Elementarereignissen. Wann immer die Parameter von MVB-Modellen als (bedingte) Wahrscheinlichkeiten latenter Ereignisse interpretiert werden, legt man implizit einen entsprechend erweiterten Wahrscheinlichkeitsraum zugrunde. Der Wahrscheinlichkeitsraum $<\mathbf{S}^{N_1} \times ... \times \mathbf{S}^{N_J}, Pr>$ beobachtbarer Ereignisse ergibt sich aus dem erweiterten Raum dadurch, daß man alle Elemente von $\mathbf{S}^*$, die das gleiche beobachtbare Elementarereignis aufweisen, jeweils zu einem Element von $\mathbf{S}$ vereinigt; analog erhält man Pr aus Pr^*, indem man jeweils über alle disjunkten Ereignisse aus Pot($\mathbf{S}^{*N_1} \times ... \times \mathbf{S}^{*N_J}$) summiert, die einem beobachtbaren Ereignis aus Pot($\mathbf{S}^{N_1} \times ... \times \mathbf{S}^{N_J}$) entsprechen.

len, der im Stamm beginnt — was hier der Versuchsbedingung *k* entspricht —, bei Gabelungen, die bestimmten latenten Ereignissen korrespondieren, jeweils einem der Wege folgt und schließlich irgendwo in der Baumkrone endet, d.h. in eine bestimmte Beobachtungskategorie einmündet. Diese Metapher macht die Bezeichnung „processing tree model" oder „Verarbeitungsbaummodell" plausibel, wenn man berücksichtigt, daß die „Astgabelungen" häufig als psychologische Ereignisse oder kognitive Prozesse interpretiert werden, die bei der Verarbeitung, der Speicherung oder beim Abruf bestimmter Informationen ablaufen.

Auf der untersten Ebene, die der ersten Aufspaltung des „Stammes" folgt, werden allgemein $L_{(1)}$ disjunkte und exhaustive latente Ereignisse E_{l1}, $l_1 = 1, ..., L_{(1)}$, unterschieden, deren Wahrscheinlichkeiten $\theta_1, \theta_2, ..., \theta_{L(1)}$ sich natürlich zu 1 aufaddieren müssen, so daß

$$\theta_{L_{(1)}} = 1 - \sum_{l_1=1}^{L_{(1)}-1} \theta_{l_1} \tag{3.4-1}$$

gilt. Im Falle des Loftus-Modells ist $L_{(1)} = 2$ und die beiden latenten Ereignisse sind „Originalinformation X in KB erinnert" (mit Wahrscheinlichkeit $\theta_1 = Pr^*(E_1) = r$) bzw. „Originalinformation X in KB nicht erinnert" (mit Wahrscheinlichkeit $\theta_2 = Pr^*(E_2) = 1-r$). Jedem Ereignis E_{l1} auf der ersten Stufe der Hierarchie folgen möglicherweise mehrere Ereignisse $E_{l2\,|\,l1}$ auf der zweiten Stufe der Hierarchie, $l_2 = 1, ..., L_{(2,l1)}$, wobei deren Anzahl $L_{(2,l1)}$ von E_{l1} abhängen kann. Die Ereignisse $E_{l2\,|\,l1}$ sind für gegebenes E_{l1} ebenfalls disjunkt und exhaustiv. Die bedingten Wahrscheinlichkeiten $\theta_{l2\,|\,l1} = Pr^*(E_{l2\,|\,l1}\,|\,E_{l1})$ müssen sich also für jedes Ereignis E_{l1} zu 1 aufaddieren, so daß gilt:

$$\theta_{L_{(2,l_1)}|l_1} = 1 - \sum_{l_2=1}^{L_{(2,l_1)}-1} \theta_{l_2|l_1}\ ,\ l_2 = 1,\ ...,L_{(1)}\ . \tag{3.4-2}$$

Für das Loftus-Modell ist z.B. $L_{(2,1)} = 2$ und die E_1 nachgeordneten Ereignisse sind „keine Gedächtnisbeeinträchtigung durch falsche Nachinformation" (mit bedingter Wahrscheinlichkeit $\theta_{1\,|\,1} = Pr^*(E_{1\,|\,1}\,|\,E_1) = 1-m$) sowie das Komplement dazu („Gedächtnisbeeinträchtigung durch falsche Nachinformation", mit bedingter Wahrscheinlichkeit $\theta_{2\,|\,1} = Pr^*(E_{2\,|\,1}\,|\,E_1) = m$).

Dieses einfache Prinzip setzt sich auf prinzipiell beliebig vielen Hierarchieebenen fort, wobei es durchaus zugelassen ist (und praktisch sogar häufig vorkommt), daß bestimmte Ereignisse auf den nachfolgenden Ebenen jeweils nur ein einziges Ereignis mit bedingter Wahrscheinlichkeit 1 nach sich ziehen. Praktisch bedeutet das, daß einige Hierarchieebenen bei Realisation bestimmter latenter Ereignisse einfach fortfallen, im Loftus-Modell z.B. bei Realisation von E_2.

Das letzte Ereignis in einem solchen Hierarchiebaum zur Versuchsbedingung k mit insgesamt H_k Ebenen ist jeweils kein latentes Ereignis, sondern eines der möglichen Beobachtungsresultate $C_{k,j}$, $j = 1, ..., J_k$. Falls eine bestimmte Sequenz E_{l1}, $E_{l2 \mid l1}$, $E_{l3 \mid l1,l2}$, ..., $E_{l(Hk\text{-}1) \mid l1,l2, ..., l(Hk\text{-}2)}$ latenter Ereignisse, d.h. ein bestimmter „Ast" des Baumes gegeben ist, wird das Beobachtungsresultat $C_{k,j}$ auf der höchsten Ebene mit der bedingten Wahrscheinlichkeit $\theta_{j \mid l1,l2,...,l(Hk\text{-}1)} = Pr^*(C_{k,j} \mid E_{l1} \cap E_{l2 \mid l1} \cap ... \cap E_{l(Hk\text{-}1) \mid l1,l2, ..., l(Hk\text{-}2)})$ erreicht. Im Falle des Loftus-Modells ist die Anzahl der Hierarchieebenen in der Experimentalbedingung beispielsweise $H_1 = 3$ und die Wahlentscheidungen X bzw. Y bilden die möglichen Beobachtungsresultate. Die bedingte Wahrscheinlichkeit einer korrekten Wahl im 2AFC-Test, gegeben eine korrekte Erinnerung von X in der KB *und* eine falschinformationsbedingte Gedächtnisbeeinträchtigung ist — um nur ein Beispiel zu nennen — $\theta_{1 \mid 1,2} = Pr^*(C_{1,1} \mid E_1 \cap E_{2,1}) = 1\text{–}b$.

Auf dem Hintergrund dieser Strukturannahmen läßt sich die Wahrscheinlichkeit eines bestimmten Weges (oder „Astes") $A_{k,j,i}$ zu einem Beobachtungsresultat $C_{k,j}$ recht einfach als Funktion der Modellparameter ausdrücken. Da der Weg-Index i durch die Indizes $l_1, l_2, ..., l_{(Hk-1)}$ der involvierten latenten Ereignisse festgelegt wird, kann man den Weg $A_{k,j,i}$ auch ausführlicher als $A_{k,j,i(l1,l2,...,l(Hk\text{-}1))}$ notieren. Bei der Berechnung von $Pr^*(A_{k,j,i}) = Pr^*(A_{k,j,i(l1,l2,...,l(Hk\text{-}1))})$ macht man sich nun einfach zunutze, daß die Wahrscheinlichkeit einer Schnittmenge von Ereignissen aufgrund der Definition bedingter Wahrscheinlichkeiten auch als Produkt von unbedingten und bedingten Wahrscheinlichkeiten dieser Ereignisse geschrieben werden kann:

$$
\begin{aligned}
&Pr(A_{k,j,i(l_1,l_2,...,l_{(H_k-1)})}) = Pr(E_{l_1} \cap E_{l_2|l_1} \cap ... \cap E_{l(H_k\text{-}1)|l_1,l_2,...,l_{(H_k\text{-}2)}} \cap C_{k,j}) = \\
&= Pr(E_{l_1}) \cdot Pr(E_{l_2|l_2} \,|\, E_{l_1}) \cdot ... \cdot Pr(C_{k,j} \,|\, E_{l_1} \cap E_{l_2|l_1} \cap ... \cap E_{l_{(H_k-1)}|l_1,l_2,...,l_{(H_k\text{-}2)}}) = \quad (3.4\text{-}3) \\
&= \theta_{l_1} \cdot \theta_{l_2|l_1} \cdot \ ... \ \cdot \theta_{j|l_1,l_2,...,l_{(H_k-1)}} \cdot
\end{aligned}
$$

Wenn eine theoretisch vorgegebene hierarchische Ordnung der latenten Ereignisse existiert, kann man also die Wahrscheinlichkeit eines Weges als Produkt der Wahrscheinlichkeiten des ersten Ereignisses, $Pr^*(E_{l1})$, des zweiten Ereignisses gegeben das erste Ereignis, $Pr^*(E_{l2} \mid E_{l1})$, usw. schreiben, wobei der letzte Term die Wahrscheinlichkeit einer bestimmten Beobachtungskategorie gegeben alle vorgeordneten Ereignisse ist. Die theoretisch vorgegebene Ordnung kann z.B. eine hypothetisch angenommene Reihenfolge sequentiell aufeinander folgender Ereignisse sein. Die zeitliche Anordnung der latenten Ereignisse ist jedoch nur ein mögliches Ordnungsprinzip. Auch andere Ordnungsprinzipien (Ähnlichkeit, Bedeutsamkeit usw.) sind prinzipiell denkbar. Insbesondere muß ein serielles Verarbeitungsmodell nicht zwangsläufig vorausgesetzt werden. MVB-Modelle sind potentiell auch auf parallel ablaufende Prozesse anwendbar.

Im Loftus-Modell ist beispielsweise das Erinnern der Originalinformation ein vorgeordnetes Ereignis mit unbedingter Wahrscheinlichkeit $Pr^*(E_1) = r$, die mögliche Gedächtnisbeeinträchtigung ein nachgeordnetes Ereignis erster Ordnung mit bedingter Wahrscheinlichkeit $Pr^*(E_{2|1} \mid E_1) = m$ und das Erraten der Originalinformation bei nachträglicher Falschinformation ein nachgeordnetes Ereignis zweiter Ordnung mit bedingter Wahrscheinlichkeit $Pr^*(E_{1|2,1} \mid E_1 \cap E_{2|1}) = (1-b)$. Man mag sich hierbei $E_{1|2,1}$ als den anderen beiden Ereignissen zeitlich nachgeordnet vorstellen: Erst wenn keine Gedächtnisspur abrufbar ist, wird der Rate- oder Rekonstruktionsprozeß initiiert, der mit Wahrscheinlichkeit b zur falschen und mit Wahrscheinlichkeit $(1-b)$ zur richtigen Antwort führt. Die Ereignisse E_1 und $E_{2|1}$ wird man dagegen sicher nicht als zeitlich aufeinander folgend auffassen. Vielmehr ist die Nachgeordnetheit von $E_{2|1}$ nur Ausdruck der Tatsache, daß wir uns für die *bedingte* Wahrscheinlichkeit einer Gedächtnisbeeinträchtigung interessieren, gegeben daß das betreffende Item in der KB abrufbar gewesen wäre. Der Parameter m ist somit als *Anteil* der in der KB erinnerbaren Originalinformationen interpretierbar, die aufgrund einer Gedächtnisbeeinträchtigung durch nachträgliche Falschinformationen in der EB nicht abrufbar sind. Diese Interpretation wäre nicht möglich, wenn die Gedächtnisbeeinträchtigung auf der ersten Ebene des Verarbeitungsbaumes angesiedelt worden wäre, weil dann der zugeordnete latente Parameter als unbedingte Wahrscheinlichkeit zu interpretieren wäre.

Aus den Ausführungen ergibt sich die Wahrscheinlichkeit des ersten der o.g. Wege als $Pr^*(A_{1,1,1}) = Pr^*(E_1) \cdot Pr^*(E_{1|1} \mid E_1) \cdot Pr^*(C_{1,1} \mid E_1 \cap E_{1|1}) = r \cdot (1-m) \cdot 1$, des zweiten als $Pr^*(A_{1,1,2}) = Pr^*(E_1) \cdot Pr^*(E_{2|1} \mid E_1) \cdot Pr^*(C_{1,1} \mid E_1 \cap E_{2|1}) = r \cdot m \cdot (1-b)$ und schließlich des dritten als $Pr^*(A_{1,1,3}) = Pr^*(E_2) \cdot Pr^*(C_{1,1} \mid E_2) = (1-r) \cdot (1-b)$. Durch deren Summierung erhält man dann die Modellgleichung für $p_{1,1}(r, m, b)$, nämlich die bereits erwähnte Gleichung (2.6-2). Die Modellgleichungen für andere Kategoriewahrscheinlichkeiten folgen dem gleichen Prinzip.

Wichtig für das korrekte Verständnis von MVB-Modellen ist die Tatsache, daß die meisten Modellparameter — abgesehen von den Wahrscheinlichkeiten der vorgeordneten („ersten") Ereignisse E_{l1} — als *bedingte* Wahrscheinlichkeiten erster oder höherer Ordnung zu interpretieren sind. Erst auf dem Hintergrund dieses Verständnisses wird klar, daß man keine stochastische Unabhängigkeitsannahme benötigt, um die Wahrscheinlichkeiten der Wege zu bestimmten Beobachtungskategorien als Produkte latenter Wahrscheinlichkeiten schreiben zu können. Gleichung (3.4-3) folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit und erfordert sonst keine weiteren Zusatzannahmen. Eine stochastische Unabhängigkeitsannnahme wäre nur dann vonnöten, wenn man z.B. die $\theta_{l2|l1}$-Parameter als *unbedingte* Wahrscheinlichkeiten von latenten Ereignissen interpretieren will. Im nächsten Kapitel wird besprochen, wie man derartige stochasti-

sche Unabhängigkeitsannahmen in Form bestimmter Parameterrestriktionen formalisieren und testen kann (vgl. Abschnitt 4.3.2).

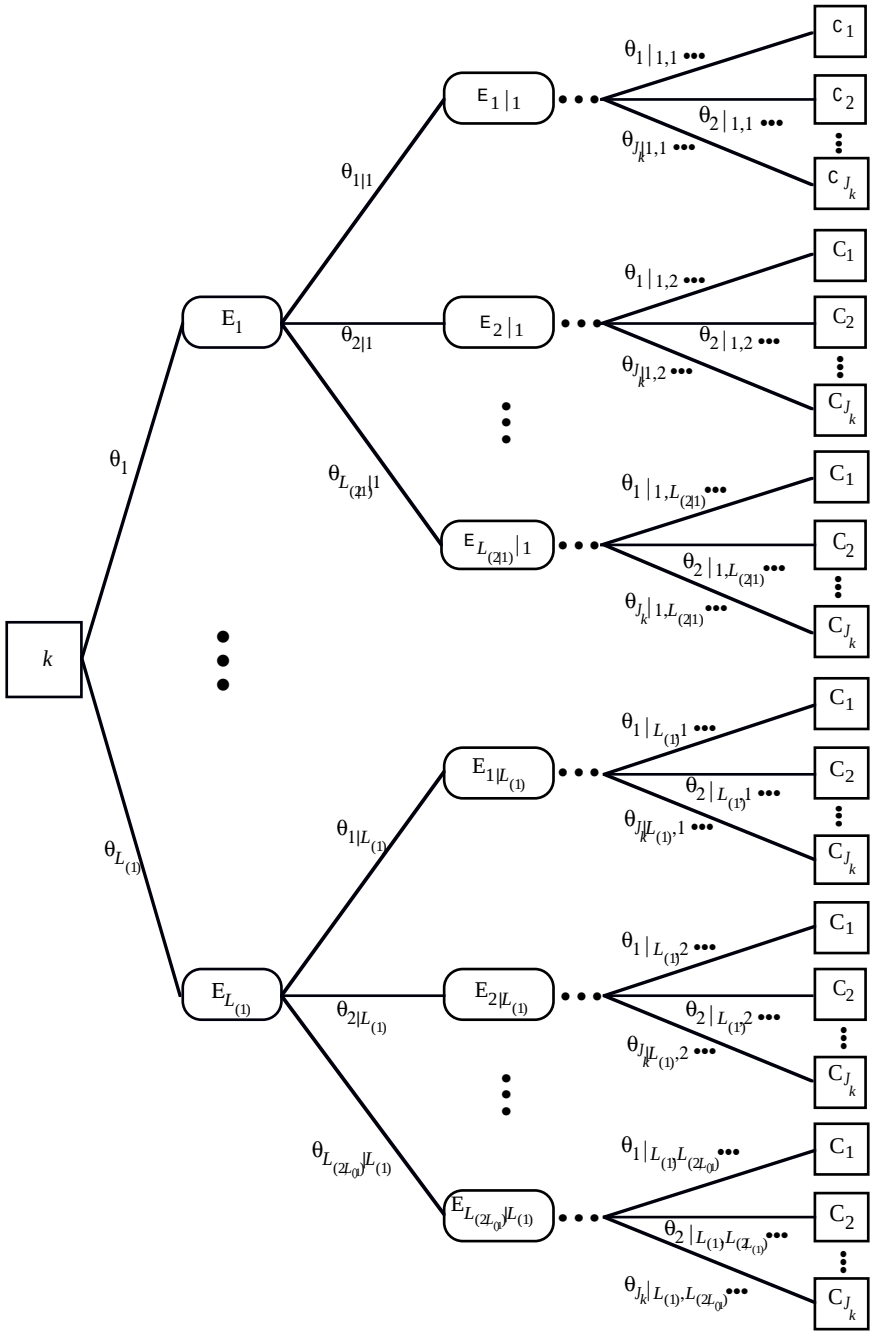

Abbildung 3.1:
Die allgemeine Struktur eines einfachen (nichtverbundenen) multinomialen Verarbeitungsbaummodells bzw. eines verbundenen multinomialen Verarbeitungsbaummodells für eine bestimmte Versuchsbedingung k mit J_k Beobachtungskategorien $C_{k,j}$. Die Symbole und Parameter werden im Text erläutert.

Nach diesen Vorbemerkungen kann nun die allgemeine Struktur von MVB-Modellen besprochen werden. Dazu ist es hilfreich, sich diese allgemeine Struktur zunächst in Form eines Baumdiagramms zu veranschaulichen. Abbildung 3.1 leistet dies für ein einfaches MVB-Modell bzw. für eine bestimmte Versuchsbedingung k eines verbundenen MVB-Modells. Verbundene MVB-Modelle

können also einfach durch mehrere Baumdiagramme der in Abbildung 3.1 dargestellten Art repräsentiert werden, wie dies in Abbildung 2.1 für den Spezialfall des Loftus-Modells veranschaulicht wird.

Die latenten Ereignisse werden im Baumdiagramm durch Felder mit abgerundeten Ecken, beobachtbare Ereignisse — d.h. die Versuchsbedingung k und die Beobachtungsresultate $C_{k,j}$, $j = 1, ..., J_k$, — dagegen durch rechteckige Felder repräsentiert. An den Verbindungslinien stehen die latenten Parameter des Modells, d.h. jeweils die Übergangswahrscheinlichkeiten vom Ereignis am linken Ende einer Verbindungslinie zum Ereignis am rechten Ende der Verbindungslinie.

MVB-Modelle sind parametrisierte multinomiale Modelle, die durch den Parameterraum $\mathbf{\Omega} \subseteq [0, 1]^S$ und eine bestimmte Abbildung $f_{\mathrm{MVB}}: \mathbf{\Omega} \rightarrow \mathbf{\Gamma}$ des Parameterraumes in den Raum der Kategoriewahrscheinlichkeiten ausgezeichnet sind. Die Abbildung f_{MVB} ergibt sich ganz einfach dadurch, daß man für jede Kategorie $C_{k,j}$ zunächst alle Äste im Baumdiagramm 3.1 identifiziert, die in die betreffende Kategorie einmünden, die Wahrscheinlichkeiten dieser Äste sodann in Termini der Modellparameter ausdrückt und dann über alle diese Äste summiert:

$$
\begin{aligned}
&\forall k, \forall j\colon p_{k,j}(\boldsymbol{\theta}) = \\
&= \sum_{l_1=1}^{L_{(1)}} \sum_{l_2=1}^{L_{(2|l_1)}} \cdots \sum_{l_{H_{(k-1)}}=1}^{L_{(H_{(k-1)}|l_1,l_2,\ldots)}} Pr^*(E_{l_1} \cap \ldots \cap E_{l_{H_{(k-1)}}|l_1,l_2,\ldots}) \cdot \\
&\qquad \cdot Pr^*(C_{k,j}|E_{l_1} \cap \ldots \cap E_{l_{H_{(k-1)}}|l_1,l_2,\ldots}) = \qquad (3.4\text{-}4) \\
&= \sum_{l_1=1}^{L_{(1)}} \theta_{l_1} \cdot \sum_{l_2=1}^{L_{(2|l_1)}} \theta_{l_2|l_1} \cdot \ldots \cdot \sum_{l_{H_{(k-1)}}}^{L_{(H_{(k-1)}|l_1,l_2,\ldots)}} \theta_{l_{H_{(k-1)}}|l_1,l_2,\ldots,l_{H_{(k-2)}}} \cdot \theta_{j|l_1,l_2,\ldots,l_{H_{(k-1)}}} \cdot
\end{aligned}
$$

Die Modellgleichung (3.4-4) hat die gleiche Struktur wie die Modellgleichung der Latent-Class-Analyse (LCA) für eine beobachtete Variable (zur LCA vgl. z.B. Goodman, 1974; Langeheine, 1988; Langeheine & Rost, 1996; Lazarsfeld & Henry, 1968). Die latenten Klassen entsprechen hierbei den verschiedenen Ästen eines MVB-Modells und die Klassenwahrscheinlichkeiten folglich den Astwahrscheinlichkeiten $Pr^*(E_{l1} \cap \ldots \cap E_{lH(k\text{-}1)\,|\,l1,\,l2,\,\ldots})$. Diese werden mit den bedingten Wahrscheinlichkeiten eines Beobachtungsresultats, gegeben ein bestimmter Ast (d.h. $Pr^*(C_{k,j}\,|\,E_{l1}, E_{l2}, \ldots, E_{lH(k-1)})$), multipliziert und dann über alle Äste (bzw. latenten Klassen) des Modells summiert. Das Ergebnis ist die Kategoriewahrscheinlichkeit $p_{k,j}(\boldsymbol{\theta})$. Der Unterschied zur „normalen" LCA besteht nur darin, daß (a) statt mehrerer beobachteter Variablen, die innerhalb der latenten Klassen stochastisch unabhängig sind, nur eine beobachtete Variable vorliegt, und daß (b) die Klassenwahrscheinlichkeiten in MVB-Modellen typischerweise restringiert sind, weil die Modellparameter θ_{l1}, $\theta_{l2\,|\,l1}$ usw., aus denen sich die Klassenwahrscheinlichkeiten

durch Produktbildung ergeben, in der Regel durch Parametergleichsetzungen oder Parameterfixierungen a priori auf bestimmte Werte festgelegt sind. Man kann also MVB-Modelle auch als „restringierte univariate Latent-Class-Modelle" auffassen (Erdfelder, 1990b), was insofern Vorteile hat, als damit das gesamte statistische Instrumentarium, das zur Parameterschätzung und zur Prüfung spezieller Hypothesen in der LCA entwickelt wurde, prinzipiell auch auf MVB-Modelle anwendbar ist. Ich werde im nächsten Kapitel darauf zurückkommen.

Die Charakterisierung von MVB-Modellen über Modellgleichungen in der Form (3.4-4) hat den Vorteil, einerseits die Beziehung zwischen MVB- und LCA-Modellen und andererseits die Interpretation der Modellparameter als bestimmte bedingte oder unbedingte Wahrscheinlichkeiten latenter Ereignisse unmittelbar deutlich werden zu lassen. Gleichzeitig hat sie jedoch den Nachteil, recht umständlich zu sein, da durch die Notation mehr Parameter eingeführt werden, als de facto frei variieren können. Für das Modell zum Loftus-Paradigma wird z.B. in der Experimentalbedingung ein Baum mit insgesamt $H_1 = 3$ Ebenen benötigt. Auf der ersten Ebene werden $L_{(1)} = 2$ latente Ereignisse E_1 („X in KB erinnert") und E_2 („X in KB nicht erinnert") und folglich auch zwei Parameter θ_1 und θ_2 eingeführt. Auf der zweiten Ebene werden konditional zu E_1 zwei Folgeereignisse $E_{1|1}$ und $E_{2|1}$ mit bedingten Wahrscheinlichkeiten $\theta_{1|1}$ und $\theta_{2|1}$ eingeführt; konditional zu E_2 wird ein (nicht explizit aufgeführtes) Folgeereignis $E_{1|2}$ mit bedingter Wahrscheinlichkeit $\theta_{1|2} = 1$ benötigt, um die fehlende zweite Hierarchieebene zu ergänzen. Auf der dritten Ebene folgen schließlich die beiden Beobachtungskategorien $C_{1,1}$ („Wahl von X") und $C_{1,2}$ („Wahl von Y"), für die pro Ast des Baumes erneut zwei bedingte Wahrscheinlichkeiten benötigt werden, bei drei Ästen also $\theta_{1|1,1}$, $\theta_{2|1,1}$, $\theta_{1|1,2}$, $\theta_{2|1,2}$, $\theta_{1|2,1}$ und $\theta_{2|2,1}$. Insgesamt ergeben sich 2 + 3 + 6 = 11 Parameter im Baumdiagramm. Nur drei davon sind jedoch funktional unabhängig, denn es muß gelten:

(1) $\theta_1 = 1-\theta_2 = r$,

(2) $\theta_{1|1} = 1-\theta_{2|1} = m$,

(3) $\theta_{2|1,2} = \theta_{2|2,1} = 1-\theta_{1|1,2} = 1-\theta_{1|2,1} = b$ und

(4) $\theta_{1|2} = \theta_{1|1,1} = 1-\theta_{2|1,1} = 1$.

Um diese Parameterredundanz zu eliminieren, hat Hu (1998; 1999a) eine kompaktere Notation für die Modellgleichungen vorgeschlagen, die hier in leichter Abwandlung wiedergegeben werden soll. Grundgedanke dieser Notation ist, die Parameter in G Gruppen mit Gruppenindex $g = 1, ..., G$ einzuteilen, und zwar derart, daß sich die Parameter pro Gruppe zu 1 aufaddieren müssen. Pro Gruppe g gibt es $L_g \geq 1$ freie Parameter $\theta_{g,l}$, $l = 1, ..., L_g$, so daß der letzte (nicht freie) Parameter in Gruppe g immer

$$1-\sum_{l=1}^{L_g}\theta_{g,l} \tag{3.4-5}$$

sein muß.[21] Der Vorteil dieser Parameternotation mit zwei Indizes *g* und *l* ist, daß nur noch die Parameter eingeführt werden, die auch wirklich funktional unabhängig sind; der Nachteil ist, daß der Notation nicht mehr angesehen werden kann, ob es sich bei dem betreffenden Parameter um eine unbedingte oder eine spezielle bedingte Wahrscheinlichkeit handelt.

Im Loftus-Modell bilden die Parameter $G = 3$ Gruppen mit jeweils einem freien Parameter ($L_1 = L_2 = L_3 = 1$), nämlich $\theta_{1,1} = r$, $\theta_{2,1} = m$ und $\theta_{3,1} = b$. Diese Parameter sollen „freie Parameter des Modells " heißen. Neben $\theta_{1,1}$, $\theta_{2,1}$ und $\theta_{3,1}$ gibt es in dem Modell nur noch restringierte Parameter. Eine Gruppe restringierter Parameter, nämlich die latenten Wahrscheinlichkeiten, welche die Gruppenwahrscheinlichkeiten zu 1 komplettieren ($1-\theta_{1,1} = 1-r$, $1-\theta_{2,1} = 1-m$ und $1-\theta_{3,1} = 1-b$), sollen „Komplettierungsparameter des Modells" genannt werden. Die verbleibenden restringierten Parameter heißen „fixierte Parameter des Modells", weil sie a priori auf bestimmte Werte fixiert sind, wie z.B. die auf 0.5 festgelegten Ratewahrscheinlichkeiten für die beiden Wahlalternativen im 2AFC-Rekognitionstest der KB.

Eine zweite Besonderheit der Notation von Hu (1998; 1999a) besteht darin, daß nicht mehr über alle Äste des Modells summiert wird, sondern pro Beobachtungskategorie $C_{k,j}$ nur noch über nichtvernachlässigbare Äste mit einer Wahrscheinlichkeit von $Pr^*(A_{k,j,i}) > 0$. Wir bezeichnen die Anzahl der nichtvernachlässigbaren Äste, die in Versuchsbedingung *k* in die Beobachtungskategorie $C_{k,j}$ einmünden mit $I_{k,j}$. Folglich ist

$$Pr(C_{k,j}) = p_{k,j} = \sum_{i=1}^{I_{k,j}} Pr^*(A_{k,j,i}) \; . \tag{3.4-6}$$

Der nächste Schritt besteht nun darin, pro Ast $A_{k,j,i}$ des Baumes auszuzählen, wie häufig die einzelnen Parameter des Modells (freie Parameter, Komplettierungsparameter und fixierte Parameter) in dem betreffenden Ast vorkommen. Die Zählvariable $a_{k,j,i,g,l}$ gibt an, wie oft der Parameter $\theta_{g,l}$, d.h. der *l*-te freie Parameter der Parametergruppe *g*, im *i*-ten Ast zur Beobachtungskategorie $C_{k,j}$ vorkommt. Kommt der Parameter in dem betreffenden Ast gar nicht vor, nimmt die Zählvariable den Wert 0 an, kommt er einmal vor, ist sie 1 usw. Analog wird in der Zählvariablen $b_{k,j,i,g}$ erfaßt, wie oft die Komplettierungswahrscheinlichkeit $1-\Sigma_l$

[21] Man beachte, daß die Parameterindizes in dieser Notation von Hu eine andere Bedeutung haben als in der zuvor dargestellten Notation. Der Laufindex *l* ist nicht mehr einer bestimmten Hierarchieebene im Modell zugeordnet, sondern läuft über alle Parameter einer Parametergruppe *g*, unabhängig von der Hierarchieebene, auf der die betreffenden Parameter angesiedelt sind.

$\theta_{g,l}$ der Parametergruppe g im i-ten Ast zur Beobachtungskategorie $C_{k,j}$ repräsentiert ist. Schließlich werden alle a priori fixierten Parameter des Astes miteinander multipliziert, und ihr Produkt wird in der Hilfsvariable $c_{k,j,i}$ zusammengefaßt. Diese Hilfsvariable kann im Gegensatz zu $a_{k,j,i,g,l}$ und $b_{k,i,g}$ nicht nur Werte aus dem Bereich der um null erweiterten natürlichen Zahlen, sondern beliebige reelle Werte annehmen.

Man kann nun mit diesen drei Gruppen von Hilfsvariablen jede Astwahrscheinlichkeit ganz einfach und kompakt als Funktion *aller* Modellparameter darstellen. Durch Summation über die Äste einer Kategorie gemäß Gleichung (3.4-6) ergibt sich dann für die Kategoriewahrscheinlichkeiten:

$$\begin{aligned} \forall k, \forall j\colon p_{k,j}(\boldsymbol{\theta}) &= \sum_{i=1}^{I_{k,j}} Pr^{*}(\mathrm{A}_{k,j,i}) = \\ &= \sum_{i=1}^{I_{k,j}} c_{k,j,i} \cdot \prod_{g=1}^{G} \left[\left(1 - \sum_{l=1}^{L_g} \theta_{g,l} \right)^{b_{k,j,i,g}} \cdot \prod_{l=1}^{L_g} \left(\theta_{g,l} \right)^{a_{k,j,i,g,l}} \right] \end{aligned} \qquad (3.4\text{-}7)$$

Gleichung (3.4-7) ist die Modellgleichung für allgemeine MVB-Modelle (general processing tree models with non-binary parameters) in der von Hu (1998; 1999a) vorgeschlagenen Form. Jede Astwahrscheinlichkeit ist dabei als Produkt von Monomen repräsentiert, d.h. von Termen, die jeweils einen mit einem bestimmten Exponenten $a_{k,j,i,g,l}$ oder $b_{k,j,i,g}$ versehenen Modellparameter enthalten; die Kategoriewahrscheinlichkeiten sind somit Polynome in den Modellparametern und ihren Komplementen. In den meisten Ästen wird der einem Parameter zugeordnete Exponent gleich null sein (d.h. der Parameter kommt in dem jeweiligen Ast nicht vor), so daß der betreffende Term den Wert 1 annimmt und sich bei der Produktbildung nicht auswirkt. Auf diese Weise wird erreicht, daß nur die Parameter in die Produktbildung eingehen, die in dem betreffenden Ast repräsentiert sind.

Zur Verdeutlichung sind in Tabelle 3.1 die entsprechenden Koeffizienten für das Loftus-Modell zusammengestellt. Die Zeilen der Tabelle entsprechen dabei den verschiedenen Ästen des Modells, und zwar sowohl für die EB (k = 1) als auch für die KB (k = 2). Wie man sieht, kommen in diesem Fall für die verschiedenen a- und b-Exponenten nur die Werte 0 und 1 und für die c-Koeffizienten nur die Werte 1 oder 0.5 vor. Das ist nicht untypisch für MVB-Modelle, weil jeder Parameter in einem Ast normalerweise entweder gar nicht oder nur einmal vorkommt. Allerdings gibt es auch Ausnahmen von dieser Regel. Im nächsten Abschnitt wird z.B. ein wichtiges MVB-Modell vorgestellt, das Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986), in dem einige Parameter pro Ast auch mehrmals vorkommen.

Tabelle 3.1:
Werte der Variablen $a_{k,j,i,g,l}$, $b_{k,j,i,g}$ und $c_{k,j,i}$ der allgemeinen MVB-Modellgleichung (3.4-7) für das Modell zum Loftus-Paradigma gemäß Abbildung 2.1. Die Laufindex k indiziert die EB (k = 1) und die KB (k = 2) des Modells, j indiziert die korrekte X-Wahl (j = 1) bzw. die inkorrekte Y-Wahl (j = 2) im 2AFC-Rekognitionstest, i ist die Nummer des Astes in eine bestimmte Beobachtungskategorie (wobei — bezogen auf Abbildung 2.1 — die Wege von oben nach unten durchgezählt werden), g der Index für die Parametergruppe (r in g = 1, m in g = 2 und b in g = 3) und l der (bei diesem Modell redundante) Laufindex für die Nummer des freien Parameters in einer Gruppe.

			$a_{k,j,i,g,l}$			$b_{k,j,i,g}$			$c_{k,j,i}$
k	j	i	g = 1, l = 1	g = 2, l = 1	g = 3, l = 1	g = 1	g = 2	g = 3	
1	1	1	1	0	0	0	1	0	1
1	1	2	1	1	0	0	0	1	1
1	1	3	0	0	0	1	0	1	1
1	2	1	1	1	1	0	0	0	1
1	2	2	0	0	1	1	0	0	1
2	1	1	1	0	0	0	0	0	1
2	1	2	0	0	0	1	0	0	0.5
2	2	1	0	0	0	1	0	0	0.5

Die für allgemeine MVB-Modelle vorgesehene Struktur ist für das Loftus-Modell unnötig komplex, da in jeder Parametergruppe immer nur ein freier Parameter vorkommt. Man kann in solchen Fällen die Modellgleichung (3.4-7) vereinfachen, wie im nächsten Abschnitt gezeigt werden soll. Die allgemeinere Struktur (3.4-7) wird jedoch als Rahmenmodell für eine Reihe anderer Modelle benötigt, z.B. für das in Abschnitt 2.5 vorgeschlagene Modell zu Blank-Trial-Prozedur von Levine (1966). Bei diesem Modell haben wir es in der allgemeinsten Version mit G = 9 Parametergruppen zu tun, nämlich vier Gruppen für die a_j- Fehlerwahrscheinlichkeiten, vier Gruppen für die b_j-Fehlerwahrscheinlichkeiten und eine Gruppe für die acht Klassenwahrscheinlichkeiten $p(h_i)$. Während die ersten acht Gruppen jeweils nur einen freien Parameter umfassen (L_g = 1, g = 1, ..., 8), beinhaltet die letztgenannte Gruppe L_9 = 7 freie Parameter. Es gibt für diese Parametergruppe also sieben Zählvariablen $a_{k,j,i,9,l}$, l = 1, ..., 7, statt nur einer wie bei den anderen Parametergruppen. Abgesehen von dieser Besonderheit kann man auch das Modell zum Levine-Paradigma analog zum Loftus-Modell in der Form der allgemeinen Modellgleichung (3.4-7) notieren.

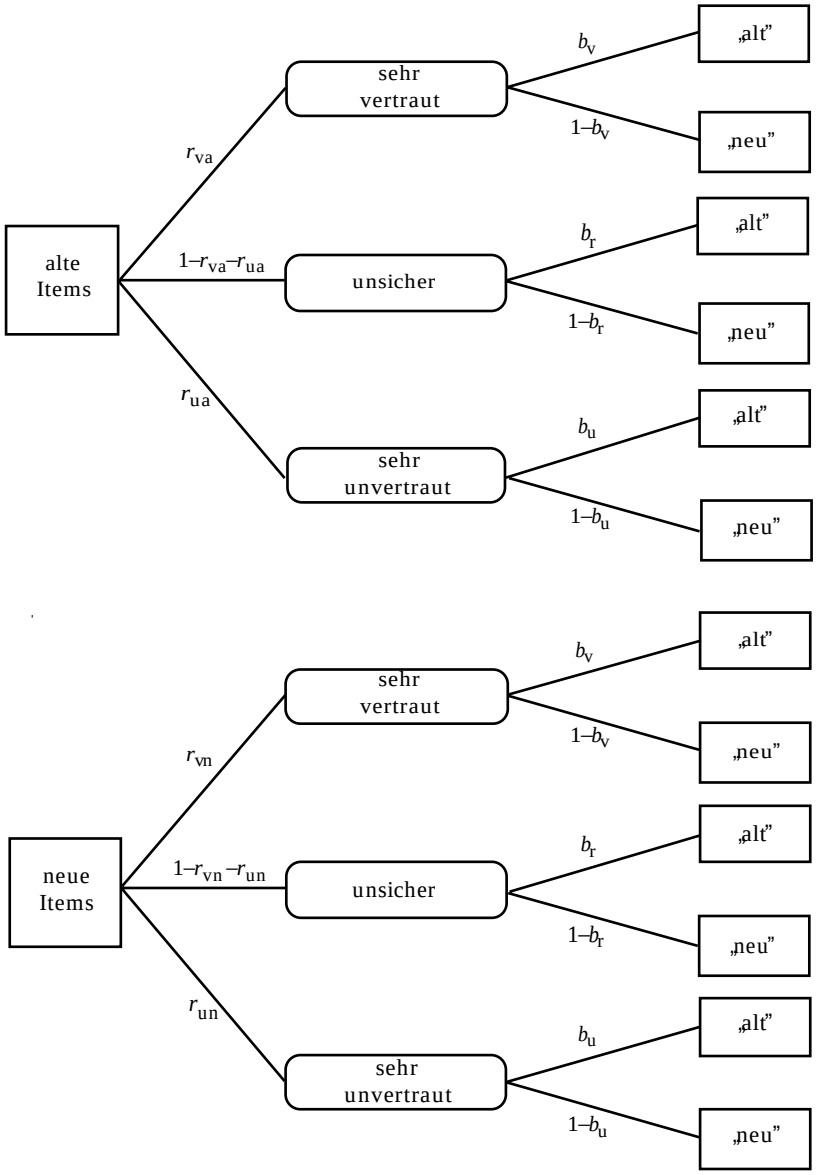

Abbildung 3.2:
Das allgemeine Schwellenmodell für Ja-Nein-Rekognitionstests oder Ja-Nein-Reizentdeckungsurteile.

Abbildung 3.2 stellt ein weiteres wichtiges MVB-Modell vor, das Parametergruppen mit mehr als einem freien Parameter umfaßt. Es handelt sich um das allgemeine Schwellenmodell (general threshold model) für Ja-Nein-Rekognitionstests. Das Modell kann als Alternative zur Signalentdeckungstheorie (SDT) aufgefaßt werden; im Gegensatz zur SDT geht das allgemeine Schwellenmodell nicht von einer kontinuierlichen, stetig verteilten Vertrautheitsvariablen aus, sondern von einer diskreten (finite-state) Variablen, die nur die drei Werte „sehr vertraut", „unsicher" und „sehr unvertraut" annehmen kann. Man kann die Idee einer kontinuierlichen Vertrautheitsvariablen mit der finite-state-Auffassung verknüpfen, indem man annimmt, daß (a) auf dem Vertrautheitskontinuum zwei Schwellenwerte definiert sind, welche die kontinuierliche Vertrautheit in eine diskrete Variable mit drei Ausprägungen überführen, und daß (b) das Antwortverhalten der

Vpn nur von der Ausprägung dieser diskreten Variablen abhängt. Aus dieser Interpretation leitet sich die Bezeichnung „Schwellenmodell" her. Tatsächlich lassen sich diverse Spezialfälle dieses allgemeinen Schwellenmodells alternativ auch aus einem verallgemeinerten SDT-Modell herleiten, das statt normalverteilter Vertrautheitsvariablen andere Dichteverteilungen annimmt (vgl. Abschnitt 6.7). Allerdings lassen sich diese modifizierten Dichteverteilungen in aller Regel psychologisch nur schlecht begründen (vgl. z.B. Macmillan & Creelman, 1991).

Das Modell umfaßt in seiner allgemeinsten Form $G = 5$ Parametergruppen. Die ersten beiden Gruppen sind die unbedingten Wahrscheinlichkeiten der drei Vertrautheitskategorien, für die $L_1 = 2$ freie Parameter r_{va} und r_{ua} bei alten Items und außerdem $L_2 = 2$ freie Parameter r_{vn} und r_{un} bei neuen Items vorgesehen sind. Darüber hinaus gibt es drei Gruppen mit jeweils einem freien Parameter, nämlich die Wahrscheinlichkeiten der „Alt"-Antworten innerhalb der drei Vertrautheitsstufen (b_v, b_r und b_u). Das Modell sieht vor, daß diese Parameter nur von der Vertrautheit abhängen, nicht aber davon, ob die betreffende Vertrautheitsstufe durch ein altes oder ein neues Item hervorgerufen wird; deshalb sind die entsprechenden Parameter in den beiden Baumdiagrammen gleichgesetzt.

Ebenso wie die Signalentdeckungstheorie kann auch das allgemeine Schwellenmodell nicht nur auf Rekognitionsurteile, sondern auch auf Reizerkennungsurteile in der Wahrnehmungspsychologie angewandt werden. Alte bzw. neue Items sind dann als Darbietungen von „Signal" bzw. „kein Signal (Rauschen)", Vertrautheitsintensitäten als Empfindungsintensitäten und „Alt"- bzw. „Neu"-Urteile als „Ja"- bzw. „Nein"-Urteile zu interpretieren. Formal ändert sich an dem Modell sonst nichts.

Das Schwellenmodell in Abbildung 3.2 ist insofern „allgemein", als es alle in der Literatur vorgeschlagenen Hoch- und Niedrigschwellen-Modelle (high- und low-threshold models) für Ja-Nein-Rekognitions- oder -Detektionstests als Spezialfälle umfaßt (vgl. z.B. Blackwell, 1953; 1963; Green & Birdsall, 1978; Kaernbach, 1991; Krantz, 1969; Luce, 1963a; 1963b).[22] In der Praxis spielt das allgemeine Modell ansonsten eine unbedeutende Rolle, da es mit einem siebendimensionalen Parameterraum bei einem zweidimensionalen Raum möglicher Kategoriewahrscheinlichkeiten stark überparametrisiert und deshalb nicht identifizierbar ist (siehe dazu Abschnitt 4.1). Nur durch bestimmte Parameterrestriktionen, die zu den o.g. Hoch-

[22]Man spricht von einem Hochschwellen-Modell (high-threshold model), wenn die obere der beiden Schwellen einen Zustand hoher Vertrautheit definiert, der nur von alten Items erreicht werden kann, oder wenn die untere Schwelle einen Zustand sehr niedriger Vertrautheit festlegt, der nur von neuen Items erreicht werden kann. Ist beides der Fall, spricht man auch von einem Zwei-Hochschwellen-Modell (two-high oder double-high threshold model, vgl. Macmillan & Creelman, 1990; 1991; Snodgrass & Corwin, 1988). Kann eine Schwelle prinzipiell von Items beider Itemklassen überschritten werden, so heißen die entsprechenden Modelle Niedrigschwellen-Modelle (low-threshold models, vgl. Luce, 1963b).

oder Niedrigschwellen-Modellen führen, werden die verbleibenden Modellparameter statistisch schätzbar. Einige dieser Modelle (bzw. Varianten davon) werden uns in den folgenden Kapiteln wiederholt begegnen.

Nach diesen Erläuterungen kann nun definiert werden, was unter einem allgemeinen multinomialen Verarbeitungsbaummodell (MVB-Modell) bzw. einem „general processing tree model with non-binary parameters" (Hu, 1999a) verstanden werden soll. Es handelt sich ganz einfach um ein parametrisiertes multinomiales Modell $\langle M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$ gemäß (3.3-5), das zwei Zusatzbedingungen erfüllt:

a) Der Parameterraum $\boldsymbol{\Omega}$ hat die Form

$$\boldsymbol{\Omega}_{\text{MVB}} = \left\{ \boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{1,L_1}, \ldots, \theta_{G,1}, \ldots, \theta_{G,L_G}) \;\middle|\; \forall g,l\colon \theta_{g,l} \geq 0;\ \forall g\colon \sum_{l=1}^{L_g} \theta_{g,l} \leq 1 \right\} \subseteq [0,\ 1]^S,$$

mit $S = \Sigma_g L_g$.

b) Die Abbildung $f\colon \boldsymbol{\Omega} \rightarrow \boldsymbol{\Gamma}$ hat die Form $f_{\text{MVB}}\colon \boldsymbol{\Omega}_{\text{MVB}} \rightarrow \boldsymbol{\Gamma}$ mit der Abbildungsvorschrift

$f_{\text{MVB}}\colon \boldsymbol{\theta}$ a $(p_{1,1}(\boldsymbol{\theta}), \ldots, p_{1,J1}(\boldsymbol{\theta}), \ldots, p_{k,j}(\boldsymbol{\theta}), \ldots, p_{K,1}(\boldsymbol{\theta}), \ldots, p_{K,JK}(\boldsymbol{\theta}))$,

wobei die Vektorelemente $p_{k,j}(\boldsymbol{\theta})$, $k = 1, \ldots, K$, $j = 1, \ldots, J_k$, der durch Gleichung (3.4-7) vorgegebenen Struktur folgen.

Ein parametrisiertes multinomiales Modell der Struktur

$$\langle M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}_{\text{MVB}}, \boldsymbol{\Gamma}, f_{\text{MVB}} \rangle \tag{3.4-8}$$

heißt allgemeines multinomiales Verarbeitungsbaummodell.

3.5 Binäre multinomiale Verarbeitungsbaummodelle

Ein sehr häufiger Spezialfall allgemeiner MVB-Modelle ist dann gegeben, wenn pro Parametergruppe des Modells nur ein freier Parameter vorliegt, so daß $L_1 = L_2 = \ldots = L_G = 1$. Ein solches Modell folgt aus der in Abbildung 3.1 dargestellten allgemeinen Baumstruktur, wenn man annimmt, daß für jedes latente Ereignis — d.h. für jede Astgabelung im Baum — maximal zwei Folgeereignisse vorgesehen sind. Man nennt solche Modelle daher auch binäre Verarbeitungsbaummodelle (Batchelder & Riefer, 1999; Hu & Batchelder, 1994).

Für binäre MVB-Modelle läßt sich die Modellgleichung (3.4-7) vereinfachen, weil das Produkt der freien Parameter innerhalb einer Parametergruppe nicht mehr gebildet werden muß. Man kann sogar das Konzept der Parametergruppen ganz fallen lassen, weil jeder freie Parameter des binären Modells eine eigene Gruppe bildet. Es ist daher möglich, die Parameter einfach von $s = 1$ bis $s = S$ durchzunumerieren und sie mit θ_s zu notieren. Es wird also nur noch ein Index s für die Parameter des Modells benötigt; der Gruppenindex g kann fortfallen. Ent-

sprechend können auch die Zählvariablen vereinfacht werden: $a_{k,j,i,s}$ gibt für binäre Modelle an, wie oft der Parameter θ_s in Versuchsbedingung k im i-ten Ast zur Beobachtungskategorie $C_{k,j}$ vorkommt; $b_{k,j,i,s}$ repräsentiert die entsprechende Häufigkeit für den komplementären Parameter $(1-\theta_s)$. Unter Benutzung dieser Notation läßt sich die Modellgleichung (3.4-7) wie folgt vereinfachen:

$$\forall k, \forall j\colon p_{k,j}(\boldsymbol{\theta}) = \sum_{i=1}^{I_{k,j}} Pr^{*}(A_{k,j,i}) = \sum_{i=1}^{I_{k,j}} c_{k,j,i} \cdot \prod_{s=1}^{S} \left[(1-\theta_s)^{b_{k,j,i,s}} \cdot (\theta_s)^{a_{k,j,i,s}} \right] \quad (3.5\text{-}1)$$

Allgemeine MVB-Modelle, die

a) einen Parameterraum $\boldsymbol{\Omega}_{BVB} = \{\boldsymbol{\theta} = (\theta_1, ..., \theta_s, ..., \theta_S) | \; 0 \le \theta_s \le 1\} = [0, 1]^S$ aufweisen und

b) durch eine Abbildung der Form $f_{BVB}\colon \boldsymbol{\Omega}_{BVB} \rightarrow \boldsymbol{\Gamma}$ gekennzeichnet sind, die der durch Gleichung (3.5-1) gegebenen Abbildungsvorschrift
$f_{BVB}\colon \boldsymbol{\theta}$ a $(p_{1,1}(\boldsymbol{\theta}), ..., p_{1,J1}(\boldsymbol{\theta}), ..., p_{k,j}(\boldsymbol{\theta}), ..., p_{K,1}(\boldsymbol{\theta}), ..., p_{K,JK}(\boldsymbol{\theta}))'$
folgt,

sollen binäre multinomiale Verarbeitungsbaummodelle (binäre MVB-Modelle) heißen. Ein binäres MVB-Modell ist also ein parametrisiertes multinomiales Modell der Art

$$\left\langle M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}),\ \boldsymbol{\Omega}_{BVB},\ \boldsymbol{\Gamma},\ f_{BVB} \right\rangle . \quad (3.5\text{-}2)$$

Zwischen binären und nichtbinären MVB-Modellen bestehen sehr enge Beziehungen. Tatsächlich läßt sich ein nichtbinäres MVB-Modell sehr häufig — wenn auch nicht immer (s.u.) — ebensogut als binäres Modell darstellen und statistisch analysieren. Bei der Überführung eines nichtbinären MVB-Modells in ein binäres geht man einfach so vor, daß man die Parameter aus Parametergruppen mit mehr als einem freien Parameter als Produkte binärer Parameter reparametrisiert. Unter einem „binären Parameter" ist in diesem Zusammenhang ein Parameter zu verstehen, der einer Gruppe g mit nur einem freien Parameter angehört ($L_g = 1$). Anschaulich entspricht der Übergang von einem nichtbinären zu einem binären Modell der Überführung jeder nichtbinären Astgabelungen mit $L_g \ge 2$ funktional unabhängigen Parametern in einen neuen Teilbaum mit L_g binären Astgabelungen. Dies kann immer über einen binären Teilbaum mit L_g Ebenen realisiert werden: Man faßt zunächst alle Äste der nichtbinären Astgabelung bis auf den ersten zu einem neuen Ast zusammen (Ebene 1), splittet diesen dann binär so auf, daß der zweite Ast der ursprünglichen nichtbinären Astgabelung gegen alle Folgeäste kontrastiert wird (Ebene 2) und fährt so fort, bis Ebene L_g erreicht ist. Auf diese Weise werden beispielsweise drei nichtbinäre Parameter $\theta_{1,1}$, $\theta_{1,2}$ und $\theta_{1,3}$, die sich zu 1 aufsummieren ($L_g = 2$), repräsentiert als $\theta_{1,1} = \theta'_1$, $\theta_{1,2} = (1-\theta'_1) \cdot \theta'_2$ und $\theta_{1,3} = (1-\theta'_1) \cdot (1-\theta'_2)$, womit man sie auf zwei binäre Parameter θ'_1 und θ'_2

zurückgeführt hätte (Batchelder, 1999; Hu & Batchelder, 1994). Natürlich nehmen die binären Parameter — von θ'_1 abgesehen — andere numerische Werte an als die nichtbinären Ausgangsparameter. Dies hat bei der statistischen Analyse des Modells insofern Konsequenzen, als Punktschätzungen und Konfidenzintervalle nicht ohne weiteres vom binären auf den nichtbinären Fall übertragbar sind. Während man die Punktschätzungen von θ'_1 und θ'_2 durch Einsetzen in die o.g. Gleichungen noch relativ problemlos in Punktschätzungen für $\theta_{1,1}$, $\theta_{1,2}$ und $\theta_{1,3}$ „rückübersetzen" kann, gelingt dies für die entsprechenden Konfidenzintervalle leider nur mit wesentlich mehr Aufwand. Dies ist ein Grund, zwischen der statistischen Analyse binärer und nichtbinärer MVB-Modelle zu unterscheiden.

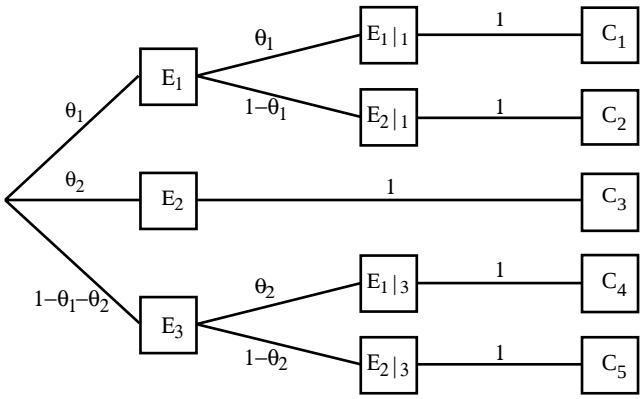

Abbildung 3.3:
Beispiel für ein allgemeines (nichtbinäres) MVB-Modell mit zwei Gleichheitsrestriktionen, das nicht in ein binäres Modell überführt werden kann.

Ein wichtigerer Grund, auf nichtbinäre MVB-Modelle nicht komplett zu verzichten, ergibt sich aus möglichen Restriktionen über die Modellparameter. Sind z.B. in einem Modell zwei oder mehr Parameter einer nichtbinären Gruppe mit Parametern anderer Parametergruppen gleichgesetzt, so läßt sich dieses Restriktionsmuster nicht adäquat in ein binäres Modell übertragen. Anders ausgedrückt: Ist man bei der „Übersetzung" eines nichtbinären Modells in ein binäres gezwungen, Äste zusammenzufassen, denen restringierte Parameter zugeordnet sind, so geht das intendierte Restriktionsmuster des nichtbinären Modells verloren. Abbildung 3.3 liefert ein Beispiel hierfür. Wären alle Äste, in denen die Parameter θ_1 und θ_2 auftauchen, mit unterschiedlichen (freien) Parametern belegt, so wäre die Übersetzung in ein binäres Modell kein Problem. Auch wenn nur einer der Parameter an zwei verschiedenen Astgabelungen auftauchen würde, der andere jedoch nicht, wäre das noch kein Problem. Man müßte dann im ersten Schritt nur die Äste der nichtbinären Parametergruppe zu einem Ast zusammenfassen, in denen kein restringierter Parameter auftaucht. Da jedoch im Modell gemäß Abbildung 3.3 sowohl θ_1 als auch θ_2 Gleichheitsrestriktionen unterworfen sind, führt

jede denkbare Astzusammenfassung zu einer Zerstörung des intendierten Restriktionsmusters.

3.5.1 Ein Beispiel: Das Enkodierung-Abruf-Modell

Das Modell zum Loftus-Paradigma, welches im letzten Abschnitt wiederholt zur Veranschaulichung herangezogen wurde, ist ein Beispiel für ein binäres MVB-Modell. Man kann es in der Form (3.5-1) notieren, wenn man die $S = 3$ freien Parameter des Modells z.B. als $\theta_1 = r$, $\theta_2 = m$ und $\theta_3 = b$ festlegt. Ein weiteres wichtiges binäres MVB-Modell ist das Enkodierung-Abruf-Modell (storage-retrieval model) von Batchelder und Riefer (1980; 1986). Da es einige Besonderheiten aufweist und eines der bislang am häufigsten angewandten kognitionspsychologisch motivierten MVB-Modelle sein dürfte, soll es hier ausführlicher vorgestellt werden.

Das Enkodierung-Abruf-Modell (EA-Modell) ist ein Modell zur freien Reproduktion von Itemlisten (z.B. Wortlisten). Es zielt primär darauf ab, die beobachtete Reproduktionsleistung in Beiträge der Enkodierungs- bzw. Speicherphase (storage) und Beiträge der Abrufphase (retrieval) zu dekomponieren. Wie weiter unten gezeigt wird, können auf diese Weise Gedächtnisdefizite, die für bestimmte Probandengruppen oder für bestimmte experimentelle Randbedingungen beobachtet werden, hinsichtlich der Beteiligung von Speicher- und Abrufprozessen „lokalisiert" werden. Dies ist eines der Kernprobleme der modernen Gedächtnispsychologie (vgl. Abschnitt 7.1). Das EA-Modell setzt eine Itemliste voraus, die teilweise aus Itempaaren (z.B. zwei Wörter einer semantischen Kategorie) und teilweise aus Einzelitems besteht (z.B. Wörter ohne semantischen Bezug zu anderen Wörtern der Liste). Itempaare und Einzelitems werden zunächst gelernt, wobei an den Anfang und an das Ende der Lernliste zusätzlich einige Füllitems plaziert werden. Diese Füllitems sollen die typischen Primacy- und Recency-Positionseffekte beim freien Reproduzieren absorbieren, die nicht Gegenstand des EA-Modells sind (vgl. z.B. Bredenkamp & Wippich, 1997b). Nach einem Retentionsintervall werden die Vpn aufgefordert, alle Items der Liste in beliebiger Reihenfolge zu reproduzieren (free recall). Das EA-Modell bezieht sich in seiner einfachsten Version auf Reproduktionsurteile für Itempaare ($k = 1$) und Einzelitems ($k = 2$). Die Füllitems sowie Intrusionen und Mehrfachnennungen bleiben unberücksichtigt. Bei Itempaaren werden vier Beobachtungkategorien unterschieden:

$C_{1,1}$: „Beide Items des Paares werden unmittelbar nacheinander reproduziert",
$C_{1,2}$: „Beide Items des Paares werden reproduziert, aber nicht nacheinander",
$C_{1,3}$: „Nur ein Item des Paares wird reproduziert" und
$C_{1,4}$: „Kein Item des Paares wird reproduziert".

Bei Einzelitems unterteilt man in die beiden trivialen Kategorien $C_{2,1}$: „Einzelitem reproduziert" bzw. $C_{2,2}$: „Einzelitem nicht reproduziert".

Natürlich sind andere Definitionen der Beobachtungskategorien zum gleichen Untersuchungsparadigma prinzipiell denkbar. Dieses Kategorienschema ist jedoch auf das Ziel der Modellkonstruktion insofern ideal zugeschnitten als es bei minimalem Aufwand genügend empirische Information für ein einfaches Speicher-Abruf-Modell bereitstellt. Für die 4 + 2 = 6 Beobachtungskategorien wird ein verbundenes multinomiales Modell M(K, $\boldsymbol{J}$, $\boldsymbol{N}$, $\boldsymbol{\pi}$) mit $K = 2$, $\boldsymbol{J} = (4, 2)$, $\boldsymbol{N} = (N_1, N_2)$ und $\boldsymbol{\pi} = (\pi_{1,1}, \pi_{1,2}, \pi_{1,3}, \pi_{1,4}, \pi_{2,1}, \pi_{2,2})$ angenommen. Das Modell basiert auf $S = 3$ freien Parametern, so daß $\boldsymbol{\Omega} = [0, 1]^3$. Die drei Parameter werden als „Wahrscheinlichkeit der Speicherung eines Itempaares als Cluster" (c), „Wahrscheinlichkeit des erfolgreichen Abrufs eines gespeicherten Itemclusters" (r) und „Wahrscheinlichkeit des erfolgreichen Abrufs eines nichtgeclusterten Items" (u) interpretiert.

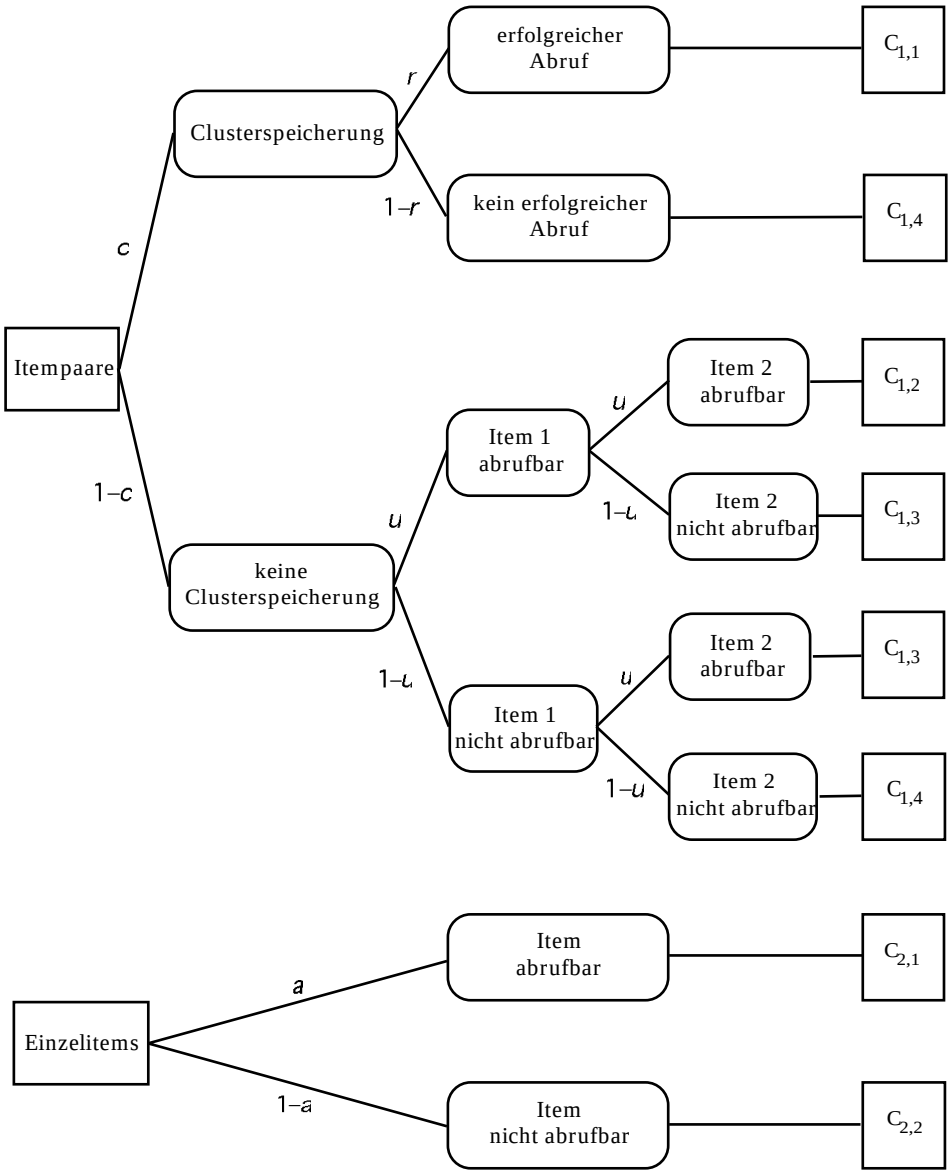

Abbildung 3.4:
Das Enkodierung-Abruf-Modell (storage-retrieval model) von Batchelder und Riefer (1980; 1986).

Abbildung 3.4 veranschaulicht das Baumdiagramm zum EA-Modell. Nach dem EA-Modell werden Itempaare mit Wahrscheinlichkeit *c* als Cluster gespeichert. Liegt eine Cluster-Speicherung vor, wird das Cluster mit Wahrscheinlichkeit *r* abgerufen, so daß die beiden Wörter unmittelbar nacheinander reproduziert werden können ($C_{1,1}$).[23] Gelingt der Abruf des Clusters mit Wahrscheinlichkeit (1–*r*) nicht, kann kein Item des Paares reproduziert werden. Geclusterte Items sind also nur als Paar oder aber gar nicht zugänglich.

Der Parameter *c* wird bei Anwendungen des EA-Modells häufig als Maß für eine besonders effektive Form der *Enkodierung* von Information, nämlich die Clusterung, interpretiert, während *r* als Maß für einen besonders effektiven *Abruf* gespeicherter Informationen dient. Aus diesen Interpretationen leitet sich die Bezeichnung des Modells her. Empirische Studien, auf die weiter unten eingegangen wird, bestätigen die psychologische Interpretation der Parameter.

Nichtgeclusterte Items eines Paares und Einzelitems sind mit Wahrscheinlichkeit *u* abrufbar. Bei Itempaaren kann das für beide Items der Fall sein, was nach dem EA-Modell zu einem $C_{1,2}$-Ereignis führen muß (vgl. Ast 3 des Baummodells). Natürlich ist es prinzipiell denkbar, daß nicht geclusterte, aber dennoch erfolgreich abgerufene Items eines Paares zufällig unmittelbar nacheinander reproduziert werden, was einem $C_{1,1}$-Ereignis entspräche. Wenn jedoch die Itemliste hinreichend lang ist, sollte dieses Ereignis wegen der unverbundenen Gedächtnisspuren beider Items eine vernachlässigbar geringe Wahrscheinlichkeit haben.

Kann nur ein Wort eines nichtgeclusterten Paares abgerufen werden, resultiert trivialerweise ein $C_{1,3}$-Ereignis (vgl. Ast 4 und 5 des Baummodells), kann kein Item abgerufen werden, ein $C_{1,4}$-Ereignis (Ast 6). Wie die Äste 3 bis 6 des Baummodells deutlich machen, werden in diesem Teil des Modells zwei weitere Annahmen zugrunde gelegt, die nicht ganz unproblematisch sind: Zum einen wird angenommen, daß beiden Items des Paares die gleiche Reproduktionswahrscheinlichkeit *u* zukommt, zum anderen — und diese Annahme ist noch kritischer —, daß die Wahrscheinlichkeit, mit der das zweite Item abgerufen wird, nicht davon abhängt, ob das erste Item erfolgreich abgerufen wurde (Ast 3) oder nicht (Ast 5). Mit anderen Worten: Es wird stochastische Unabhängigkeit des Abrufs nichtgeclusterter Items eines Paares unterstellt.

Einzelitems werden nach dem EA-Modell mit Wahrscheinlichkeit *a* abgerufen ($C_{2,1}$) und mit der komplementären Wahrscheinlichkeit 1–*a* nicht abgerufen ($C_{2,2}$). Wenn Einzelitems und Itempaare sich in bezug auf Salienz, Bildhaftigkeit,

[23] Denkbar ist natürlich, daß beim Abruf eines Clusters Intrusionen aus der gleichen (z.B. semantischen) Kategorie vorkommen. Man sollte deshalb bei der Auswertung eines Reproduktionsprotokolls zwischen verschiedenen Typen von Intrusionen unterscheiden. Tritt zwischen den beiden Items eines Paares eine Intrusion der gleichen Kategorie auf, sollte man dies als $C_{1,1}$-Ereignis werten. Ist dagegen zwischen den reproduzierten Items eines Paares eine Intrusion aus einer fremden Kategorie zu beobachten, erscheint eine Bewertung als $C_{1,2}$-Ereignis angemessener.

Bedeutungshaltigkeit und andere gedächtnisrelevante Störvariablen nicht systematisch unterscheiden, sollten sich Einzelitems eigentlich wie nicht geclusterte Items aus Itempaaren verhalten, so daß $a = u$ gilt. Man kann diese Gleichheitsrestriktion auch von vornherein in das EA-Modell einbauen und erhält dann ein testbares Modell, das 3 + 1 = 4 funktional unabhängige Kategoriewahrscheinlichkeiten mit drei Parametern (c, r, u) erklärt. Anders als c und r kann der u-Parameter des Modells weder als reiner Enkodierungs- noch als reiner Abrufparameter interpretiert werden, da er erfolgreiche Speicherung *und* erfolgreichen Abruf eines Einzelwortes widerspiegelt. Die gedächtnispsychologische Interpretation dieses Parameters ist nicht zuletzt aus diesem Grunde bislang weitgehend ungeklärt.

Tabelle 3.2:
Werte der Variablen $a_{k,j,i,s}$, $b_{k,j,i,s}$ und $c_{k,j,i}$ der binären MVB-Modellgleichung (3.5-1) für das Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986) gemäß Abbildung 3.4. Die Laufindex k indiziert die Itempaare (k = 1) und die Einzelitems (k = 2) des Modells, j die Nummer der Beobachtungskategorie pro Baum und i die Nummer des Astes in eine bestimmte Beobachtungskategorie (wobei — bezogen auf Abbildung 3.4 — die Äste von oben nach unten durchgezählt und die beiden Äste in die Beobachtungskategorie $C_{1,3}$ zu einem Ast zusammengefaßt werden). Der Index s ist der Laufindex für die Nummer des freien Parameters, wobei c als s = 1, r als s = 2 und u als s = 3 kodiert wird.

			$a_{k,j,i,s}$			$b_{k,j,i,s}$			$c_{k,j,i}$
k	j	i	s = 1	s = 2	s = 3	s = 1	s = 2	s = 3	
1	1	1	1	1	0	0	0	0	1
1	4	1	1	0	0	0	1	0	1
1	2	1	0	0	2	1	0	0	1
1	3	1	0	0	1	1	0	1	2
1	4	2	0	0	0	1	0	2	1
2	1	1	0	0	1	0	0	0	1
2	2	1	0	0	0	0	0	1	1

Bei dem EA-Modell handelt es sich offenkundig um ein binäres MVB-Modell, da jedes Ereignis im Baum maximal zwei Folgeereignisse mit komplementären Wahrscheinlichkeiten nach sich zieht. Man kann es folglich in Form der Modellgleichung (3.5-1) darstellen. Zur Verdeutlichung wird in Tabelle 3.2 die Matrix der Koeffizienten $a_{k,j,i,s}$ und $b_{k,j,i,s}$ angegeben, die aus Abbildung 3.4 abgeleitet werden kann. Da die beiden Äste in die Kategorie $C_{1,3}$ (Äste 4 und 5 des oberen Baumes) die gleiche Wahrscheinlichkeit $(1-c)\cdot u\cdot(1-u)$ haben, können sie auch zu einem Ast mit dem Gewicht $c_{1,3,1} = 2$ zusammengefaßt werden (vgl. Abb. 1 in Batchelder & Riefer, 1999).

Das EA-Modell ist methodisch in verschiedener Hinsicht interessant. Zunächst einmal haben wir es hier anders als im Falle des Loftus-Modells, der Signalentdeckungstheorie und des allgemeinen Schwellenmodells nicht mit einem produktmultinomialen Modell zu tun, dessen Kategoriensystem lediglich Beobachtungen einer kategorialen Variablen — z.B. richtige versus falsche Vp-Antworten — unter verschiedenen Versuchsbedingungen umfaßt. Vielmehr handelt es sich um ein verbundenes multinomiales Modell, dessen Beobachtungskategorien sich zwischen den Versuchsbedingungen qualitativ und quantitativ unterscheiden. Darüber hinaus ergeben sich die vier Beobachtungskategorien für Itempaare nicht unmittelbar aus dem Vp-Verhalten, sondern sie sind im Hinblick auf die Zielrichtung der Modellkonstruktion von Batchelder und Riefer (1980) gezielt so ausgewählt worden, daß sie eine Erfassung von Cluster-Enkodierungs- und Cluster-Abrufprozessen auf möglichst einfache, d.h. parameterarme Weise ermöglichen. Man erkennt an diesem Beispiel, daß die theoriegeleitete Konstruktion multinomialer Modelle nicht nur die Entwicklung von Modellgleichungen zu vorgegebenen Beobachtungskategorien umfaßt; vielmehr beginnt sie bereits bei der Spezifizierung des experimentellen Paradigmas und der problembezogenen Auswahl von geeigneten Beobachtungskategorien (vgl. Kapitel 11).

Wie Tabelle 3.2 deutlich macht, hebt sich das EA-Modell von den zuvor besprochenen MVB-Modellen auch dadurch ab, daß (a) bestimmte Parameter pro Ast mehrfach auftauchen (vgl. Zeilen 3 und 5 der Tabelle) und (b) der multiplikative Faktor $c_{k,j,i}$ größer als 1 wird (vgl. Zeile 4). Schließlich wird anhand dieses Modells auch deutlich, wie man stochastische Unabhängigkeitsannahmen zu den modellierten latenten Ereignissen im Rahmen eines MVB-Modells formalisieren kann, nämlich durch Gleichsetzung von Parametern in verschiedenen Ästen des Baummodells (vgl. Abschnitt 4.3.2). Das EA-Modell hat also eine etwas komplexere Struktur als die bislang besprochenen Modelle und offenbart dadurch mehr über die Möglichkeiten, die sich im Rahmen der multinomialen Verarbeitungsbaummodellierung bieten.

Wichtiger als diese formalen Aspekte sind jedoch die zahlreichen substanzwissenschaftlichen Anwendungen, die das Modell bislang gefunden hat. Sie bestätigen einerseits die psychologische Validität der Enkodierungs- und Abrufparameter des Modells und helfen andererseits, offene Forschungsfragen zu klären, die auf rein designbasiertem Wege — d.h. ohne Verwendung mathematischer Modelle (vgl. Abschnitt 2.6) — bislang nicht befriedigend geklärt werden konnten.

Batchelder und Riefer (1980; 1986) konnten mit Hilfe des EA-Modells die Widersprüche zwischen verschiedenen Untersuchungen zum Effekt des Abstandes semantisch zusammengehöriger Wörter in der Lernliste (sog. within-category spacing) auf die anschließende Reproduktionsleistung aufklären (zusammenfassend: Bredenkamp & Wippich, 1977b, S. 45 f.): Sie zeigten, daß mit zunehmendem

Abstand zusammengehöriger Items einerseits die Cluster-Enkodierungswahrscheinlichkeit *c* sinkt, andererseits jedoch die Cluster-Abrufwahrscheinlichkeit *r* ansteigt. Ersteres kann über die Erleichterung der Kategorieidentifikation aufgrund raum-zeitlicher Kontiguität erklärt werden, letzteres z.B. über Abruferleichterung durch Enkodierungsvariabilität: Je unterschiedlicher die Enkodierungskontexte gemeinsam abgespeicherter Gedächtnisinhalte sind, um so mehr Möglichkeiten gibt es in der Abrufsituation, Zugang zu den Gedächtnisinhalten zu finden (zusammenfassend: Bredenkamp & Wippich, 1977b, S. 66 f).

Ein kleines Problem der von Batchelder und Riefer (1986) vorgeschlagenen Maximum-Likelihood-Schätzer für die Modellparameter besteht allerdings darin, daß sie immer nur für *eine* Menge von Itempaaren und *eine* dazugehörige Menge von Einzelitems konzipiert sind. Im Falle des Datensatzes von Batchelder und Riefer (1980) liegen aber vier Mengen von Itempaaren (mit einem Abstand von 0, 4, 12 und 14 Items zwischen den beiden Items einer Kategorie) bei nur einer Menge von Einzelitems vor. Batchelder und Riefer (1986) „vervierfachten" deshalb den Datensatz zu den Einzelitems künstlich und schätzten entsprechend für jede der o.g. Itemabstands-Bedingungen einen separaten *u*-Parameter, obwohl de facto immer der gleiche Rohdatensatz vorlag. Glücklicherweise ergaben sich hierbei jedoch keine großen Unterschiede zwischen den vier $\hat{u}$-Schätzern, und die Auswirkungen auf die Schätzer der anderen Parameter sowie die Modellanpassungsstatistiken erweisen sich als marginal, wenn man die Ergebnisse bei optimaler Analyse des Datensatzes zum Vergleich heranzieht (Erdfelder, 1990b; 1992b). Dies muß jedoch nicht bei jeder Anwendung des EA-Modells so sein. Deshalb empfiehlt es sich, bei künftigen Modellanwendungen flexiblere Methoden der Parameterschätzung einzusetzen, welche die künstliche „Vervielfältigung" von Teildatensätzen (z.B. für die Einzelitems) unnötig macht. Neben den noch zu besprechenden Allzweck-Computerprogrammen für MVB-Modelle (AppleTree, MBT, GPT oder die „Solver"-Funktion aus EXCEL) erfüllt das speziell auf das EA-Modell zugeschnittene und sehr einfach handhabbare Programm „SRM" (Erdfelder, 1992b) diese Bedingung (vgl. Abschnitt 4.7).

Eine interessante methodische Erweiterung des EA-Modells stammt von Riefer (1982). Er zeigte, wie man das Modell auf Experimente mit kategorisierten Itemlisten verallgemeinern kann, die mehr als zwei Elemente pro Kategorie umfassen. Allerdings geht diese Erweiterung mit einem erheblichen Anstieg der Parameteranzahl einher, wenn die Itemkategorien groß sind. Einige dieser Parameter lassen sich psychologisch zudem nur schwer interpretieren. Nicht zuletzt aus diesem Grund dürften Anwendungen des EA-Modells von Riefer (1982) bislang rar geblieben sein. Publikationen, die explizit auf diese Version des EA-Modells Bezug nehmen, sind mir nicht bekannt.

Das EA-Modell ist nicht nur zur Analyse des Spacing-Effekts, sondern auch zur Analyse des Altersdefizits im episodischen Gedächtnis eingesetzt worden. Riefer und Batchelder (1991a) fanden einen deutlichen Unterschied in den *r*-Schätzern, jedoch keinen signifikanten Unterschied in den *c*-Schätzern zwischen alten und jungen Probanden. Qualitativ ähnlich war das Ergebnis von Erdfelder und Bayen (1991), das ausführlich von Bayen (1990) referiert wird. Auch hier zeigte sich in den *r*-Parameterschätzungen eine signifikante und praktisch bedeutsame Cluster-Abrufbeeinträchtigung bei älteren im Vergleich zu jüngeren Probanden und ein vergleichsweise kleiner Unterschied im Cluster-Enkodierungsparameter *c*. Im Gegensatz zur Arbeit von Riefer und Batchelder (1991a) erwies sich jedoch auch dieser kleine Unterschied noch als statistisch signifikant, so daß man schließen muß, daß dem Altersdefizit im episodischen Gedächtnis zwar in erster Linie eine Abrufbeeinträchtigung zugrunde liegt, in zweiter Linie aber auch eine nicht ganz so stark ausgeprägte Cluster-Enkodierungsbeeinträchtigung. Da die deskriptiv-statistischen Ergebnisse — d.h. die Parameterschätzungen — bei Riefer und Batchelder (1991a) sowie Erdfelder und Bayen (1991) sehr ähnlich ausfielen, dürften die unterschiedlichen inferenzstatistischen Ergebnisse vor allem auf die kleineren Stichprobenumfänge bei Riefer und Batchelder (1991a) zurückzuführen sein: Sie untersuchten 23 ältere (64- bis 92-jährige) und 23 jüngere (21- bis 41-jährige) Probanden, während Erdfelder und Bayen (1991) 40 ältere (63- bis 92-jährige) und 40 jüngere (17 bis 26 jährige) Personen untersucht hatten. Folglich ist — gleiches Signifikanzniveau unterstellt — die statistische Power der Tests auf Altersunterschiede in der letztgenannten Untersuchung größer (vgl. Abschnitt 4.3.3).

Abgesehen von diesem statistisch erklärbaren Ergebnisunterschied lieferten beide Untersuchungen konvergierende Evidenz zur Validität der Parameter des EA-Modells. In beiden Untersuchungen stiegen bei dreifacher Wiederholung des Listenlernexperiments alle drei Modellparameterschätzungen monoton mit der Lernerfahrung an. Weiterhin untersuchten Erdfelder und Bayen (1991) die Effekte einer präsentierten versus nicht präsentierten Cluster-Enkodierungsinstruktion („Versuchen Sie, semantisch zusammengehörige Wortpaare zu erkennen und gemeinsam abzuspeichern") und einer unmittelbaren versus um fünf Minuten verzögerten Reproduktion der Itemliste. Erwartungsgemäß zeigte sich ein signifikanter Effekt der Enkodierungsinstruktion nur auf den *c*-Parameter und nur bei solchen Itempaaren, die in der Lernliste separiert waren (Abstand: 15 Items); Itempaare, die unmittelbar nacheinander dargeboten worden waren (Abstand: 0 Items) offenbarten keinen Effekt der Enkodierungsinstruktion. Interpretieren läßt sich dieses Ergebnis folgendermaßen: Die Cluster-Enkodierungsinstruktion ist bei unmittelbar nacheinander dargebotenen Items einer Kategorie vermutlich redundant, bei schwer erkennbaren Kategorien, deren Items in der Lernliste um 15 Items separiert sind, jedoch möglicherweise hilfreich und effektiv. Alte und junge Pro-

banden profitierten von diesem Effekt gleichermaßen. Hingegen zeigte die Manipulation des Retentionsintervalls nur einen Effekt auf den *u*-Parameter bei älteren Probanden, obwohl zusätzlich ein Effekt auf den Abrufparameter *r* erwartet worden war. Vermutlich war die Manipulation des Retentionsintervalls einfach nicht stark genug, um den erwarteten Effekt zu produzieren. Von diesem Teilergebnis abgesehen, fielen jedoch alle Resultate hypothesenkonform aus.

Ein weiteres wichtiges Anwendungsfeld des EA-Modells sind Untersuchungen zu den Determinanten retroaktiver Hemmung. Riefer und Batchelder (1988, S. 331f) fanden, daß zusätzliche Itemlisten, die zwischen der zu reproduzierenden Liste und dem Abrufzeitpunkt dargeboten werden, den Cluster-Abrufparameter *r* herabsetzen, nicht jedoch den *c*-Parameter. Dieses Ergebnis wurde von Bäuml (1991c) repliziert, und zwar sowohl für semantische Wortkategorien als auch für phonetisch-visuelle Wortkategorien (Wortpaare, die sich reimen). Bäuml (1991a) zeigte allerdings, daß ein gewisser, wenn auch nicht sehr starker Effekt auf den Cluster-Speicherparameter *c* zu beobachten ist, wenn man anstelle einer massierten Darbietung zusammengehöriger Wörter (Abstand: 0 Items) eine verteilte Darbietung mit Zufallsabständen wählt. In einer weiteren Arbeit (Bäuml, 1991b) zeigte sich ein noch stärkerer Interferenzeffekt auf den *c*-Parameter, wenn den interpolierten Listen zwar andere Itempaare, aber die gleichen semantischen Kategorien zugrunde lagen wie der zu reproduzierenden Liste. Man kann also demnach retroaktive Hemmung nicht einfach auf eine reine Abrufhemmung reduzieren, die durch geeignete Abrufhilfen immer eliminiert werden kann (Tulving & Psotka, 1971). Wenn Bedingungen geschaffen werden, unter denen die zu reproduzierenden und die konkurrierenden Informationen aufgrund ihrer Ähnlichkeit oder aufgrund der Darbietungsbedingungen stark interferieren, ist auch ein bedeutsamer retroaktiver Hemmungseffekt im *c*-Parameter feststellbar.

Bäuml (1996) hat das EA-Modell zu einem Markoff-Ketten-Modell erweitert, das nicht nur die Analyse der Daten eines Reproduktionsdurchganges, sondern die gemeinsame Analyse der Ergebnisse wiederholter Reproduktionsdurchgänge ermöglicht. Ein wesentlicher Vorteil dieses Modells besteht darin, daß es reine Enkodierungseffekte (Wird ein Itempaar in der Akquisitionsphase als Cluster enkodiert?) und reine Speichereffekte nach einer erfolgreichen Cluster-Enkodierung (Geht die Gedächtnisrepräsentation eines enkodierten Clusters später wieder verloren?) voneinander trennen kann, da im Modell unterschiedliche Modellparameter für die ursprüngliche Cluster-Enkodierung ($\alpha_{P,1} + \alpha_{P,2}$) und den Verlust eines gespeicherten Clusters (β) enthalten sind. Da zusätzlich auch noch ein Parameter für die Abrufhemmung bei vorliegender Cluster-Speicherung existiert (Parameter γ) — das Pendant zum *r*-Parameter also — gibt es im Markoff-EA-Modell Parameter für alle drei Phasen, die in Gedächtnisuntersuchungen typischer-

weise unterschieden werden: Enkodierung, Speicherung und späterer Abruf (vgl. z.B. Bredenkamp & Erdfelder, 1996).

Bäuml (1996) hat mit dem Markoff-EA-Modell die Daten der oben bereits geschilderten Experimente reanalysiert und die ursprünglichen Schlußfolgerungen bestätigen können. Das Markoff-Modell paßte zu den Daten aller drei Experimente (Bäuml, 1991a; 1991b; 1991c), was zeigt, daß auch Veränderungen von Gedächtnisrepräsentationen (Verlust gespeicherter Informationen) oder Veränderungen der Zugänglichkeit von Gedächtnisrepräsentationen (Abrufbeeinträchtigungen) durch die recht einfachen Annahmen eines stationären Markoff-Ketten-Modells beschrieben werden können. Formal ist interessant, daß auch die Markoff-Version des EA-Modells selbst wieder ein binäres MVB-Modell ist, weil — wie sich zeigen läßt — die Modellgleichungen in der Form (3.5-1) geschrieben werden können. Wir kommen auf die Beziehung zwischen Markoff-Ketten-Modellen und MVB-Modellen in Abschnitt 6.5 zurück.

Als Fazit bleibt festzuhalten, daß das EA-Modell — einschließlich seiner Erweiterungen durch Bäuml (1996) und Riefer (1982) — eines der bislang fruchtbarsten MVB-Modelle in der kognitionspsychologischen Forschung ist. Wünschenswert sind Folgeuntersuchungen, die einerseits die Validität des Abrufparameters *r* weiter untermauern — hierzu hatte die Untersuchung von Erdfelder und Bayen (1991) noch kein befriedigendes Ergebnis erbracht — und die andererseits die gedächtnispsychologische Bedeutung des hybriden *u*-Parameters etwas erhellen. Denkbar ist, daß die Unterscheidung item-spezifischer und relationaler Enkodierungsprozesse (item-specific and relational processing) wie sie von verschiedenen Autoren vorgeschlagen wurde (z.B. Einstein & Hunt, 1980; Hunt & Einstein, 1981) hierbei weiterhilft: Wenn Manipulationen der relationalen Enkodierungsstärke ausschließlich auf den *c*-Parameter, Manipulationen der item-spezifischen Enkodierungsstärke dagegen ausschließlich auf *u* wirken, wäre die Bedeutung des *u*-Parameters zumindest partiell erhellt.

3.5.2 Weitere Beispiele: Hochschwellen-Modelle

Neben dem EA-Modell sind vor allem diverse Spezialfälle des allgemeinen Schwellenmodells gemäß Abbildung 3.2 für die praktische Anwendung sehr wichtige und auch schon weithin bekannte Beispiele für binäre MVB-Modelle. Dies gilt etwa für das klassische Ein-Hochschwellen-Modell (one-high threshold model, 1HT-Modell), das schon von Blackwell (1953, S. 16) als Korrekturmethode für Trefferraten in Signalentdeckungstests propagiert wurde.[24] Es braucht uns vorerst

[24]Blackwell (1953, S. 16) hat die Korrekturformel der Trefferrate gemäß 1HT-Modell explizit angegeben, das Modell selbst aber nur sehr knapp und irreführend geschildert: „This correction may be shown to depend upon the assumption that “Yes” responses based upon sensory differences, and other “Yes” responses, based, for example, upon expectations, are statistically independent.”

nicht zu interessieren, daß dieses Modell wegen offenkundiger Schwächen immer wieder heftiger Kritik ausgesetzt war (z.B. von Kinchla, 1994; Kintsch, 1970; Luce, 1963a; Murdock, 1974; Snodgrass & Corwin, 1988; Tanner & Swets, 1954; Vorberg & Schmidt, 1975). Diese Schwächen und Methoden ihres Nachweises werden in Abschnitt 4.5 behandelt. Ich führe das 1HT-Modell hier nur deshalb kurz ein, weil es sich aufgrund seiner Einfachheit zur Veranschaulichung bestimmter statistischer Konzepte sehr gut eignet (vgl. Abschnitt 4.1) und weil es als „Baustein" in anderen MVB-Modellen, etwa in Quellendiskriminationsmodellen (vgl. Batchelder, Hu & Riefer, 1994; Batchelder & Riefer, 1990; Batchelder, Riefer & Hu, 1994; Riefer, Hu & Batchelder, 1994), immer wieder auftaucht.

Man erhält das Ein-Hochschwellen-Modell aus dem allgemeinen Schwellenmodell, wenn man eine hohe Schwelle für den Zustand hoher Vertrautheit annimmt, deren Überschreitung mit Wahrscheinlichkeit $b_{\mathrm{v}} = 1$ ein „Alt"-Urteil zur Folge hat. Ferner basiert das Modell auf den Annahmen, daß (a) alte Items nie in den Zustand ausgeprägter Unvertrautheit geraten ($r_{\mathrm{ua}} = 0$), während (b) neue Items weder sehr vertraut noch sehr unvertraut sind ($r_{\mathrm{vn}} = r_{\mathrm{un}} = 0$), so daß sie automatisch den „unsicheren" Zustand zur Folge haben. Folglich umfaßt das Ein-Hochschwellen-Modell nur noch $S = 2$ freie Parameter, nämlich einen Parameter r_{va} für die Rekognitionsrate (überschwellige Vertrautheit) bei alten Items (im folgenden mit $r_{1\mathrm{ht}}$ abgekürzt) und einen Antworttendenzparameter b_{r} (Wahrscheinlichkeit, im Zustand der Unsicherheit „alt" zu raten, im folgenden mit $b_{1\mathrm{ht}}$ abgekürzt). Die beiden Modellgleichungen, welche die Trefferrate und die Rate falschen Alarms als Funktion der Modellparameter ausdrücken, ergeben sich somit aus Abbildung 3.2 wie folgt:

$$p_{1,1}(r_{1\mathrm{ht}}, b_{1\mathrm{ht}}) = Pr(\text{„alt"}; \text{alte Items}) = r_{1\mathrm{ht}} + (1 - r_{1\mathrm{ht}}) \cdot b_{1\mathrm{ht}} \qquad (3.5.2\text{-}1)$$

$$p_{2,1}(r_{1\mathrm{ht}}, b_{1\mathrm{ht}}) = Pr(\text{„alt"}; \text{neue Items}) = b_{1\mathrm{ht}} \qquad (3.5.2\text{-}2)$$

Ein dem 1HT-Modell eng verwandter Spezialfall des allgemeinen Schwellenmodells ist das Zwei-Hochschwellen-Modell (two-high oder double-high threshold model, 2HT-Modell), das erstmals von Egan (1958, zit. nach Macmillan & Creelman, 1991) vorgeschlagen und später dann von verschiedenen Autorinnen und Autoren in unterschiedlichen Kontexten diskutiert wurde (Green & Swets, 1966; Macmillan & Creelman, 1990; Macmillan & Creelman, 1991; Snodgrass & Corwin, 1988; Swets, 1986). Wie die Bezeichnung suggeriert, basiert das 2HT-Modell auf der Annahme zweier hoher Schwellen, eine für den Zustand „sehr vertraut" und eine für den Zustand „sehr unvertraut". Bezogen auf das allgemeine Schwellenmodell gemäß Abbildung 3.2 wird angenommen, daß alte Items nie die Schwelle für den Zustand „sehr unvertraut" überschreiten ($r_{\mathrm{ua}} = 0$) während neue Items nie die Schwelle für den Zustand „sehr vertraut" überschreiten ($r_{\mathrm{vn}} = 0$). Im Zustand „sehr

vertraut" wird immer die Antwort „alt" (b_v = 1), im Zustand „sehr unvertraut" entsprechend immer die Antwort „neu" gegeben (b_u = 0). Schließlich wird unterstellt, daß die Rate „echter" Rekognitionen für alte Items gleich der Rate „echter" Distraktorentdeckungen für neue Items ist, so daß $r_{va} = r_{un} = r_{2ht}$. Insgesamt verbleiben so ein Diskriminations-Parameter (r_{2ht}) und ein Rateparameter b_r im Modell, der im folgenden mit b_{2ht} bezeichnet werden soll. Die aus Abbildung 3.2 leicht herleitbaren Modellgleichungen für die Trefferrate und die Rate falschen Alarms lauten:

$$p_{1,1}(r_{2ht}, b_{2ht}) = Pr(\text{„alt"}; \text{alte Items}) = r_{2ht} + (1 - r_{2ht}) \cdot b_{2ht} \qquad (3.5.2\text{-}3)$$

$$p_{2,1}(r_{2ht}, b_{2ht}) = Pr(\text{„alt"}; \text{neue Items}) = (1 - r_{2ht}) \cdot b_{2ht} \qquad (3.5.2\text{-}4)$$

Die Annahme der Gleichheit von r_{va} und r_{un} ist sicherlich die problematischste Annahme des 2HT-Modells. Sie wurde m.W. nie theoretisch, sondern immer nur pragmatisch gerechtfertigt: Diese Restriktion führt zu einem saturierten Zwei-Parameter-Modell, das statistisch einfach handhabbar ist, während das 2HT-Modell ohne diese Restriktion überparametrisiert wäre. Man kann die Angemessenheit der Gleichheitsrestriktion jedoch testen, indem man unter sonst gleichen Bedingungen die Ratewahrscheinlichkeit experimentell manipuliert, so daß eine empirische Schätzung für die sogenannte receiver-operating characteristic curve (ROC-Kurve) resultiert. Unter diesen Umständen ist das restringierte Modell auf ein 2HT-Modell erweiterbar, das auf die Gleichsetzung von r_{va} und r_{un} verzichtet, ohne deshalb überparametrisiert zu sein. Erstaunlicherweise zeigte sich wiederholt, daß diese Erweiterung im allgemeinen empirisch nicht nötig ist, d.h. das 2HT-Modell paßt auch ohne diese Erweiterung recht gut auf die ROC-Daten (vgl. z.B. die Datensätze bei Bernbach, 1967; Grier, 1971; Snodgrass & Corwin, 1988).

Die Parameter r_{2ht} und b_{2ht} des 2HT-Modells schnitten in Evaluationsstudien immer deutlich besser als die Parameter des 1HT-Modells und teilweise auch besser als signalentdeckungstheoretisch begründete Parameter ab (vgl. z.B. Macmillan & Creelman, 1990; Snodgrass & Corwin, 1988). Darüber hinaus fand und findet es nach wie vor häufig implizit Anwendung — d.h. ohne explizite Bezugnahme auf die Modellannahmen —, nämlich immer dann, wenn als Maß für die Rekognitionsleistung die schon von Woodworth (1938) vorgeschlagene Ratekorrektur „Treffer minus falsche Alarme" benutzt wird: Wie man durch Subtraktion der beiden Modellgleichungen (3.5.2-3) und (3.5.2-4) leicht herleiten kann, ist diese „Ratekorrektur" nichts anderes als ein Schätzer für den Rekognitionsparameter r_{2ht} des 2HT-Modells. Trotz der positiven Eigenschaften und der weiten Verbreitung der 2HT-Modellparameter als Maße für die Rekognitionsleistung und den Antwortbias stößt das Modell — wie auch viele andere Schwellenmodelle — immer wieder auf Kritik, vor allem wegen der implizierten nichtregulären ROC-

„Kurve", einer Geraden, die für $r_{2ht} > 0$ nicht durch die Punkte (0, 0) und (1, 1) geht (z.B. Macmillan & Creelman, 1991; Swets, 1986). Entschieden ist die Debatte um das „beste" Modell zu Ja-Nein-Rekognitions- und Signalentdeckungsaufgaben nach wie vor nicht, und denkbar ist auch, daß es ein für alle Anwendungskontexte optimales Modell gar nicht gibt.

3.6 Ereignisbaumrepräsentation von Verarbeitungsbaummodellen

In den letzten Abschnitten wurde gezeigt, daß parametrisierte multinomiale Modelle, die sich durch Ereignisbaumdiagramme der in Abbildung 3.1 dargestellten Struktur repräsentieren lassen, immer auch multinomiale Verarbeitungsbaummodelle sind, d.h. Modellgleichungen der Form (3.4-7) (allgemeine MVB-Modelle) bzw. (3.5-1) aufweisen (binäre MVB-Modelle). Es existiert eine Abbildung der Menge möglicher Ereignisbaumdiagramme auf die Menge möglicher MVB-Modelle, d.h. einem beliebigen Ereignisbaumdiagramm kann immer eine ganz bestimmte — für MVB-Modelle charakteristische — Abbildung $f_{MVB}: \mathbf{\Omega} \rightarrow \mathbf{\Gamma}$ in eindeutiger Weise zugeordnet werden. Der Nachweis dieser Behauptung ist trivial: Bei gegebenem Ereignisbaumdiagramm muß man nur die einzelnen Astwahrscheinlichkeiten durch Produktbildung der Parameter ausrechnen und dann die Astwahrscheinlichkeiten zu einer bestimmten Kategorie summieren, um die Kategoriewahrscheinlichkeit zu erhalten. Das entspricht formal zwangsläufig einer Gleichung des Typs (3.4-7) bzw. (3.5-1). Die einzige Bedingung, die ein Ereignisbaumdiagramm erfüllen muß, damit ein MVB-Modell daraus ableitbar ist, ist die Nebenbedingung, daß sich die Astwahrscheinlichkeiten pro Baum für beliebige Werte der freien Parameter zu Eins aufaddieren.[25]

Nun soll die umgekehrte Frage gestellt werden: Existiert auch eine Abbildung möglicher MVB-Modelle auf die Menge möglicher Ereignisbäume? Sind binäre oder allgemeine MVB-Modelle immer eindeutig durch einen bestimmten Ereignisbaum der in Abbildung 3.1 dargestellten Form repräsentierbar? Dies ist eindeutig zu verneinen. Wenn man zum Beispiel einen binären Zwei-Ebenen-Ereignisbaum mit vier Ästen betrachtet, der auf der ersten Ebene zwei latente Ereignisse mit Wahrscheinlichkeit $Pr^*(E_1) = \theta_1$ und $Pr^*(E_2) = 1-\theta_1$ und auf der zweiten Ebene jeweils zwei Folgeereignisse mit gleichen bedingten Wahrscheinlichkeit $Pr^*(E_{1|1}) = Pr^*(E_{1|2}) = \theta_2$ und $Pr^*(E_{2|1}) = Pr^*(E_{2|2}) = 1-\theta_2$ einführt, dann kann man wegen der Kommutativität der Multiplikation die Parameter auf den beiden Ebenen (bzw. die dadurch repräsentierten psychologische Prozesse) auch vertauschen, ohne daß sich dadurch an den MVB-Modellgleichungen etwas ändert. Schon dies

[25]Diese Nebenbedingung ist in den o.g. Definitionen allgemeiner und binärer MVB-Modelle implizit mitgenannt: Da MVB-Modelle eine Teilklasse parametrisierter multinomialer Modelle bilden und diese per definitionem der Summationsregel folgen (vgl. (3.3-4)), muß dies auch für MVB-Modelle gelten.

allein zeigt, daß man sehr vorsichtig sein muß, wenn man MVB-Modelle als sequentielle Verarbeitungsmodelle interpretiert: Die Modellgleichungen reflektieren diese Interpretation nicht zwangsläufig.

Ein weiteres Beispiel für die Uneindeutigkeit der Ereignisbaumrepräsentation ergibt sich dann, wenn (z.B. durch Parameterrestriktionen) zwei oder mehr Äste, die in unterschiedliche Kategorien einmünden, gleiche Astwahrscheinlichkeiten haben. In diesem Fall kann man die Kategorien an den Astenden beliebig permutieren, ohne daß sich die Modellgleichungen dadurch ändern.

Sieht man einmal von Baumäquivalenzen ab, die sich bei Permutationen der Kategorien an den Astenden oder bei Permutationen der Parameter an den Astsegmenten ergeben, so stellt sich die Frage, ob wenigstens die Astanzahl und die Aststruktur eines Ereignisbaumdiagramms durch die MVB-Modellgleichungen eindeutig determiniert sind. Auch dies muß verneint werden (vgl. Hu & Phillips, 1999). Betrachten wir als Beispiel[26] die folgende Modellgleichungen:

$$\begin{aligned} p_1(\theta) &= \theta^3 \\ p_2(\theta) &= \theta \cdot (1-\theta) \\ p_3(\theta) &= \theta \cdot (1-\theta) \\ p_4(\theta) &= \theta \cdot (1-\theta) \\ p_5(\theta) &= (1-\theta)^3 \end{aligned} \qquad (3.6\text{-}1)$$

Offenkundig handelt es sich um ein einfaches binäres MVB-Modell mit nur einem freien Parameter θ, das sich auf eine Versuchsbedingung mit fünf Beobachtungskategorien C_1 bis C_5 bezieht. Jede der Gleichungen genügt der Form (3.5-1) und die fünf Kategoriewahrscheinlichkeiten summieren sich für beliebiges θ auch zu 1 auf, da $(1-\theta)^3 = 1 - 3\cdot\theta\cdot(1-\theta) - \theta^3$ gilt. Wie man sich leicht klarmachen kann, impliziert jedes der beiden Baumdiagramme (a) und (b) in Abbildung 3.5 diese MVB-Modellgleichungen, obwohl sich die Astanzahl (acht versus sechs Äste) und auch die Aststruktur unterscheidet. Aus jedem der Baumdiagramme lassen sich weitere äquivalente Baumdiagramme erzeugen, wenn man die Kategorielabels der Kategorien C_2, C_3 und C_4 permutiert.

Als Fazit bleibt festzuhalten, daß Tests von MVB-Modellen im allgemeinen nicht als Tests spezieller Ereignisbaumdiagramme interpretiert werden können. Es kann durchaus vorkommen, daß unterschiedliche Baumdiagramme, die aus unterschiedlichen psychologischen Ideen resultieren, zu denselben MVB-Modellgleichungen und damit zu denselben MVB-Modellen führen. Man wird die zugrundeliegenden psychologischen Ideen also nicht ohne Erweiterung des Modellrahmens und der Datenbasis gegeneinander testen können. In Abschnitt 6.8 werden

[26] Den Hinweis auf dieses besonders informative Beispiel verdanke ich Xiangen Hu (persönliche Mitteilung, 8. 9. 1997).

wir eine naheliegende Erweiterung von MVB-Modellen auf Antwort- bzw. Reaktionszeiten kennenlernen, mit der die in den Ereignisbaumdiagrammen (a) und (b) skizzierten Theorien potentiell empirisch diskriminierbar werden.

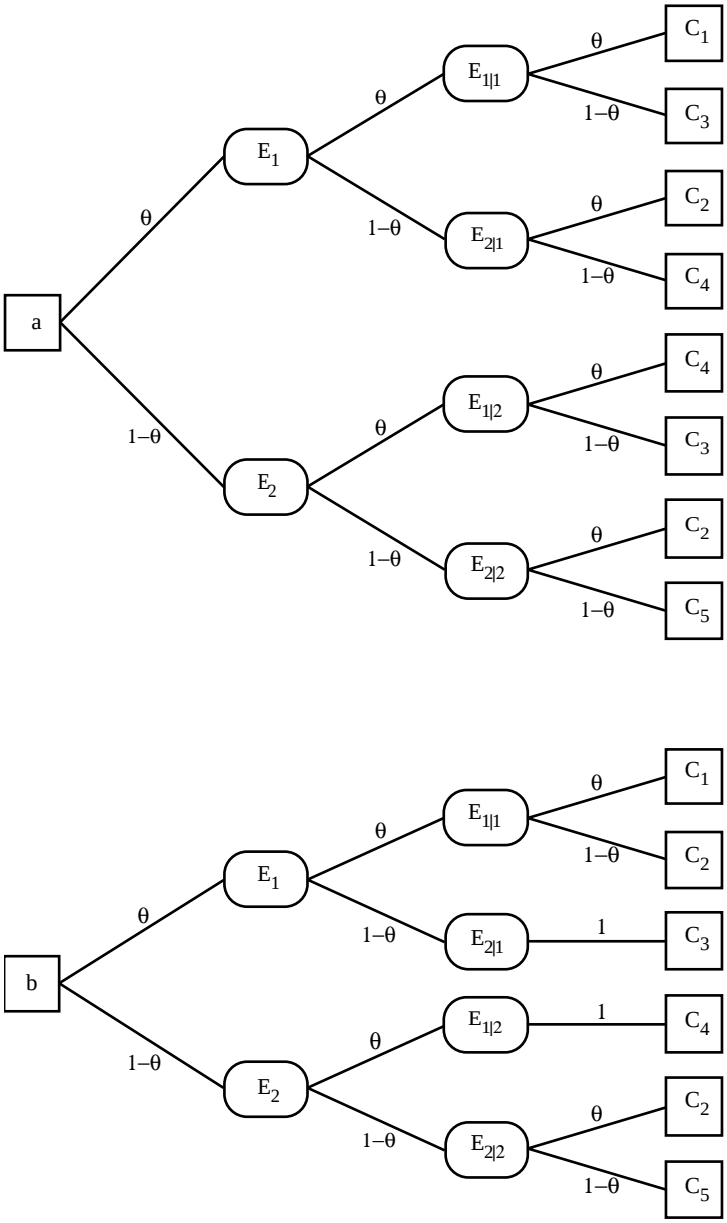

Abbildung 3.5:
Zwei unterschiedliche Ereignisbaumdiagramme (a) und (b), die zum gleichen MVB-Modell (3.6-1) führen.

3.7 Zusammenfassung

In diesem Kapitel wurden zunächst die Anforderungen geschildert, die multinomiale Modelle an die Daten stellen. Es wurde gezeigt, daß diese Anforderungen insofern realistisch sind, als sie der psychologischen Forschungspraxis weitgehend entsprechen. Es wurden ferner Verteilungsannahmen geschildert, die mit einfa-

chen oder verbundenen multinomialen Modellen (für eine bzw. *K* Gruppen) verbunden sind (Abschnitt 3.1 und 3.2). Diese Annahmen sind z.T. kritisch, weil kognitionspsychologische Daten typischerweise aus der Konfrontation mehrerer Vpn mit mehreren Aufgaben oder Items resultieren. Derartige Daten können wegen möglicher interindividueller Unterschiede und Schwierigkeitsdifferenzen der Items nicht ohne weiteres als unabhängig und identisch multinomialverteilt aufgefaßt werden. Die Robustheit multinomialer Datenanalysen gegenüber Verletzungen dieser Annahmen ist daher ein wichtiges Thema, das nicht vernachlässigt werden darf.

Das Konzept „parametrisiertes multinomiales Modell" wurde in Abschnitt 3.3 ebenso eingeführt wie die Begriffe der Saturiertheit, Überparametrisiertheit und Unterparametrisiertheit von Modellen. Allgemeine wie auch binäre MVB-Modelle sind spezielle parametrisierte multinomiale Modelle, die in den Abschnitten 3.4 und 3.5 sowohl formal als auch anhand von konkreten Beispielen dargestellt wurden. Unterschiedliche Notationen und Baumdiagramm-Repräsentationen für MVB-Modelle wurden vorgestellt. Abschnitt 3.6 schließlich machte deutlich, daß Ereignisbaumdiagramme zwar eindeutig in MVB-Modelle, aber MVB-Modelle nicht immer eindeutig in Baumdiagramme überführbar sind. Folglich können die Ergebnisse statistischer Analysen von MVB-Modellen nicht ohne weiteres in Termini eines ganz bestimmten Ereignisbaumdiagramms interpretiert werden.

4. Analyse und Evaluation multinomialer Modelle

Im vorliegenden Kapitel wird die statistische Analyse parametrisierter multinomialer Modelle unter den „Standardannahmen" geschildert, d.h. unter Annahme der Gültigkeit statistischer Voraussetzungen, die zum Teil im letzten Kapitel — insbesondere in den Abschnitten 3.1 und 3.2 — bereits geschildert wurden, zum Teil im vorliegenden Kapitel noch darzustellen sind. Die naheliegende Frage, ob die vorzustellenden Analyseverfahren sich auch bei Verletzungen der zugrundeliegenden Annahmen noch adäquat (d.h. „robust") verhalten, ist nicht Gegenstand dieses Kapitels. Ich komme in Kapitel 5 darauf zurück.

Die ersten beiden Abschnitte des vorliegenden Kapitels beschäftigen sich mit dem Problem der Bestimmung bzw. Schätzung der Modellparameterwerte, wobei das zugrundeliegende parametrisierte multinomiale Modell zunächst einmal als gültig unterstellt wird. Es wird also angenommen, daß Modellparameterwerte $\boldsymbol{\theta}$ existieren, welche die in einem konkreten Anwendungsfall gegebenen Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ „vorhersagen", wenn man sie in die Modellgleichungen einsetzt. Abschnitt 4.1 behandelt die Bedingungen, unter denen die Modellparameter eindeutig bestimmt werden können, sofern die Kategoriewahrscheinlichkeiten bekannt sind. Das scheint auf den ersten Blick ein für die Anwendungspraxis irrelevantes Problem zu sein, weil die Kategoriewahrscheinlichkeiten typischerweise unbekannt sind; stattdessen liegen lediglich Kategoriehäufigkeiten $\boldsymbol{y}$ aus einer bestimmten Stichprobe vor. Dennoch ist Abschnitt 4.1 für die Anwendungspraxis wichtig, denn es werden die Begriffe der globalen und lokalen Identifizierbarkeit eines Modells eingeführt, die für die folgenden Abschnitte von zentraler Bedeutung sind. Darauf aufbauend behandelt Abschnitt 4.2 das Problem der Ableitung möglichst guter Schätzwerte $\hat{\boldsymbol{\theta}}$ der Modellparameter $\boldsymbol{\theta}$ aus Stichprobenhäufigkeiten $\boldsymbol{y}$ (Punktschätzungen) und das Problem der Bestimmung von Bereichsschätzern (Konfidenzintervallen) für die Modellparameter.

Abschnitt 4.3 hat das Problem der statistischen Modellgeltungsprüfung zum Gegenstand, das insofern zentral ist, als Modellparameter bzw. Parameterschätzungen nur bei Modellgültigkeit sinnvoll interpretiert werden können. Globale Modellgeltungstests und Tests von spezifischen Parameterhypothesen und Submodellen werden vorgestellt. Hierbei wird besonderes Gewicht auf die simultane Kontrolle der α- und der β-Fehlerwahrscheinlichkeiten bei der statistischen Entscheidung gelegt. Die folgenden Abschnitte 4.4 und 4.5 beschäftigen sich mit dem Problem der Entscheidung zwischen Modellen, die in Modellgeltungstests gleichermaßen gut oder nahezu gleich gut abschneiden. Die Probleme der Äquivalenz und der Datenäquivalenz von Modellen werden ebenso dargestellt wie das eng verwandte Problem der Implikationsbeziehung zwischen Modellen mit gleicher Parameteranzahl (Abschnitt 4.4). Anschließend wird die systematische Modellvali-

dierung als zusätzliche Evaluationsmethode vorgestellt, um Probleme der Datenäquivalenz und verwandte Probleme zu lösen (Abschnitt 4.5). Schließlich wird in Abschnitt 4.6 untersucht, welche konstruktiven Lösungsvorschläge angeboten werden können, wenn ein bestimmtes parametrisiertes multinomiales Modell zwar psychologisch gut begründbar ist, jedoch keine eindeutigen Parameterschätzungen erlaubt. Abschnitt 4.7 gibt eine knappe Übersicht über die wichtigsten Computerprogramme, die derzeit für die Analyse und Evaluation kognitionspsychologisch motivierter multinomialer Modelle — insbesondere für MVB-Modelle — zur Verfügung stehen. Eine Zusammenfassung schließt das Kapitel ab (Abschnitt 4.8).

4.1 Identifizierbarkeit

Parametrisierte multinomiale Modelle sind durch eine bestimmte Abbildung f des Parameterraumes $\boldsymbol{\Omega}$ in den Raum der Kategoriewahrscheinlichkeiten $\boldsymbol{\Gamma}$ definiert (vgl. Abschnitt 3.3). Von der Art dieser Abbildung hängt es ab, ob das Modell identifizierbar ist: Wenn unterschiedliche Parametervektoren $\boldsymbol{\theta}$ und $\boldsymbol{\theta}'$ aus $\boldsymbol{\Omega}$ immer auch auf unterschiedliche Kategoriewahrscheinlichkeitsvektoren $\boldsymbol{p} = f(\boldsymbol{\theta})$ und $\boldsymbol{p}' = f(\boldsymbol{\theta}')$ aus $\boldsymbol{\Gamma}$ abgebildet werden, dann heißt das Modell „global identifizierbar". Sofern ein parametrisiertes multinomiales Modell global identifizierbar ist, kann man daher für jeden Vektor $\boldsymbol{p}$ von Kategoriewahrscheinlichkeiten, der bei Gültigkeit des Modells vorkommen kann, eindeutig angeben, welche Modellparameter $\boldsymbol{\theta}$ die betreffenden Kategoriewahrscheinlichkeiten „erzeugt" haben (allgemein: Bamber & van Santen, 2000; Hurwicz, 1950; Koopmans & Reiersøl, 1950).

Identifizierbarkeit ist eine ganz wesentliche Modelleigenschaft, denn bei Anwendungen eines Modells gilt das primäre Interesse meistens den zugrundeliegenden Modellparameterwerten. Wie der nächste Abschnitt zeigen wird, stellt die globale Identifizierbarkeit eines Modells nicht nur sicher, daß ein eindeutiger Rückschluß von den (zumeist unbekannten) Kategoriewahrscheinlichkeiten auf den „erzeugenden" Parametervektor $\boldsymbol{\theta}$ möglich ist; sie impliziert zusätzlich, daß für beliebige beobachtete Häufigkeiten $\boldsymbol{y}$ in einer Stichprobe fast sicher eindeutige Schätzungen $\hat{\boldsymbol{\theta}}$ der Modellparameter $\boldsymbol{\theta}$ existieren. Ist umgekehrt globale Identifizierbarkeit nicht gegeben, existieren oft keine eindeutigen Parameterschätzungen zu gegebenen Daten, so daß der Sinn der Modellkonstruktion — von Ausnahmefällen abgesehen (vgl. u.a. Abschnitt 4.6) — in Frage gestellt ist. Der Prüfung der Identifizierbarkeit und der Unterbreitung von Vorschlägen zur Überwindung von Identifizierbarkeitsproblemen kommt daher eine zentrale Rolle bei der Anwendung multinomialer Modelle zu.

4.1.1 Globale versus lokale Identifizierbarkeit

Globale Identifizierbarkeit eines parametrisierten multinomialen Modells kann unter Rückgriff auf das Konzept der Injektivität (Eineindeutigkeit) einer Abbildung definiert werden (vgl. Abschnitt 2.4). Ein parametrisiertes multinomiales Modell $\langle \mathrm{M}(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$ ist demnach global identifizierbar, wenn f injektiv ist, d.h. wenn gilt

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in V\colon f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}') \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}' \ . \qquad (4.1.1\text{-}1)$$

Ist $f(\boldsymbol{\Omega}) \subseteq \boldsymbol{\Gamma}$ das Bild von $\boldsymbol{\Omega}$ unter f, d.h. die Teilmenge aller Kategoriewahrscheinlichkeitsvektoren, auf die ein Element des Parameterraumes abgebildet wird,

$$f(\boldsymbol{\Omega}) = \left\{ \boldsymbol{p} \in \boldsymbol{\Gamma} \mid \exists \boldsymbol{\theta} \in \Omega\colon f(\boldsymbol{\theta}) = \boldsymbol{p} \right\} \subseteq \boldsymbol{\Gamma} \ , \qquad (4.1.1\text{-}2)$$

so sichert die Injektivität von f die Existenz der Umkehrabbildung $f^{-1}\colon f(\boldsymbol{\Omega}) \rightarrow \boldsymbol{\Omega}$, die jedem Element des Bildes von $\boldsymbol{\Omega}$ sein Urbild zuordnet. Die inverse Abbildung ist durch die Abbildungsvorschrift $f^{-1}\colon \boldsymbol{p} = f(\boldsymbol{\theta})$ a $\boldsymbol{\theta}$ definiert.

Von der globalen Identifizierbarkeit muß eine schwächere Form der Identifizierbarkeit unterschieden werden: die sogenannte lokale Identifizierbarkeit (Bamber & van Santen, 2000; McHugh, 1956; Wald, 1950). Ein parametrisiertes multinomiales Modell $\langle \mathrm{M}(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$ ist an der Stelle $\boldsymbol{\theta^o} \in \boldsymbol{\Omega}$ lokal identifizierbar, wenn es in einer offenen Umgebung $\mathbf{U}$ von $\boldsymbol{\theta^o}$ identifizierbar ist. Etwas ausführlicher: Gibt es eine Teilmenge $\mathbf{U}$ des Parameterraums, die $\boldsymbol{\theta^o}$ enthält (d.h. $\boldsymbol{\Omega} \supseteq \mathbf{U} \supset \{\boldsymbol{\theta^o}\}$), und ist f bezüglich dieses Argumentbereichs $\mathbf{U}$ injektiv,

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbf{U}\colon f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}') \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}' \ , \qquad (4.1.1\text{-}3)$$

dann heißt das Modell $\langle \mathrm{M}(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$ „lokal identifizierbar an der Stelle $\boldsymbol{\theta^o} \in \boldsymbol{\Omega}$". Ein Vergleich von (4.1.1-1) und (4.1.1-3) zeigt, daß globale Identifizierbarkeit automatisch lokale Identifizierbarkeit für einen beliebigen Punkt $\boldsymbol{\theta} \in \boldsymbol{\Omega}$ impliziert, da man ja einfach $\mathbf{U} = \boldsymbol{\Omega}$ wählen kann. Umgekehrt muß lokale Identifizierbarkeit für einen bestimmten Punkt $\boldsymbol{\theta^o} \in \boldsymbol{\Omega}$ natürlich nicht zwangsläufig globale Identifizierbarkeit nach sich ziehen. Selbst wenn für alle denkbaren Parametervektoren $\boldsymbol{\theta} \in \boldsymbol{\Omega}$ lokale Identifizierbarkeit gegeben ist, garantiert das noch nicht automatisch globale Identifizierbarkeit. Man betrachte beispielsweise ein einfaches Binomialmodell $\langle \mathrm{M}(1, 2, N, \pi), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$, mit eindimensionalem Parameterraum $\boldsymbol{\Omega} = \mathbb{R}$, Raum der Kategoriewahrscheinlichkeiten $\boldsymbol{\Gamma} = [0, 1]$ und einer Abbildung $f\colon \boldsymbol{\Omega} \rightarrow \boldsymbol{\Gamma}$, die durch die Modellgleichung

$$f(\theta) = p_{1,1}(\theta) = \theta - g(\theta) \qquad (4.1.1\text{-}4)$$

definiert wird. Hierbei soll $g(\theta)$ den ganzzahligen Anteil von θ bezeichnen (die „Zahl vor dem Komma"), sofern $\theta \geq 0$; ist dagegen $\theta < 0$, so sei $g(\theta)$ der ganzzahlige Anteil von θ minus 1. Der Graph dieser Modellgleichung entspricht anschaulich

einem „Sägezahnmuster" entlang der θ-Achse, welches nur Werte im halboffenen Intervall [0, 1) annimmt. Das Modell ist offensichtlich nicht global identifizierbar, da auf jede Kategoriewahrscheinlichkeit π im Intervall [0, 1) immer unendlich viele unterschiedliche Parameterwerte abgebildet werden: Ist $f(\theta) = \pi$, so ist auch $f(\theta) + z = \pi$, sofern z eine beliebige ganze Zahl ist. Dennoch ist das Modell für alle $\theta \in \mathbf{\Omega}$ lokal identifizierbar, denn man kann immer eine Umgebung **U** zu jedem θ so wählen, daß die durch Gleichung (4.1.1-4) definierte Abbildung bezüglich des Argumentbereichs **U** injektiv wird. Das Intervall $\mathbf{U} = (\theta - 0.4, \theta + 0.4)$ erfüllt beispielsweise diese Bedingung.

Etwas weniger exotische Beispiele sind Latent-Class-Modelle ähnlich dem in Abschnitt 2.5 besprochenen Modell zum Experiment von Levine. Nehmen wir der Einfachheit halber an, wir würden ein unrestringiertes Latent-Class-Modell mit zwei latenten Klassen für eine 2×2×2×2-Kontingenztafel mit 15 freien Kategoriewahrscheinlichkeiten formulieren. Dieses Modell umfaßt für jede der beiden Klassen vier Antwortwahrscheinlichkeiten für die vier dichotomen Variablen und zusätzlich eine freie Klassenwahrscheinlichkeit, insgesamt also neun freie Parameter. Obwohl das Modell deutlich unterparametrisiert ist, ist es nicht global identifizierbar. Durch Permutation der Klassenindizes (d.h. durch Änderung der Numerierung der beiden latenten Klassen) erhält man nämlich problemlos neue Parametervektoren, welche die gleichen Antwortmusterwahrscheinlichkeiten vorhersagen wie die ursprünglichen Parametervektoren. Dennoch kann man zu fast allen Vektoren $\boldsymbol{\theta^0} \in \mathbf{\Omega}$ eine Umgebung **U** finden, so daß die Modellgleichungen diesbezüglich eine injektive Abbildung definieren. Das Modell ist folglich fast überall im Parameterraum lokal identifizierbar, aber nicht global identifizierbar. (Die Einschränkung „fast überall" ist hier notwendig, weil nicht nur die Parameterwerte 0 und 1 am Rande des Parameterraumes ausgeschlossen werden müssen, sondern auch Fälle, in denen (a) die beiden latenten Klassen exakt gleich groß sind oder (b) die Antwortwahrscheinlichkeit innerhalb der beiden Klassen auf alle vier Variablen exakt gleich sind.)

Die Unterscheidung lokaler und globaler Identifizierbarkeit ist insofern wichtig, als lokale Identifizierbarkeit in der näheren Umgebung des (zumeist unbekannten) wahren Parametervektors $\boldsymbol{\theta^0} \in \mathbf{\Omega}$, der einer Stichprobe zugrundeliegt, zur statistischen Begründung der Parameterschätzung und der Modellprüfung ausreichen kann, wenn noch einige weitere Regularitätsbedingungen erfüllt sind. Ich werde in den folgenden Abschnitten hierauf zurückkommen (vgl. die Abschnitte 4.2.2 und 4.3.1). Vorerst ist nur wichtig, daß globale Identifizierbarkeit nicht die conditio sine qua non für statistische Analysen multinomialer Modelle ist. Lokale Identifizierbarkeit reicht für viele Zwecke aus. Natürlich ist der analytische Nachweis der globalen Identifizierbarkeit insofern dennoch hilfreich und nützlich, als damit zum einen die lästige empirische Überprüfung der lokalen

Identifizierbarkeit überflüssig wird und zum zweiten so gut wie sicher ist, daß das betreffende Modell auf beliebige Datensätze so anwendbar ist, daß eindeutige Parameterschätzungen resultieren.

Obwohl der Begriff der globalen Identifizierbarkeit eines parametrisierten multinomialen Modells mit dem in Abschnitt 2.4 eingeführten Begriff der globalen Identifizierbarkeit eines deterministischen Modells insofern Ähnlichkeit hat, als in beiden Konzepten die Injektivität einer Abbildung eine zentrale Rolle spielt, handelt es sich um ganz unterschiedliche Konzepte, die keinesfalls miteinander verwechselt werden dürfen. Die Mengen, die aufeinander abgebildet werden, sind bei deterministischen und stochastischen Modellen grundverschieden. Im Falle eines deterministischen psychologischen Modells — wie es in Abschnitt 2.4 exemplarisch vorgestellt wurde — wird eine Menge latenter Zustände (z.B. ein „Hypothesenraum") auf eine Menge manifester Beobachtungstatbestände (z.B. einen „Antwortraum") abgebildet; beide Mengen sind nichtnumerisch, und zum Antwortraum ist auch keine Wahrscheinlichkeitsverteilung definiert. Bei stochastischen Modellen handelt sich jedoch um zwei Zahlenmengen $\mathbf{\Omega}$ und $\mathbf{\Gamma}$, die aufeinander abgebildet werden. Eine Verbindung zu den Konzepten „latenter Zustandsraum" und „Antwortraum" besteht insofern, als die Elemente von $\mathbf{\Omega}$ — jedenfalls bei MVB-Modellen — als Wahrscheinlichkeiten latenter Zustände und die Elemente von $\mathbf{\Gamma}$ als Wahrscheinlichkeiten manifester Beobachtungstatbestände interpretierbar sind. Das bedeutet aber keineswegs, daß die Identifizierbarkeit eines stochastischen Modells mit der Identifizierbarkeit des „korrespondierenden" deterministischen Modells gleichzusetzen ist; es bedeutet nicht einmal, daß die Existenz einer Abbildung von $\mathbf{\Omega}$ in $\mathbf{\Gamma}$ — sei sie nun injektiv oder nicht — die Existenz einer Abbildung zwischen den korrespondierenden Zustands- und Antworträumen voraussetzt.

Betrachten wir ein Beispiel. Um es ganz einfach zu halten, analysieren wir nur die Situation für das Ein-Hochschwellen-Modell (vgl. Abschnitt 3.5.2) und auch hier nur den Fall alter Items in der Testsituation. Neue Items sollen vorerst unberücksichtigt bleiben. Der psychologische Zustandsraum $\mathbf{Z}$ umfaßt in diesem Fall zwei latente Vertrautheitszustände, $\mathbf{Z}$ = {überschwellig, unterschwellig}, und der Verhaltensraum $\mathbf{R}$ umfaßt ebenfalls zwei potentiell beobachtbare Antworten, $\mathbf{R}$ = {„alt", „neu"}. Das 1HT-Modell nimmt offenbar *keine* Abbildung zwischen den Mengen $\mathbf{Z}$ und $\mathbf{R}$ an, denn dem unterschwelligen Vertrautheitszustand ist keine eindeutige Antwort zugeordnet; beide Antworten aus $\mathbf{R}$ sind in diesem Fall möglich. Man kann folglich auch nicht davon sprechen, daß die Vertrautheitszustände aus $\mathbf{Z}$ anhand der Antworten aus $\mathbf{R}$ „operationalisiert" werden, sei es nun durch bilaterale Reduktionssätze oder eine andere Methode (vgl. Abschnitt 2.4). Dennoch kann sehr wohl eine Abbildung zwischen dem zweidimensionalen Parameterraum $\mathbf{\Omega} = \{(r_{1\mathrm{ht}}, b_{1\mathrm{ht}}) \mid 0 \leq r_{1\mathrm{ht}}, b_{1\mathrm{ht}} \leq 1\}$ und dem (bei Beschränkung auf alte Items)

eindimensionalen Raum der Kategoriewahrscheinlichkeiten $\mathbf{\Gamma} = \{p_{11} \mid 0 \leq p_{11} \leq 1\}$ gegeben sein, beispielsweise die durch Gleichung (3.5.2-1) definierte Abbildung.

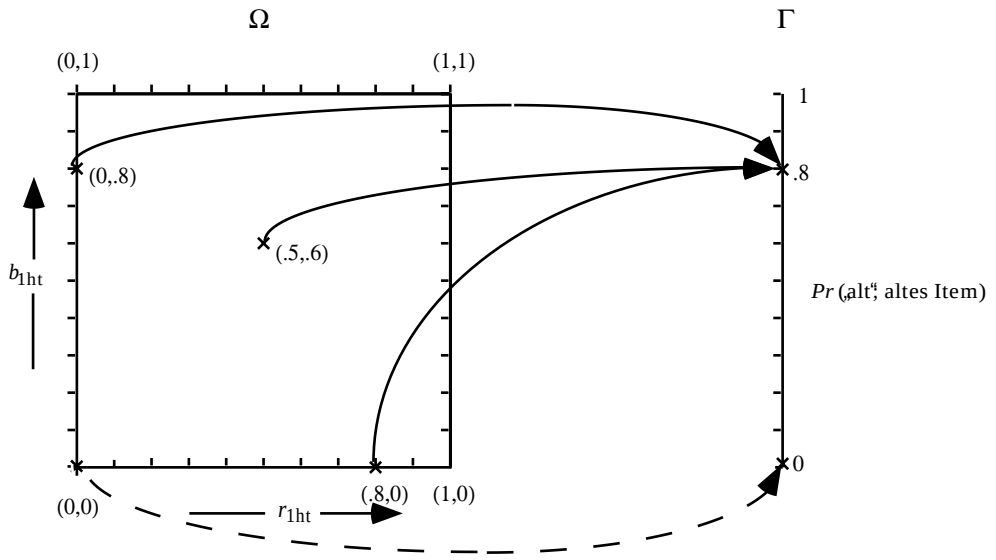

Abbildung 4.1:
Veranschaulichung der Tatsache, daß das Ein-Hochschwellen-Modell für Ja-Nein-Rekognitionstests bei Beschränkung auf alte Items nicht identifizierbar ist.

Die durch Gleichung (3.5.2-1) definierte Abbildung ist nicht injektiv, weil z.B. der Punkt $p_{11} = .8 \in \mathbf{\Gamma}$ von ganz unterschiedlichen Werten des Parameterraumes aus erreicht werden kann (vgl. Abb. 4.1). Folglich ist das 1HT-Modell bei Beschränkung auf alte Items nicht global identifizierbar. Ist wenigstens lokale Identifizierbarkeit gegeben? Dies ist ebenfalls nicht der Fall, auch wenn der Punkt $p_{11} = .0 \in \mathbf{\Gamma}$ nur ausgehend vom Punkt $(0, 0) \in \mathbf{\Omega}$ erreicht werden kann. Lokale Identifizierbarkeit würde verlangen, daß es weitere Parameterwerte „in der Nähe" von (0, 0) gibt, die ebenfalls eineindeutig bestimmten Trefferraten zugeordnet werden. Ist jedoch mindestens einer der Parameter r_{1ht} oder b_{1ht} größer als null, so gibt es zwangsläufig immer mehrere Paare von Parameterwerten (r_{1ht}, b_{1ht}), denen die gleiche Trefferwahrscheinlichkeit p_{11} zugeordnet ist.

4.1.2 Notwendige und hinreichende Bedingungen für Identifizierbarkeit

Bislang wurde die Beurteilung der globalen und lokalen Identifizierbarkeit anhand konkreter Modelle illustriert. Nützlich ist es, wenn man notwendige und/oder hinreichende Bedingungen der globalen oder lokalen Identifizierbarkeit angegeben kann, die auf möglichst große Modellklassen zutreffen. Generelle und zugleich einfach anwendbare Resultate zur Identifizierbarkeit parametrisierter multinomialer Modelle sind aber leider rar. Wenn wir uns zunächst auf MVB-Modelle beschränken, so ist intuitiv klar, daß ein solches Modell im allgemeinen nur für Parameterwerte aus einer offenen Teilmenge $\mathbf{\Omega}^* \subseteq (0, 1)^S$ global identifizierbar sein

kann, d.h. die Extremwerte 0 und 1 der Parameter müssen ausgeschlossen werden. Wie Abbildung 4.1 zeigt, gibt es zwar gelegentlich einzelne Punkte am Rand des Parameterraumes, die eineindeutig bestimmten Kategoriewahrscheinlichkeiten zugeordnet werden. Allerdings sind dies Ausnahmefälle. Wenn mindestens ein Parameter in einem Produktterm (d.h. in einem Ast eines Verarbeitungsbaumes) den Wert 0 annimmt, können alle anderen Parameter, die ausschließlich in diesem Ast vorkommen, beliebige Werte annehmen, ohne daß sich dadurch an den vorhergesagten Kategoriewahrscheinlichkeiten etwas ändert. Folglich ist ein Modell mit diesen Eigenschaften nicht identifizierbar. Wenn umgekehrt mindestens ein Parameter den Wert 1 annimmt, muß es irgendwo anders im Baum mindestens einen komplementären Parameter mit dem Wert 0 geben; das zuvor genannte Argument greift dann erneut.

Eine sehr allgemeine Bedingung, die auch für Modelle außerhalb des MVB-Rahmens gilt, ist folgende: Modelle, die global identifizierbar sind, sind niemals überparametrisiert, d.h. eine notwendige Bedingung der globalen Identifizierbarkeit ist

$$S \leq M = \sum_{k=1}^{K} (J_k - 1) \,. \qquad (4.1.2\text{-}1)$$

Mit anderen Worten: die Dimensionalität S des Parameterraumes darf nicht größer als die Dimensionalität des Raumes möglicher Kategoriewahrscheinlichkeiten sein (vgl. Abschnitt 3.3), sonst ist das Modell zwangsläufig nicht identifizierbar.[27] Dieses Resultat impliziert beispielsweise, daß nicht nur das 1HT-Modell, sondern auch jedes beliebige andere parametrisierte multinomiale Modell zum Ja-Nein-Rekognitionstest (beschränkt auf alte Items) nicht identifizierbar ist, sofern es $S \geq 2$ Modellparameter umfaßt.

Die Bedingung (4.1.2-1) ist zwar notwendig, jedoch keineswegs zugleich auch hinreichend für globale Identifizierbarkeit. Ein Beispiel hierfür ist das Quellendiskriminationsmodell 6a von Batchelder und Riefer (1990, S. 552), das 3·(3-1) = 6 funktional unabhängige Kategoriewahrscheinlichkeiten mit S = 6 funktional unabhängigen Parametern beschreibt. Obwohl die Bedingung (4.1.2-1) eingehalten wird, ist dieses Modell dennoch nicht identifizierbar, und zwar weil zwei der sechs Modellgleichungen identisch sind, so daß die betreffenden Kategoriewahrscheinlichkeiten p_{12} und p_{23} (vgl. Gleichung 10c und 11c bei Batchelder und Riefer, 1990, S. 551) ebenfalls identisch sein müssen. Als Konsequenz aus dieser Modelleigenschaft kann das Bild des Parameterraumes $f_{6a}(\mathbf{\Omega}) \subset \mathbf{\Gamma}$ im Gegensatz zu $\mathbf{\Omega}$ und $\mathbf{\Gamma}$

[27] Wäre es identifizierbar, so müßte eine Umkehrabbildung f^{-1}: $f(\mathbf{\Omega}) \rightarrow \mathbf{\Omega}$ existieren, die surjektiv ist, so daß auf *jedes* Element in $\mathbf{\Omega}$ ein Element aus $f(\mathbf{\Omega})$ abgebildet wird. Die Abbildung f^{-1} kann aber nicht surjektiv sein, da wegen der geringeren Dimensionalität von $f(\mathbf{\Omega}) \subset \mathbf{\Gamma}$ im Vergleich zu $\mathbf{\Omega}$ das Urbild $f^{-1}(f(\mathbf{\Omega}))$ eine echte Teilmenge von $\mathbf{\Omega}$ sein muß.

nicht sechs-, sondern nur fünfdimensional sein. Die modelldefinierende Abbildung f_{6a} ist folglich zwangsläufig nicht injektiv. Ich werde auf dieses und weitere Quellendiskriminationsmodelle in Kapitel 7.1.2 zurückkommen.

Für binäre MVB-Modelle haben Hu und Batchelder (1994) eine *hinreichende* Bedingung der globalen Identifizierbarkeit angegeben: Ein MVB-Modell ist immer dann global identifizierbar, wenn (1) alle Äste des Modells in unterschiedliche Kategorien $C_{k,j}$ einmünden und (2) deren Wahrscheinlichkeiten sämtlich positiv sind, so daß Gleichung (3.5-1) mit $I_{k,j} = 1$ und $Pr(C_{k,j,}) > 0$ für alle Versuchsbedingungen k sowie Kategorien j gilt. Dieses Ergebnis läßt sich problemlos auf allgemeine MVB-Modelle übertragen (vgl. Abschnitt 3.5). Die globale Identifizierbarkeit von MVB-Modellen mit „beobachtbaren Ästen" folgt aus der Tatsache, daß sich deren Parameter immer eindeutig als Funktion der Astwahrscheinlichkeiten $Pr^*(A_{k,j,i})$ darstellen lassen, sofern diese sämtlich größer null sind:

$$\theta_s = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{i=1}^{I_{k,j}} a_{k,j,i,s} \cdot Pr^*(A_{k,j,i})}{\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{i=1}^{I_{k,j}} (a_{k,j,i,s} + b_{k,j,i,s}) \cdot Pr^*(A_{k,j,i})} \qquad (4.1.2\text{-}2)$$

In Gleichung (4.1.2-2) sind $a_{k,j,i,s}$ und $b_{k,j,i,s}$ die bereits in Abschnitt 3.5 eingeführten Zählvariablen. Der Beweis dieser Gleichung folgt unmittelbar aus einem Resultat von Hu und Batchelder (1994, S. 26, Gleichung 10).[28] Aufgrund der Voraussetzung „beobachtbarer Äste" kann man für die dort angegebenen Astwahrscheinlichkeiten auch Kategoriewahrscheinlichkeiten $p_{k,j}$ schreiben, womit die Umkehrfunktion f^{-1} angegeben wäre, die den Kategoriewahrscheinlichkeiten die zugrundeliegenden Parameterwerte zuordnet. Die Existenz von f^{-1} impliziert die Injektivität von f. Folglich muß ein Modell, das die genannten Bedingungen erfüllt, global identifizierbar sein.

Diese hinreichende Bedingung für die globale Identifizierbarkeit von MVB-Modellen ist sehr nützlich, sofern die Prämisse „beobachtbarer Äste" erfüllt ist. Für die Praxis dürfte sie dennoch nicht so bedeutsam wie die zuvor genannte notwendige Bedingung sein, da MVB-Modelle mit beobachtbaren Ästen sehr seltene Ausnahmefälle sind. Tatsächlich ist mir kein einziges Modell aus der einschlägigen Literatur bekannt, das diese Bedingung erfüllt. Dennoch ist das Ergebnis hilfreich, denn der skizzierte Beweis macht ein allgemein anwendbares Prinzip des globalen Identifizierbarkeitsnachweises deutlich: Man kann versuchen, eine Umkehrabbil-

[28] Hu und Batchelder (1994, Gleichung 10) haben gezeigt, daß Gleichung (4.1.2-2) gilt, wenn man die Parameter θ_s und die Astwahrscheinlichkeiten $Pr(A_{k,j,i})$ durch ihre Maximum-Likelihood-Schätzer ersetzt. Da Maximum-Likelihood-Schätzer konsistent sind (vgl. Abschnitt 4.2), muß Gleichung (4.1.2-2) auf für die Beziehung zwischen den zugrundeliegenden Parametern gelten.

dung zu den Modellgleichungen zu konstruieren, gegebenenfalls unter Zugrundelegung bestimmter Randbedingungen. Gelingt dies, ist damit der Beweis der Identifizierbarkeit eines Modells unter den zugrunde gelegten Randbedingungen erbracht. Bei voll parametrisierten Modellen kann dies durchaus eine recht einfache Übung sein: Man muß versuchen, die Modellgleichungen nach den Parametern aufzulösen, so daß jeder Modellparameter als Funktion der Kategoriewahrscheinlichkeiten dargestellt wird. Computerprogramme wie beispielsweise die SOLVE-Funktion aus „Mathematica" (Wolfram, 1992) können hierbei hilfreiche Dienste leisten. Konkrete Anwendungen dieser Methode des Identifizierbarkeitsnachweises findet man z.B. bei Buchner, Erdfelder und Vaterrodt-Plünnecke (1995), Erdfelder und Buchner (1995) sowie Buchner, Steffens, Erdfelder und Rothkegel (1997). In allen genannten Arbeiten wurde die Methode zum Nachweis der Identifizierbarkeit verschiedener Prozeßdissoziationsmodelle herangezogen (vgl. dazu auch Kapitel 7.1.3).

Gelingt die Konstruktion einer Umkehrabbildung zu den Modellgleichungen in Form eines Gleichungssystems nicht, so muß mit der Möglichkeit gerechnet werden, daß das Modell nicht identifizierbar ist. Zwingend ist dies allerdings nicht: Ein Modell kann global identifizierbar sein, ohne daß ein explizites Gleichungssystem, das die Parameter als Funktion der Kategoriewahrscheinlichkeiten darstellt, existieren muß. In solchen Fällen lassen sich die Werte der Umkehrabbildung f^{-1} nur durch ein iteratives Rechenverfahren bestimmen (vgl. Abschnitt 4.2.1).

Bevor man versucht, den Identifizierbarkeitsbeweis über den Nachweis der Injektivität gemäß (4.1.1-1) zu führen, kann es nützlich sein, zunächst zu prüfen, ob die Identifizierbarkeit des betreffenden Modells a priori ausgeschlossen werden kann. Eine hinreichende Bedingung für die Nichtidentifizierbarkeit von stochastischen Modellen haben Bamber und van Santen (1985; 2000) angegeben: Wenn (a) die Modellgleichungen nach den Parametern stetig differenzierbar sind und (b) der maximale Rang ρ_{max} der Jacobischen Matrix $\mathbf{J}(\boldsymbol{\theta}, f)$ der Modellgleichungen kleiner als S ist,

$$\rho_{max} = \max_{\theta \in \Omega} \cdot \rho\left[\mathbf{J}(\theta, f)\right] < S \,, \qquad (4.1.2\text{-}3)$$

dann ist das Modell an keiner Stelle lokal identifizierbar und folglich auch nicht global identifizierbar (vgl. Bamber & van Santen, 1985, S. 464, Proposition 20)[29]. Die

[29]Bamber und van Santen (1985, S. 464f) beweisen explizit nur die schwächere Behauptung, daß das Modell unter der Bedingung (4.1-7) an solchen Stellen des Parameterraumes nicht lokal identifizierbar ist, an denen die Jacobische Matrix ihren maximalen Rang $\rho_{max} < S$ annimmt. Die stärkere Behauptung, daß das Modell an *keiner* Stelle des Parameterraumes lokal identifizierbar ist, folgt jedoch aus den gleichen Voraussetzungen. Hierzu beweist man, daß wenn ein Modell an keiner Stelle mit Rang i der Jacobischen Matrix lokal identifizierbar ist, es dann auch an keiner Stelle mit Rang i–1 der Jacobischen Matrix identifizierbar sein kann. Über vollständige Induktion läßt sich dann für Punkte $\boldsymbol{\theta}$ des Parameterraumes mit beliebigem Rang $\rho[\mathbf{J}(\boldsymbol{\theta}, f)] < S$ zeigen, daß das Modell an der Stelle $\boldsymbol{\theta}$ nicht lokal identifizierbar ist. Dies ist aber gleichbedeutend mit der Aussage, daß das Modell an keiner

Jacobische Matrix $\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)$ umfaßt die ersten Ableitungen der Modellgleichungen nach den Modellparametern, evaluiert an der Stelle $\boldsymbol{\theta}^{\mathbf{o}}$ des Parameterraumes (einführend z.B. Smith, 1998, Appendix A). Wenn $\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)$ vollen Rang S hat, ist f in einer offenen Umgebung von $\boldsymbol{\theta}^{\mathbf{o}}$ invertierbar (Smith, 1998, S. 133). Ein voller Rang $\rho[\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)] = S$ ist also eine hinreichende Bedingung für die lokale Identifizierbarkeit des Modells an der Stelle $\boldsymbol{\theta}^{\mathbf{o}}$ (vgl. auch Formann, 1984, S. 21 ff). Umgekehrt ist ein voller Rang jedoch keine notwendige Bedingung für lokale Identifizierbarkeit an der Stelle $\boldsymbol{\theta}^{\mathbf{o}}$. Das einparametrige Modell $f(\theta) = \theta^3$ mit $\theta \in \mathbb{R}$ und $S = 1$ ist beispielsweise nicht nur lokal, sondern sogar global identifizierbar, obwohl für $\theta = 0$ die Jacobische Matrix nicht vollen Rang hat, da $\rho[\mathbf{J}(\theta = 0, f)] = 0 < S$. Ist der Rang der Jacobischen Matrix also nur an einigen Stellen des Parameterraumes kleiner als S, so kann man daraus noch nicht auf Nichtidentifizierbarkeit schließen. Erst wenn der Rang für alle möglichen Parameterwerte — der maximale Rang — kleiner als S ist (vgl. Gleichung (4.1.2-3)), so folgt daraus die (lokale und globale) Nichtidentifizierbarkeit des Modells.

Die o.g. Differenzierbarkeitsbedingung (a), welche für das Nichtidentifizierbarkeitsargument vorausgesetzt werden muß, ist bei binären und nicht binären MVB-Modellen aufgrund der Definition der Modellgleichungen automatisch erfüllt, sofern wir uns auf den offenen Parameterraum $\boldsymbol{\Omega}^* \subseteq (0, 1)^S$ beschränken, also die Extremwerte 0 und 1 der Parameter ausschließen.

Die hier als Bedingung (4.1.2-3) wiedergegebene Proposition 20 von Bamber und van Santen (1985) ist ein sehr allgemeines Resultat. Sie impliziert u.a. die weiter oben aufgestellte Bedingung (4.1.2-1), wonach identifizierbare Modelle niemals überparametrisiert sein dürfen. Ist nämlich $S > M$, so ist zwangsläufig auch $S > \rho_{max}$, woraus nach (4.1.2-3) die Nichtidentifizierbarkeit folgt.

4.1.3 Entscheidungen über Identifizierbarkeitsfragen

Welche Vorgehensweise empfiehlt sich, wenn man in der Praxis Entscheidungen über die lokale oder globale Identifizierbarkeit von multinomialen Modellen fällen will? Proposition 20 von Bamber und van Santen (1985) kann hier ein nützlicher Ausgangspunkt sein. Sie ist natürlich nur dann hilfreich, wenn man ρ_{max} in Gleichung (4.1.2-3), den maximalen Rang der Jacobischen Matrix, numerisch bestimmen kann. Im Falle parametrisierter multinomialer Modelle stößt man hierbei jedoch selten auf Probleme. Sei $\boldsymbol{\theta} = (\theta_1, ..., \theta_S, ..., \theta_S) \in \boldsymbol{\Omega}$ ein beliebiger Parametervektor und $f(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_J(\boldsymbol{\theta})) \in \boldsymbol{\Gamma}$ der Vektor der Kategoriewahrscheinlichkeiten, auf den $\boldsymbol{\theta}$ abgebildet wird. Bei verbundenen multinomialen Modellen sei J dabei die Gesamtzahl der Kategoriewahrscheinlichkeiten über alle K Bedingungen

---

Stelle des (offenen) Parameterraumes lokal identifizierbar ist. Ich danke Erasmus Gerigk für den Hinweis auf diese Beweisidee und die explizite Ausformulierung des Induktionsbeweises.

hinweg, d.h. $J = \Sigma_k J_k$. Wie schon gesagt: $\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)$ ist die Matrix der partiellen Ableitungen der Modellgleichungen $p_1(\boldsymbol{\theta})$, ..., $p_J(\boldsymbol{\theta})$ nach den S Modellparametern θ_1, ..., θ_s, ..., θ_S, evaluiert an der Stelle $\boldsymbol{\theta}^{\mathbf{o}} \in \boldsymbol{\Omega}$. Sie hat die Form:

$$\boldsymbol{J}(\boldsymbol{\theta}^{\mathbf{o}}, f) = \begin{bmatrix} \partial p_1(\boldsymbol{\theta}) / \partial\theta_1 & \ldots & \partial p_1(\boldsymbol{\theta}) / \partial\theta_S \\ \ldots & \partial p_j(\boldsymbol{\theta}) / \partial\theta_s & \ldots \\ \partial p_J(\boldsymbol{\theta}) / \partial\theta_1 & \ldots & \partial p_J(\boldsymbol{\theta}) / \partial\theta_S \end{bmatrix}_{\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathbf{o}}} \qquad (4.1.3\text{-}1)$$

Der Rang $\rho[\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)]$ dieser Matrix ist die Maximalzahl linear unabhängiger Spalten- oder Zeilenvektoren von $\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)$. Obwohl die derzeit verfügbaren Computerprogramme zu MVB-Modellen keine Bestimmung von $\rho[\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)]$ erlauben (vgl. Abschnitt 4.7), könnte diese Lücke leicht geschlossen werden. Wie man aus der Produktregel[30] herleiten kann, hat die an der Stelle $\boldsymbol{\theta}^{\mathbf{o}}$ evaluierte partielle Ableitung der j-ten Modellgleichung nach dem s-ten Modellparameter im Falle binärer MVB-Modelle beispielsweise die Form

$$\left.\frac{\partial p_j(\boldsymbol{\theta})}{\partial\theta_s}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\mathbf{o}}} = \sum_{i=1}^{I_j} c_{j,i} \cdot \left(\begin{array}{l} a_{j,i,s} \cdot \prod_{l=1}^{S} \left[(1-\theta^{\mathrm{o}}{}_l)^{b_{j,i,l}} \cdot \left(\theta^{\mathrm{o}}{}_l\right)^{a^*{}_{j,i,l}} \right] \\ + \; b_{j,i,s} \cdot \prod_{l=1}^{S} \left[(1-\theta^{\mathrm{o}}{}_l)^{b^*{}_{j,i,l}} \cdot \left(\theta^{\mathrm{o}}{}_l\right)^{a_{j,i,l}} \right] \end{array} \right). \qquad (4.1.3\text{-}2)$$

Alle Konstanten sind hierbei wie in Gleichung (3.5-1) definiert, mit der Ausnahme von $a^*_{j,i,l}$ und $b^*_{j,i,l}$. Es gilt:

$$a^*{}_{j,i,l} = \begin{cases} a_{j,i,l}, \text{falls } l \neq s \\ a_{j,i,l} - 1, \text{falls } l = s \end{cases} \text{ und } b^*{}_{j,i,l} = \begin{cases} b_{j,i,l}, \text{falls } l \neq s \\ b_{j,i,l} - 1, \text{falls } l = s \end{cases}. \qquad (4.1.3\text{-}3)$$

Die beiden Indizes k und j aus der Modellgleichung (3.5-1) wurden außerdem durch einen einzigen Index j ersetzt, der über alle $J = \Sigma_k J_k$ Modellgleichungen läuft.

Wenn man jede der $J \cdot S$ Zellen der Jacobischen Matrix anhand von Gleichung (4.1.3-2) bestimmt hat, benötigt man nur noch ein numerisch exaktes Computerprogramm zur Berechnung von $\rho[\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f_{\mathrm{BVB}})]$, beispielsweise die von Bamber und van Santen (1985, S. 459) empfohlene Routine von Nash (1979) oder die Routine „SCDCMP" von Press, Teukolsky, Vetterling und Flannery (1992, S. 59ff). Das Ergebnis informiert uns — wie oben bereits ausgeführt — zunächst nur über die *lokale* Identifizierbarkeit des Modells an der evaluierten Stelle $\boldsymbol{\theta} = \boldsymbol{\theta}^{\mathbf{o}}$ des Parameterraumes: Hat die Jacobische Matrix vollen Rang $\rho[\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f_{\mathrm{BVB}})] = S$, so ist das Mo-

[30] Sind $u(x)$,$v(x)$ und $f(x) := u(x) \cdot v(x)$ differenzierbare reellwertige Funktionen einer reellwertigen Variablen x, so gilt für die Ableitung der Produktfunktion $f(x)$ nach x: $f'(x) = v(x) \cdot u'(x) + u(x) \cdot v'(x)$.

dell zumindest dort lokal identifizierbar. Um lokale und globale Identifizierbarkeit ausschließen zu können, muß man allerdings den *maximalen* Rang gemäß Gleichung (4.1.2-3) bestimmen und zeigen, daß dieser kleiner als S ist. Glücklicherweise ist dies bei analytischen, beliebig oft differenzierbaren Modellgleichungen — wie im hier gegebenen Fall — kein Problem, denn die Jacobische Matrix hat unter diesen Umständen maximalen Rang beinahe überall in $\boldsymbol{\Omega}^*$. Man wählt also einfach einen Parametervektor $\boldsymbol{\theta}^* \in \boldsymbol{\Omega}^*$ zufällig aus, bestimmt für diesen $\rho[\mathbf{J}(\boldsymbol{\theta}^*, f_{\mathrm{BVB}})]$ und nimmt an, daß dieser Rang mit $\rho_{\max}$ identisch ist. „Although it is theoretically possible for this procedure to give an incorrect value for the maximum, the probability of this happening is zero" (Bamber & van Santen, 1985, S. 457; eine genauere Formulierung und einen Beweis des Arguments findet man in Anhang B, a.a.O., S. 470-472).

Das hier für binäre Verarbeitungsbaummodelle illustrierte Vorgehen läßt sich problemlos auch auf allgemeine MVB-Modelle anwenden. Hierbei ist die Ableitung (4.1.3-2) lediglich durch die entsprechende Ableitung der MVB-Modellgleichung zu ersetzen. Analog geht man bei anderen parametrisierten multinomialen Modellen vor.

Der Rang der Jacobischen Matrix ist — wie man sieht — eine ausgesprochen nützliche Informationsquelle, um über die lokale Identifizierbarkeit bzw. die Nichtidentifizierbarkeit von Modellen Entscheidungen zu fällen. Schon aus diesem Grund wäre es wünschenswert, wenigstens $\rho[\mathbf{J}(\hat{\boldsymbol{\theta}}, f_{\mathrm{BVB}})]$ — den Rang der Jacobischen Matrix für die gegebenen Parameterschätzungen — routinemäßig in Computerprogrammen zu BVB-Modellen zu berechnen, wie das in einigen Computerprogrammen zur LCA bereits geschieht (Clogg, 1977; Formann, 1984). Das Ergebnis würde uns im Regelfall Auskunft über die Eindeutigkeit der Parameterschätzungen geben; zusätzlich erleichtert es die Entscheidung darüber, ob gewisse Standardverfahren der Modellprüfung zur Anwendung kommen können (vgl. Abschnitt 4.3).[31] Auch wenn sich herausstellen sollte, daß Identifizierbarkeit nicht gegeben ist, kann der Rang der Jacobischen Matrix sehr hilfreich sein, um alternative Methoden der Modellprüfung zu entwickeln. Ich werde in Abschnitt 4.6 hierauf zurückkommen.

Stellt sich über eine Analyse des Ranges der Jacobischen Matrix heraus, daß ein Modell möglicherweise global identifizierbar ist, und kann der Beweis der Identifizierbarkeit nicht durch Konstruktion der Umkehrabbildung zu f erbracht werden, so bleiben zwei Möglichkeiten. Einmal kann man versuchen, den Beweis „direkt" zu erbringen, indem man zeigt, daß aus der Gleichheit der Kategoriewahrscheinlichkeiten $p_{k,j}(\theta_1, ..., \theta_S) = p_{k,j}(\theta_1', ..., \theta_S')$ für $k = 1, ..., K$, $j = 1, ..., J_k$, $\theta_s \in (0,$

[31] Indirekt erhält man diese Information allerdings auch aus der sogenannten Fisher-Informationsmatrix bzw. aus der Antwort auf die Frage, ob das Inverse der Fisher-Informationsmatrix existiert oder nicht (vgl. dazu Abschnitt 4.2.5).

1) und $\theta'_s \in (0, 1)$ die Gleichheit der erzeugenden Parameterwerte folgt (d.h., $\theta_S = \theta_S'$, für $s = 1, ..., S$). Demonstrationen dieses Vorgehens, das nicht sehr aufwendig sein muß, findet man z.B. bei Rouder und Batchelder (1998, Anhang B) und bei Meiser und Bröder (1999, Anhang A). Gelingt der Identifizierbarkeitsnachweis für ein bestimmtes Modell, so folgt natürlich automatisch die Identifizierbarkeit aller Submodelle dieses Modells. Im Falle hierarchischer Modellfamilien sollte man daher zunächst versuchen, die Identifizierbarkeit saturierter Modelle zu beweisen, weil sich im Erfolgsfalle separate Nachweise für die Submodelle erübrigen.

Gelingt der Identifizierbarkeitsnachweis auch auf diese Weise nicht, so bleibt noch ein pragmatischer Ausweg: Man simuliert sehr viele zufällig ausgewählte Parametervektoren, berechnet durch Einsetzen in die Modellgleichungen die zugehörigen Kategoriewahrscheinlichkeiten aus $\boldsymbol{\Gamma}$ und bestimmt dann dazu irgendeinen BAN-Schätzer der Modellparameter. Weicht auch nur einer der Schätzwerte von dem bei der Erzeugung der Kategoriewahrscheinlichkeiten zugrunde gelegten Parameter ab, so ist das Modell offenbar nicht global identifizierbar. Gelingt dagegen die Reproduktion der erzeugenden Parameterwerte — bis auf eventuelle Rundungsungenauigkeiten — in jedem Fall, so ist die globale Identifizierbarkeit des Modells sehr wahrscheinlich. Rouder und Batchelder (1998, S. 207) nennen dies den „Simulationsnachweis der Identifizierbarkeit" (simulated identifiability). Natürlich ist der Simulationsnachweis kein Beweis; er kann aber dennoch sehr nützlich sein, wenn man die Existenz eindeutiger Parameterschätzungen für beliebige modellkonforme Datensätze plausibel machen will.

4.2 Parameterschätzung

Die im letzten Abschnitt dargestellten Analysen betrafen die Frage, ob bzw. unter welchen Bedingungen die Parameterwerte eindeutig bestimmbar sind, wenn (a) das Modell empirisch gilt und (b) die Kenntnis der Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ vorausgesetzt werden kann. Dies ist nicht die Situation, mit der Anwender konfrontiert sind. Sie kennen $\boldsymbol{\pi}$ in aller Regel nicht und können folglich $\boldsymbol{\theta}$ auch nicht „bestimmen", selbst wenn ein Modell global identifizierbar und gültig sein sollte. Stattdessen gilt es, $\boldsymbol{\theta}$ aus

$$J = M + K = \sum_{k=1}^{K} J_k \qquad (4.2\text{-}1)$$

vorliegenden Stichprobenhäufigkeiten für die verschiedenen Beobachtungskategorien zu schätzen. Diese Stichprobenhäufigkeiten werden zu einem Vektor $\boldsymbol{y}$ zusammengefaßt (vgl. Gleichung (3.2-1)). Unter einem Parameterschätzer oder kurz „Schätzer" (estimator) eines (möglicherweise vektoriellen) Parameters $\boldsymbol{\theta}$ versteht man eine Funktion κ, die jedem möglichen Vektor von Stichprobenhäufigkeiten

$\boldsymbol{Y} = \boldsymbol{y}$ einen Parameterwert (oder Parametervektor) $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\theta}}$ zuordnet. Ebenso wie $\boldsymbol{Y}$ ist auch der Schätzer $\hat{\boldsymbol{\Theta}} = \kappa(\boldsymbol{Y})$ eine Zufallsvariable bzw. ein Zufallsvektor mit einer Verteilung, der sogenannten Stichprobenverteilung von $\hat{\boldsymbol{\Theta}}$, die im Prinzip für gegebene Beobachtungsgesamtzahl N bzw. gegebene Gesamtzahlen $\boldsymbol{N} = (N_1, N_2, ..., N_K)$ aus dem multinomialen Rahmenmodell in Abhängigkeit von $\boldsymbol{\theta}$ herleitbar ist (vgl. Abschnitt 3.1 und 3.2). Der einem konkreten Stichprobenvektor $\boldsymbol{y}$ durch den Schätzer $\hat{\boldsymbol{\Theta}} = \kappa(\boldsymbol{Y})$ zugeordnete Wert bzw. Vektor $\hat{\boldsymbol{\theta}} = \kappa(\boldsymbol{y})$ heißt auch Parameterschätzung oder kurz „Schätzwert" (estimate) von $\boldsymbol{\theta}$.

Offenbar sind aus der Vielzahl denkbarer Funktionen κ nicht alle interessant, sondern nur solche, die in einem noch zu konkretisierenden Sinne „gute" Schätzer des Parameters $\boldsymbol{\theta}$ sind. Die Güte eines Schätzers mißt man an der Abweichung zwischen $\boldsymbol{\theta}$ und $\hat{\boldsymbol{\Theta}}$, quantifiziert durch eine Verlustfunktion $W_{\boldsymbol{\theta}}(\hat{\boldsymbol{\Theta}})$, beispielsweise durch die quadratische Verlustfunktion $W_{\boldsymbol{\theta}}(\hat{\boldsymbol{\Theta}}) = (\hat{\boldsymbol{\Theta}}-\boldsymbol{\theta})^2$. Der Erwartungswert des Verlustes eines Schätzers κ bei gegebenem Parameter $\boldsymbol{\theta}$, $R_{\boldsymbol{\theta}}(\kappa) := E_{\theta}(W_{\boldsymbol{\theta}}(\hat{\boldsymbol{\Theta}} = \kappa(\boldsymbol{Y})))$, ist das sogenannte Risiko des Schätzers κ. Ein Schätzer κ wäre offenbar ideal im Sinne der spezifizierten Verlustfunktion, wenn dessen Risiko für alle möglichen Parameterwerte $\boldsymbol{\theta} \in \boldsymbol{\Omega}$ und alle möglichen Stichprobenumfänge $\boldsymbol{N}$ niemals größer als das Risiko eines denkbaren alternativer Schätzers κ' wäre und für wenigstens einen Parameterwert kleiner. Leider gibt es einen in diesem Sinne generell besten Schätzer im allgemeinen nicht (zusammenfassend: Klauer, 1996a).

Alternativ bietet es sich an, Familien von Schätzern auszuzeichnen, deren Stichprobenverteilungen möglichst viele „positive" Eigenschaften auf sich vereinigen. Ein wichtiges Konzept in diesem Zusammenhang ist das Konzept des „besten asymptotisch normalverteilten Schätzers" (best asymptotically normal estimator, BAN-Schätzer), das von Neyman (1949) eingeführt wurde. BAN-Schätzer haben asymptotisch, d.h. für gegen unendlich strebende Stichprobenumfänge, eine Reihe wünschenswerter Eigenschaften. Sei N der Gesamtstichprobenumfang eines verbundenen multinomialen Modells, d.h. es gilt $\boldsymbol{N} = (N \cdot g_1, ..., N \cdot g_K)$ für positiv-reellwertige Konstanten g_k mit $\Sigma_k\, g_k = 1$. Dann hat ein BAN-Schätzer $\hat{\boldsymbol{\Theta}}$ des Parameters $\boldsymbol{\theta}$ in einem parametrisierten multinomialen Modell folgende Eigenschaften (vgl. z.B. Elandt-Johnson, 1971; Read & Cressie, 1988):

- Er ist *konsistent*, d.h.

$$\lim_{N\to\infty} Pr\left(\left\{\left|\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}\right| < \boldsymbol{\epsilon}\right\}\right) = 1, \text{ für beliebig kleines } \boldsymbol{\epsilon} > \boldsymbol{0}; \tag{4.2-2}$$

- er ist *asymptotisch erwartungstreu*, d.h. für den Erwartungswert $E(\hat{\boldsymbol{\Theta}})$ des Schätzers gilt

$$\lim_{N\to\infty} E(\hat{\boldsymbol{\Theta}}) = \boldsymbol{\theta}\ ; \tag{4.2-3}$$

- er ist *asymptotisch effizient*, d.h. für $N \to \infty$ tendiert die Varianz $Var(\hat{\boldsymbol{\Theta}})$ des Schätzers gegen die untere Grenze der Varianz, die sogenannte Cramer-Rao-Schranke (einführend: Klauer, 1996a);
- er ist *asymptotisch normalverteilt*, d.h. die Statistik

$$Z_{\hat{\boldsymbol{\Theta}}} = \frac{\hat{\boldsymbol{\Theta}}}{\sqrt{\lim_{N \to \infty} Var(\hat{\boldsymbol{\Theta}})}} \qquad (4.2\text{-}4)$$

folgt asymptotisch einer Standardnormalverteilung; diese Eigenschaft kann beispielsweise für die Konstruktion von asymptotischen Konfidenzintervallen genutzt werden kann (s.u.).

Die Vielzahl positiver asymptotischer Eigenschaften von BAN-Schätzern darf nicht darüber hinwegtäuschen, daß sie im realistischen Fall finiter Stichprobenumfänge durchaus auch „schlechte" Schätzer in dem Sinne sein können, daß sie beispielsweise nicht erwartungstreu sind. Es ist intuitiv klar, daß dieses Problem um so gravierender wird, je kleiner N ist.

Dennoch wird man einen Schätzer, der kein BAN-Schätzer ist, nur in besonders begründeten Ausnahmefällen einem BAN-Schätzer vorziehen. Wenn der Stichprobenumfang hinreichend groß ist, so daß die exakte Stichprobenverteilung eines BAN-Schätzers $\hat{\boldsymbol{\Theta}}$ problemlos durch dessen asymptotische Verteilung approximiert werden kann, und wenn gleichzeitig die Berechnung des BAN-Schätzers praktikabel ist, gibt es eigentlich keinen Grund, auf ihn zu verzichten. Ich werde mich deshalb in diesem Kapitel im wesentlichen auf BAN-Schätzer für parametrisierte multinomiale Modelle beschränken und auf andere Schätzer nur kurz eingehen.

4.2.1 Maximum-Likelihood-Schätzer

Der wohl bekannteste BAN-Schätzer ist der Maximum-Likelihood-Schätzer (ML-Schätzer). Tatsächlich hat Neyman das Konzept des BAN-Schätzers u.a. deshalb ausgearbeitet, um diejenigen asymptotischen Eigenschaften von ML-Schätzern zu spezifizieren, mit denen man die Bevorzugung von ML-Schätzern gegenüber anderen Schätzern rational begründen kann: „To the writer´s knowledge, these three important properties of the ML estimates (gemeint sind die BAN-Eigenschaften Konsistenz, asymptotische Normalverteilung und asymptotische Effizienz) form the only rational basis for preferring these estimates to others" (Neyman, 1949, S. 241).[32] ML-Schätzern kommt die BAN-Eigenschaft nicht nur im Falle parametri-

[32]Ein weiteres Motiv Neymans war, innerhalb der Klasse der Schätzer mit gleichen asymptotischen Eigenschaften möglicherweise solche zu finden, die sich einfacher berechnen lassen als ML-Schätzer (Neyman, 1949, S. 241). Wie im folgenden deutlich werden wird, kann dieses Motiv, nach Alternativen zu ML-Schätzern zu suchen, aus heutiger Sicht nicht mehr überzeugen.

sierter multinomialer Modelle zu. Nicht zuletzt deshalb sind sie zu einer Art „Allzweck-Werkzeug" in der angewandten Statistik geworden, auch wenn dies gelegentlich — speziell von seiten der Bayesschen Statistik — auf Kritik gestoßen ist. Wie Rao (1962; 1963) gezeigt hat, besitzen ML-Schätzer in parametrisierten multinomialen Modellen zusätzlich Effizienz zweiter Ordnung, d.h. sie haben nach geeigneter Biaskorrektur auch im Falle endlicher Stichprobenumfänge die geringste Varianz (vgl. Read & Cressie, 1988, Anhang A1, S. 154f). Diese Eigenschaft hebt sie von anderen BAN-Schätzern ab.[33] Berücksichtigt man darüber hinaus, daß ML-Schätzer mit gegenwärtig verfügbaren Computerprogrammen leicht und bequem berechenbar sind, so ist dies Grund genug, die Diskussion der ML-Schätzer an den Anfang zu stellen.

Der ML-Schätzer des Parameters $\boldsymbol{\theta}$ in einem parametrisierten multinomialen Modell zeichnet sich dadurch aus, daß er die Likelihood-Funktion

$$L(\boldsymbol{\theta};\boldsymbol{y}) = \prod_{k=1}^{K} \frac{N_k!}{y_{k,1}!\cdot y_{k,2}!\cdot\ldots\cdot y_{k,J_k}!} p_{k,1}(\boldsymbol{\theta})^{y_{k,1}} \cdot p_{k,2}(\boldsymbol{\theta})^{y_{k,2}} \cdot\ldots\cdot p_{k,J_k}(\boldsymbol{\theta})^{y_{k,J_k}} = \\ = \prod_{k=1}^{K} N_k! \cdot \prod_{j=1}^{J_k} \frac{p_{k,j}(\boldsymbol{\theta})^{y_{k,j}}}{y_{k,j}!} \quad . \tag{4.2.1-1}$$

maximiert. Die Likelihood-Funktion hat dieselbe Form wie die Wahrscheinlichkeitsfunktion (3.2-3), wenn man die Modellgleichungen $p_{k,j}(\boldsymbol{\theta})$ für die Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ einsetzt.[34] Allerdings ist $L(\boldsymbol{\theta}; \boldsymbol{y})$ eine Funktion der Parameter $\boldsymbol{\theta}$ für ein fixiertes Stichprobenresultat $\boldsymbol{Y} = \boldsymbol{y}$, während es bei der Wahrscheinlichkeitsfunktion $Pr(\boldsymbol{y}; \boldsymbol{\theta})$ genau umgekehrt ist. Aus der Definition von ML-Schätzern ergibt sich unmittelbar, daß sie einer Nullstelle der ersten partiellen Ableitungen der Likelihood-Funktion nach den S Modellparametern korrespondieren müssen. Diese Bedingung ist notwendig, aber nicht hinreichend, da die Likelihood-Funktion prinzipiell auch mehrere Extrema (Maxima und Minima) aufweisen kann. Die ML-Schätzer entsprechen dann den Parameterwerten, die zum größten Maximum führen.

Die Regularitätsbedingungen, die für die Existenz eindeutiger ML-Schätzer vorausgesetzt werden müssen, entsprechen denen von BAN-Schätzern allgemein (vgl. Birch, 1964; Read & Cressie, 1988, Anhang A5). Abgesehen von Differenzierbarkeits- und Stetigkeitsbedingungen sind hierbei hauptsächlich zwei Forderungen zu beachten. Zum einen darf der wahre Parameter $\boldsymbol{\theta}^{\mathbf{o}}$ nicht auf dem Rand des Pa-

[33] Allerdings gibt es seit Anfang der 80er Jahre eine intensive Debatte über die Adäquatheit des Konzepts der Effizienz zweiter Ordnung. Nicht alle Statistiker sehen darin ein nützliches Kriterium zum Vergleich von Schätzern (vgl. Read & Cressie, 1988, Anhang 1).

[34] Man nennt die M^* verschiedenen Modellgleichungen $p_{k,j}(\boldsymbol{\theta})$ auch Skalarkomponenten der modelldefinierenden Abbildung f: $\boldsymbol{\Omega} \rightarrow \boldsymbol{\Gamma}$ (vgl. Bamber & van Santen, 1985, S. 454).

rameterraumes liegen. Für MVB-Modelle zieht dies den Ausschluß der Extremwerte null und eins nach sich, so daß $\boldsymbol{\theta}^{\mathbf{o}}$ einer offenen Menge $\boldsymbol{\Omega}^{*} \subseteq (0, 1)^{S}$ entstammen muß. Ferner muß die Jacobische Matrix (vgl. Abschnitt 4.1.1) an der Stelle des wahren Parameters $\boldsymbol{\theta}^{\mathbf{o}}$ vollen Rang $\rho[\mathbf{J}(\boldsymbol{\theta}^{\mathbf{o}}, f)] = S$ besitzen, das Modell muß mithin in einer offenen Umgebung von $\boldsymbol{\theta}^{\mathbf{o}}$ lokal identifizierbar sein. Dies schließt u.a. auch überparametrisierte Modelle mit $S > M$ aus (vgl. Abschnitt 4.1.1). Für praktische Zwecke ist es oft ausreichend, den Rang der Jacobischen Matrix an der Stelle eines Schätzwerts $\hat{\boldsymbol{\theta}}$ zu prüfen, der die Likelihood-Funktion (4.2.1-1) maximiert. Denn wenn $\mathbf{J}(\hat{\boldsymbol{\theta}}, f)$ vollen Rang aufweist, dann muß die Jacobische Matrix auch in einer Umgebung von $\hat{\boldsymbol{\theta}}$ vollen Rang aufweisen, in der $\boldsymbol{\theta}$ — hinreichend große Stichprobenumfänge vorausgesetzt — sehr wahrscheinlich liegen wird (vgl. Bamber & van Santen, 1985, Appendix B).

Polson und Huizinga (1974, S. 13, Definition 10) haben vorgeschlagen, ein Modell relativ zu ein konkreten Datensatz $\boldsymbol{Y} = \boldsymbol{y}$ „empirisch identifizierbar" zu nennen, wenn der ML-Schätzwert für diese Daten eindeutig ist. Zu beachten ist, daß bei kleinen Stichprobenumfängen empirische Identifizierbarkeit nicht automatisch gesichert ist, wenn die o.g. Bedingungen erfüllt sind. Lediglich für $N \rightarrow \infty$ sind die ML-Schätzwerte unter den genannten Voraussetzungen eindeutig (Polson & Huizinga, 1974, S. 13).

Die Parameterwerte, welche die Likelihood-Funktion maximieren, maximieren natürlich auch jede monotone Transformation der Likelihood-Funktion; folglich kann man statt $L(\boldsymbol{\theta}; \boldsymbol{y})$ auch die Log-Likelihood-Funktion $ln(L(\boldsymbol{\theta}; \boldsymbol{y}))$ verwenden, was bei der Herleitung der partiellen Ableitungen oft Vorteile hat. Ein weiteres äquivalentes Kriterium ist die Minimierung der Log-Likelihood-Quotienten-Statistik (z.B. Bishop et al., 1975; Read & Cressie, 1988):

$$G^2(\boldsymbol{\theta};\boldsymbol{y}) = -2 \cdot \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \ln\left(\frac{N_k \cdot p_{k,j}(\boldsymbol{\theta})}{y_{k,j}}\right) = 2 \cdot \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \ln\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta})}\right) \quad (4.2.1\text{-}2)$$

Die Äquivalenz der beiden Schreibweisen von $G^2(\boldsymbol{\theta}; \boldsymbol{y})$ in Gleichung (4.2.1-2) ist offensichtlich, wenn sowohl die empirischen Häufigkeiten $y_{k,j}$ als auch erwarteten Häufigkeiten $N_k \cdot p_{k,j}(\boldsymbol{\theta})$ größer als null sind. Um den bei kleinen Stichproben gelegentlich auftretenden Fall zu berücksichtigen, daß eine Kategoriehäufigkeit oder auch mehrere Kategoriehäufigkeiten den Wert $y_{k,j} = 0$ annehmen, definiert man $\ln(N_k \cdot p_{k,j}(\boldsymbol{\theta}) \,/\, 0) := 0$ bzw. $\ln(0 \,/\, N_k \cdot p_{k,j}(\boldsymbol{\theta})) := 0$.

Wie kann die Berechnung der ML-Schätzwerte praktisch ablaufen? Eine ganz einfache Situation ist gegeben, wenn wir ein vollparametrisiertes Modell verwenden, dessen Parameterzahl S der Anzahl M funktional unabhängiger Kategoriewahrscheinlichkeiten entspricht und das zudem global identifizierbar ist. Wie bereits Bailey (1951) gezeigt hat, erhält man in einem solchen Fall ML-Schätzer für

reellwertige, nicht notwendig auf das Intervall (0, 1) eingeschränkte Parameter θ_1, ..., θ_S, ... θ_S ganz einfach dadurch, daß man zunächst ML-Schätzer $\hat{\pi}_{k,j}$ der Kategoriewahrscheinlichkeiten $\pi_{k,j}$ bestimmt, diese anstelle der $p_{k,j}$ in die Modellgleichungen einsetzt und das resultierende Gleichungssystem nach den Parametern auflöst, d.h. die Umkehrabbildung zu den Modellgleichungen bestimmt. Wie man leicht zeigen kann (z.B. Krengel, 1991, S. 64), entsprechen die ML-Schätzer der unrestringierten Kategoriewahrscheinlichkeiten $\pi_{k,j}$ einfach den relativen Häufigkeiten der Kategorien in der vorliegenden Stichprobe,

$$\hat{\pi}_{k,j} = \frac{y_{k,j}}{N_k},\ k = 1,\ \ldots,\ K,\ j = 1,\ \ldots,\ J_k, \tag{4.2.1-3}$$

so daß die Berechnung der ML-Schätzer für die Parameter wirklich sehr einfach wird, wenn die Umkehrabbildung zu den Modellgleichungen bekannt ist. Ein Beispiel mag dies illustrieren: Das 2HT-Modell aus Abschnitt 3.5.2 ist ein global identifizierbares Modell mit $S = M = 2$ freien Parametern $r_{2\text{ht}}$ und $b_{2\text{ht}}$. Durch Bildung der Umkehrfunktion zu den Gleichungen (3.5.2-3) und (3.5.2-4) erhält man die ML-Schätzer der Parameter nach der Bailey-Methode einfach wie folgt[35]:

$$\hat{r}_{2\text{ht}} = \hat{\pi}_{1,1} - \hat{\pi}_{2,1} = y_{1,1} / N_1 - y_{2,1} / N_2 \tag{4.2.1-4}$$

$$\hat{b}_{2\text{ht}} = \frac{\hat{\pi}_{2,1}}{1 - \hat{r}_{2\text{ht}}} = \frac{y_{2,1} / N_2}{1 - (y_{1,1} / N_2 - y_{2,1} / N_2)} \tag{4.2.1-5}$$

Bei unterparametrisierten Modellen funktioniert diese einfache Methode leider nicht. Der dann zunächst naheliegende Weg ist die Bestimmung der ersten partiellen Ableitungen einer der oben eingeführten Funktionen $L(\boldsymbol{\theta}; \boldsymbol{y})$, $ln(L(\boldsymbol{\theta}; \boldsymbol{y}))$ oder $G^2(\boldsymbol{\theta}; \boldsymbol{y})$ nach den Modellparametern θ_1, ..., θ_S, ... θ_S und die Nullsetzung dieser partiellen Ableitungen für gegebene Daten. Für viele Modellklassen liegen die ersten Ableitungen bereits in analytischer Form vor, so etwa für binäre (Hu & Batchelder, 1994) und nichtbinäre MVB-Modelle (Hu, 1999a). Ergeben sich mehrere Nullstellen, so sucht man einfach das größte Maximum heraus.

Abgesehen davon, daß dieser Standardweg der Bestimmung von ML-Schätzern trotz bereits vorliegender analytischer Resultate aufwendig und häufig auch fehlerträchtig ist, ist er noch aus zwei weiteren Gründen problematisch:

(1) Nicht immer existieren explizite Lösungen der ML-Schätzgleichungen, d.h. nicht immer läßt sich aus den Nullstellen der Ableitungen von $L(\boldsymbol{\theta}; \boldsymbol{y})$, $ln(L(\boldsymbol{\theta}; \boldsymbol{y}))$ oder $G^2(\boldsymbol{\theta}; \boldsymbol{y})$ ein System von Gleichungen herleiten, das die ML-Schätzwerte $\hat{\theta}_1, \hat{\theta}_2$, ..., $\hat{\theta}_S$ explizit als Funktionen der Stichprobenhäufigkei-

[35] Wegen der Einschränkung des Parameterraums auf das Intervall (0, 1) ist in diesem Fall allerdings vorauszusetzen, daß $y_{1,1} / N_1 > y_{2,1} / N_2$ gilt; sonst resultieren unzulässige Parameterschätzungen. Wie in einem solchen Fall zu verfahren ist, wird weiter unten besprochen.

ten $\boldsymbol{y}$ darstellt. Man kann die ML-Schätzwerte in solchen Fällen also nicht einfach durch Einsetzung der $y_{k,j}$ in eine Formel errechnen.

(2) Selbst wenn explizite Lösungen für MVB-Schätzgleichungen existieren — wie beispielsweise im Falle des Enkodierung-Abruf-Modells von Batchelder und Riefer (1986) oder des oben behandelten 2HT-Modells — so beinhalten diese Lösungen zunächst nicht die Einschränkung der Parameter auf das Intervall [0, 1]; die Schätzer können daher prinzipiell auch Werte kleiner als null oder größer als eins annehmen. Wenn dieser Fall eintritt, weist dies tendenziell auf Verletzungen der Modellannahmen hin. Die bloße Tatsache eines Schätzwerts größer als eins oder kleiner als null kann aber noch kein Grund sein, das Modell zu verwerfen, denn auch Stichprobenfehler können ein solches Resultat generieren, wenn die wahren Parameter in der Nähe eines Extremwerts liegen.[36] Es ist in solchen Fällen unbedingt wünschenswert, über ein Verfahren zu verfügen, das die Ungleichheitsrestriktion $0 < \hat{\theta}_s < 1$, $s = 1, ..., S$, für alle Schätzwerte erzwingt. Die Frage, ob ein Modell im gegebenen Anwendungsfall noch gültig ist, wäre dann anhand der Modellanpassungsstatistik für ein solchermaßen restringiertes Modell zu entscheiden, nicht anhand der Frage, ob die Schätzwerte im zulässigen Bereich liegen (vgl. dazu Abschnitt 4.3).

Problem (2) läßt sich prinzipiell dadurch lösen, daß man reellwertige Hilfsparameter $\theta'_1, ..., \theta'_s, ... \theta'_S$ zu den Modellparametern einführt, die mit den interessierenden Parametern z.B. über logistische Funktionen verknüpft sind (vgl. Formann, 1984):

$$\theta_s = \frac{\exp(\theta'_s)}{1+\exp(\theta'_s)},\ s = 1,\ ...,\ S. \qquad (4.2.1\text{-}6)$$

Die Hilfsparameter können beliebige reelle Werte annehmen und werden durch die logistische Transformation auf Parameterwerte im Bereich (0, 1) „gestaucht". Man reformuliert also die Likelihood-Funktion einfach in Termini der Hilfsparameter, indem man jeden Parameter durch einen logistischen Term der Form (4.2.1-6) ersetzt. Dann werden zunächst ML-Schätzer der Hilfsparameter bestimmt, die anschließend durch Einsetzen in Gleichung (4.2.1-6) in Schätzwerte der interessierenden Parameter „rückübersetzt" werden. Die resultierenden Schätzwerte liegen zwangsläufig im reellen Intervall [0, 1].

---

[36] Eine Monte-Carlo-Studie von Rouder und Batchelder (1998) macht dies eindrucksvoll deutlich. Sie generierten wiederholt Stichproben zu einem gültigen MVB-Modell, bei dem einige Parameterwerte in der Nähe des Parameterraumrandes lagen. Hierzu wurden unrestringierte ML-Schätzer der Parameter berechnet. Trotz der Modellgültigkeit lagen bei kleinem N teilweise 66% der Parameterschätzungen außerhalb des [0, 1]-Intervalls! Hätte man jeden dieser Fälle zum Anlaß einer Modellverwerfung genommen, wäre die faktische α-Fehlerwahrscheinlichkeit also um mehr als .60 größer als das nominelle $\alpha = .05$ (sic!).

Problem (1) kann durch eine Vielzahl unterschiedlicher numerischer Verfahren gelöst werden, die Maxima und Minima von Funktionen numerisch bestimmen. Hierzu sind zunächst „möglichst gute" Ausgangswerte (sogenannte Startwerte) für die Parameterschätzung vom Anwender festzulegen; wenn nichts über die wahren Parameterwerte bekannt ist, können das im Prinzip auch zufällig ausgewählte Werte aus dem Parameterraum sein. Ein iteratives Rechenverfahren führt dann über mehr oder minder viele Rechenschritte zu den Endwerten, die — im günstigsten Fall — die Likelihood-Funktion maximieren. Einige numerische Verfahren setzen nicht einmal die Kenntnis der partiellen Ableitungen voraus; sie benötigen als Input neben den Daten lediglich eine zu maximierende oder zu minimierende Funktion. Spezifiziert man die Likelihood-Funktion als zu maximierende Funktion oder die Log-Likelihood-Quotienten-Statistik $G^2(\boldsymbol{\theta}; \boldsymbol{y})$ als zu minimierende Funktion, so erhält man als Ergebnis eine Schätzung des Parametervektors $\boldsymbol{\theta}$, die einem lokalen oder globalen Maximum der Likelihood-Funktion entspricht. Durch mehrere Wiederholungen des Algorithmus für unterschiedliche Startwerte kann plausibel gemacht werden, daß die gefundene Lösung tatsächlich einem globalen Maximum bzw. Minimum entspricht. Gegenfurtner (1992) hat ein recht einfach handhabbares Programm PRAXIS vorgestellt, das genau dies leistet. Es basiert auf einer Modifikation der Gradientenmethode der Funktionenminimierung von Powell (1962) durch Brent (1973). Dieses Programm wurde u.a. von Erdfelder und Buchner (1998b) zur Parameterschätzung für das „Dual-Process-Signal-Detection Model" (DPSDM) von Yonelinas, Regehr und Jacoby (1995) eingesetzt, eine Erweiterung des SDT-Modells auf die Prozeßdissoziationsprozedur (vgl. Kapitel 7.1.3). Dieses Modell fällt zwar in die Klasse parametrisierter multinomialer Modelle, nicht aber in die Klasse der MVB-Modelle, so daß es nicht mit speziell auf MVB-Modelle zugeschnittenen Computerprogrammen wie GPT, MBT oder AppleTree (vgl. Abschnitt 4.7 und den Anhang) analysiert werden kann. Vielleicht noch benutzerfreundlicher als PRAXIS ist die SOLVER-Routine aus dem verbreiteten EXCEL-Tabellenkalkulationsprogramm, das für sehr viele verschiedene Betriebssysteme verfügbar und einfach zu bedienen ist (Dodson, Prinzmetal & Shimamura, 1998). Ähnlich wie PRAXIS arbeitet die SOLVER-Routine aus EXCEL mit einer Gradientenmethode (vgl. Kapitel 4.7) und kann zur ML-Schätzung für beliebige parametrisierte multinomiale Modelle eingesetzt werden.

Andere numerische Verfahren setzen die Kenntnis der ersten Ableitungen voraus, beispielsweise die von Elandt-Johnson (1971, S. 307) vorgeschlagene Methode, die auf einer Taylor-Erweiterung basiert, oder die verschiedenen Varianten des Newton-Raphson-Verfahrens (Ortega & Rheinboldt, 1970; Press et al., 1992), welche in vielen Programmen zur Analyse multinomialverteilter Daten bereits Anwendung finden (z.B. Formann, 1984). Alle genannten Verfahren haben den Nachteil, numerisch recht instabil zu sein, d.h. in Abhängigkeit von den gegebe-

nen Daten und von den verwendeten Startwerten entweder gar nicht zu konvergieren oder aber zu Endwerten, die nicht den ML-Schätzern entsprechen. Ein konkretes Beispiel für die Nichtkonvergenz eines Newton-Raphson-Verfahrens auf die ML-Schätzwerte hat z.B. Weir (1990, S. 63) im Kontext der Schätzung von Allelhäufigkeiten zum AB0-Blutgruppensystem angegeben. Er zeigte zugleich, daß das Problem durch geschickte Wahl von Startwerten in der Nähe der ML-Schätzwerte vermieden werden kann. Leider ist aber nur in wenigen Anwendungsfällen a priori klar, wodurch sich optimale Startwerte auszeichnen.

Glücklicherweise gibt es ein Verfahren, welches (a) bei vielen parametrisierten multinomialen Modellen anwendbar ist — u.a. bei allen MVB-Modellen — und (b) alle diese Probleme nicht aufweist: den Expectation-Maximization (EM) Algorithmus. Der EM-Algorithmus ist ein generelles Verfahren der Parameterschätzung bei Problemen mit unvollständigen oder fehlenden Daten („incomplete data problems") bzw. bei stochastischen Modellen, die man als Modelle für unvollständige Datensätze auffassen kann. Er ist ursprünglich aus konkreten Parameterschätzproblem im Bereich der Genetik hervorgegangen (Ceppellini, Siniscalco & Smith, 1955) und inzwischen auf eine Vielzahl unterschiedlicher Probleme verallgemeinert worden. EM-Verfahren bilden derzeit eines der expansivsten Gebiete der angewandten Statistik (Dempster, Laird & Rubin, 1977; Goodman, 1974; 1979; Little & Rubin, 1987; McLachlan & Krishnan, 1997; Mooijaart & van der Heijden, 1992; Rost, 1988; Rubin, 1991; Weir, 1990; Wu, 1983). Für den Bereich von MVB-Modellen haben vor allem Hu und Batchelder (1994) und Hu (1999a) gezeigt, wie sich der EM-Algorithmus einfach und gewinnbringend einsetzen läßt.

Der Grundgedanke des EM-Verfahrens ist — wie schon erwähnt — die Konzeptualisierung des Parameterschätzproblems als Problem fehlender Daten. Die fehlenden Daten sind bei der Analyse von MVB-Modellen die Asthäufigkeiten. Wären diese in jedem Anwendungsfall gegeben — hätte man es also ausschließlich mit „beobachtbaren Ästen" zu tun —, dann ließen sich die ML-Schätzer ganz einfach aus Gleichung (4.1.2-2) errechnen: Setzt man statt der Astwahrscheinlichkeiten deren relative Häufigkeiten in einer Stichprobe ein, so erhält man ML-Schätzer der Modellparameter (Hu & Batchelder, 1994, S. 26, Beobachtung 3).

Nun sind aber die Häufigkeiten einzelner oder aller Äste im Regelfall unbekannt. Bekannt sind nur bestimmte Summen der Asthäufigkeiten, nämlich die Kategoriehäufigkeiten in einer Stichprobe. Der EM-Algorithmus besteht aus zwei iterativ wiederholten Schritten, dem Expectation-Schritt (E-Schritt) und dem Maximization-Schritt (M-Schritt). Im E-Schritt werden die erwarteten Häufigkeiten des vollständigen Datensatzes — der einzelnen Äste also — aus vorläufigen Schätzungen der Modellparameter und gegebenen Kategoriehäufigkeiten errechnet. Hierbei müssen die erwarteten Häufigkeiten der Äste, die in eine bestimmte Kategorie $C_{k,j}$ einmünden, die Randbedingung erfüllen, daß sie sich zur Kategoriehäu-

figkeit $y_{k,j}$ aufaddieren. Mit anderen Worten: Im E-Schritt wird die „bestmögliche" Ergänzung der vorliegenden Daten zu einem vollständigen Datensatz vorgenommen, und zwar „bestmöglich" im Sinne der vorläufigen Parameterschätzungen. Im M-Schritt wird dann der Datensatz erwarteter Häufigkeiten, der aus dem E-Schritt resultiert, wie ein vollständiger empirischer Datensatz behandelt, d.h. es werden ML-Schätzer der Modellparameter durch Einsetzen der erwarteten Häufigkeiten in Gleichung (4.1.2-2) errechnet. Konkret sieht der EM-Algorithmus im Falle binärer MVB-Modelle wie folgt aus (vgl. Hu & Batchelder, 1994, S. 27):

0) Setze den Iterationszähler auf $l = 0$.

1) Setze die Schätzer der Parameter auf beliebige Startwerte $\hat{\theta}_1^{(l)}, \hat{\theta}_2^{(l)}, \ldots, \hat{\theta}_S^{(l)}$, die größer als null und kleiner als eins sind.

2) *E-Schritt*: Berechne die erwarteten Häufigkeiten $E(\#(\mathrm{A}_{k,j,i}))^{(l)}$ der Äste $\mathrm{A}_{k,j,i}$ aus den vorliegenden Schätzungen $\hat{\theta}_1^{(l)}, \hat{\theta}_2^{(l)}, \ldots, \hat{\theta}_S^{(l)}$ unter der Randbedingung, daß $\Sigma_i\, E(\#(\mathrm{A}_{k,j,i}))^{(l)} = y_{k,j}$ für alle Kategorien $\mathrm{C}_{k,j}$ gilt:

$$E(\#(\mathrm{A}_{k,j,i}))^{(l)} = y_{k,j} \cdot \frac{c_{k,j,i} \cdot \prod_{s=1}^{S}\left[(1-\theta_s^{(l)})^{b_{k,j,i,s}} \cdot \left(\theta_s^{(l)}\right)^{a_{k,j,i,s}}\right]}{\sum_{i=1}^{I_{k,j}} c_{k,j,i} \cdot \prod_{s=1}^{S}\left[(1-\theta_s^{(l)})^{b_{k,j,i,s}} \cdot \left(\theta_s^{(l)}\right)^{a_{k,j,i,s}}\right]}. \qquad (4.2.1\text{-}7)$$

Der Zähler im Quotienten der Gleichung (4.2.1-7) entspricht der geschätzten Wahrscheinlichkeit des Astes $\mathrm{A}_{k,j,i}$, der Nenner der geschätzten Wahrscheinlichkeit der Kategorie $\mathrm{C}_{k,j}$:, jeweils unter Zugrundelegung der vorläufigen Schätzwerte. Der Quotient ist also der erwartete *Anteil* der Kategoriehäufigkeit $y_{k,j}$, der in den Ast $\mathrm{A}_{k,j,i}$ fällt.

3) Inkrementiere den Iterationszähler um 1: Setze $l = l + 1$.

4) *M-Schritt*: Berechne die ML-Schätzer der Parameter für den vollständigen Datensatz erwarteter Kategoriehäufigkeiten, wie er in Schritt 2) errechnet wurde. Aus Beobachtung 3 von Hu und Batchelder (1994, S. 26) folgt, daß dies die Schätzwerte

$$\theta_s^{(l)} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{i=1}^{I_{k,j}} a_{k,j,i,s} \cdot E(\#(\mathrm{A}_{k,j,i}))^{(l-1)}}{\sum_{k=1}^{K}\sum_{j=1}^{J_k}\sum_{i=1}^{I_{k,j}} (a_{k,j,i,s} + b_{k,j,i,s}) \cdot E(\#(A_{k,j,i}))^{(l-1)}} \qquad (4.2.1\text{-}8)$$

sind.

5) Berechne für ein festgelegtes (z.B. euklidisches) Distanzmaß die Distanz zwischen den Schätzwerten $\hat{\theta}_1^{(l)}$ und den vorherigen Schätzwerten $\hat{\theta}_1^{(l-1)}$. Ist sie geringer als das Konvergenzkriterium $\varepsilon > 0$ (z.B. $\varepsilon = 10^{-8}$), so beende die Pro-

zedur und akzeptiere die Schätzwerte $\hat{\theta}_1^{(l)}, \hat{\theta}_2^{(l)}, ..., \hat{\theta}_S^{(l)}$ als endgültige Werte; andernfalls gehe zurück zu Schritt 2.

Dempster et al. (1977) haben bewiesen, daß dieser Algorithmus unter sehr allgemeinen Randbedingungen monoton auf einen stationären Punkt (Maximum oder Sattelpunkt) der Likelihood-Funktion hin konvergiert, d.h. es gilt $L(\boldsymbol{\theta}^{(l)}; \boldsymbol{y}) \geq L(\boldsymbol{\theta}^{(l-1)}; \boldsymbol{y})$ für $l = 1, ..., L$, $L \in \mathbb{N}$. Hu und Batchelder (1994, S. 29, Beobachtung 4) konnten zeigen, daß dieses Resultat im Fall der Anwendung auf MVB-Modelle sogar noch verschärft werden kann: Hier ist monotone Konvergenz auf ein (mindestens) lokales Maximum der Likelihood-Funktion garantiert, Sattelpunkte sind ausgeschlossen. Allerdings ist die Möglichkeit mehrerer lokaler Maxima der Likelihood-Funktion prinzipiell gegeben. Man sollte daher den EM-Algorithmus von unterschiedlichen Zufallsstartwerten aus mehrmals starten und die resultierende Schätzwerte erst dann als ML-Schätzungen akzeptieren, wenn sie mehrmals repliziert werden konnten. Ist das der Fall, so ist das natürlich noch keine Garantie, daß die akzeptierte Lösung die ML-Lösung ist; aber man kann sich durch weitere Replikationen gegebenenfalls noch mehr Sicherheit verschaffen. Setzt man die Wahrscheinlichkeit der Konvergenz zum globalen Maximum bei Zufallsstartwerten für die Parameterschätzungen mit x an, so errechnet sich die Wahrscheinlichkeit, bei n unabhängigen Wiederholungen des EM-Algorithmus wenigstens einmal das globale Maximum zu erwischen, nach der geometrischen Verteilung als $1-(1-x)^n$. Ist z.B. $x = 0.5$, so garantieren $n = 5$ Wiederholung bereits, daß mit Wahrscheinlichkeit 0.969 das globale Maximum gefunden wird.

Die praktischen Erfahrungen mit MVB-Modellen zeigen jedoch, daß $x = 0.5$ deutlich zu skeptisch angesetzt ist. Multiple Maxima scheint es bei MVB-Modellen sehr selten zu geben. Mathematisch ist das auch durchaus plausibel: Die Log-Likelihood-Funktion von MVB-Modellen ist ein logarithmiertes Polynom in den Modellparametern, und zwar typischerweise ein logarithmiertes Polynom niedrigen (d.h. ersten oder zweiten) Grades. Es ist unwahrscheinlich, daß eine solche Funktion für den Argumentbereich $[0, 1]^S$ mehrere Maxima aufweist (vgl. Hu & Batchelder, 1994, S. 29). Lediglich wenn derselbe Modellparameter in einem „Ast" des zugehörigen Ereignisbaummodells sehr oft auftaucht, resultiert ein Polynom höheren Grades, so daß die Möglichkeit lokaler Maxima ernsthaft in Betracht zu ziehen ist. Dies ist beispielsweise bei Binomialmischverteilungsmodellen der Fall, die ebenfalls zur Familie der MVB-Modelle gehören (vgl. dazu Abschnitt 6.5).

Im übrigen läßt sich das Problem möglicher multipler Maxima auch durch keines der anderen numerischen Verfahren vermeiden, die oben erwähnt wurden; im Gegenteil: hier ergeben sich zusätzliche Probleme durch fehlende oder nicht monotone Konvergenz bzw. durch Konvergenz auf Endwerte, die keinem lokalen oder globalen Maximum entsprechen. Die einzige Möglichkeit, das Problem multipler Maxima mit absoluter Sicherheit auszuschließen, ist die analyti-

sche Bestimmung aller Nullstellen der ersten partiellen Ableitungen der Likelihood-Funktion. Dies ist ein im allgemeinen sehr aufwendiger Weg.

Wenn bei mehreren Wiederholungen des EM-Algorithmus für den gleichen Datensatz identische Ergebnisse resultieren, kann man mit hoher Sicherheit annehmen, daß (a) ein globales Maximum gefunden wurde und daß (b) das Modell „empirisch identifizierbar", d.h. an der Stelle der Lösung lokal identifizierbar ist. Vorliegende Computerprogramme zur Analyse von MVB-Modellen benutzen aus diesem Grunde Wiederholungen des EM-Algorithmus für einen Datensatz auch als grobe Methode der Identifizierbarkeitsprüfung (identifiability check).

Falls bei mehreren Wiederholungen des EM-Algorithmus unterschiedliche Ergebnisse resultieren, so können folgende Fälle unterschieden werden:

1) Der Wert der Likelihood-Funktion ist trotz unterschiedlicher Schätzwerte immer gleich. Dies läßt mit hoher Wahrscheinlichkeit ein Identifizierbarkeitsproblem vermuten. Theoretisch möglich sind aber auch mehrere gleich hohe lokale Maxima, wobei das Modell an jeder dieser Stellen lokal identifizierbar ist. Sicherheit läßt sich nur erzielen, wenn man den Rang der Jacobischen Matrix zu den verschiedenen Lösungen bestimmt.
2) Der Wert der Likelihood-Funktion unterscheidet sich zwischen verschiedenen Lösungen. Dies ist ein klares Indiz für mehrere lokale Maxima der Likelihood-Funktion. Die ML-Lösung korrespondiert dann dem höchsten Maximum. Man beachte, daß dieser Fall nicht mit einem Identifizierbarkeitsproblem einhergehen muß. Ein Modell kann selbstverständlich lokal oder sogar global identifizierbar sein und trotzdem mehrere lokale Maxima der Likelihood-Funktion aufweisen. Auch hierbei kann der Rang der Jacobischen Matrix an der Stelle der Lösung zur Klärung beitragen.

Es gibt also eine Reihe von Gründen, den Rang der Jacobischen Matrix zu jeder Lösung des EM-Verfahrens zu berechnen; zusammen mit dem Wert der Likelihoodfunktion schafft das Ergebnis Klarheit darüber, wie die möglichen Lösungen zu interpretieren sind.

In bezug auf die Anzahl erforderlicher Iterationsschritte gehört das EM-Verfahren im allgemeinen nicht zu den schnellsten numerischen Verfahren. Das ist aber aus zwei Gründen kein Problem: Erstens kann man einen Akzelerationsfaktor (speed adjuster) in den EM-Algorithmus einbauen, der die Anzahl erforderlicher Iterationsschritte reduziert (vgl. Hu & Batchelder, 1994, S. 37f.; McLachlan & Krishnan, 1997, S. 141-165). Aber auch wenn man davon keinen Gebrauch macht, ist der faktische Zeitbedarf bis zur Konvergenz auch bei einer Vielzahl erforderlicher Iterationsschritte äußerst gering, wenn moderne Computer genutzt werden. Innerhalb eines Iterationszyklus´ können die Statistiken (4.2.1-7) und (4.2.1-8) extrem schnell berechnet werden, so daß die Anzahl der Iterationen nicht gravierend zu Buche schlägt. Auch für parametrisierte multinomiale Modelle, die keine

MVB-Modelle sind, existieren im allgemeinen leicht anwendbare EM-Varianten, und zwar selbst dann, wenn innerhalb des M-Schritts die Berechnung der ML-Schätzer nicht mittels expliziter Gleichungen erfolgen kann. Sogenannte generalisierte EM-Verfahren verwenden iterative Verfahren (z.B. Newton-Raphson) zur Likelihood-Maximierung innerhalb des M-Schritts (vgl. McLachlan & Krishnan, 1997, S. 28-32). Im Prinzip kann auf eine Maximierung der Likelihood pro M-Schritt sogar verzichtet werden. Für die Konvergenz eines generalisierten EM-Verfahrens auf ein lokales oder globales Maximum der Likelihood-Funktion ist es ausreichend, wenn die Likelihood von einem M-Schritt zum nächsten nicht abfällt (Dempster et al., 1977; McLachlan & Krishnan, 1997). Dies belegt die Robustheit des Verfahrens besonders deutlich.

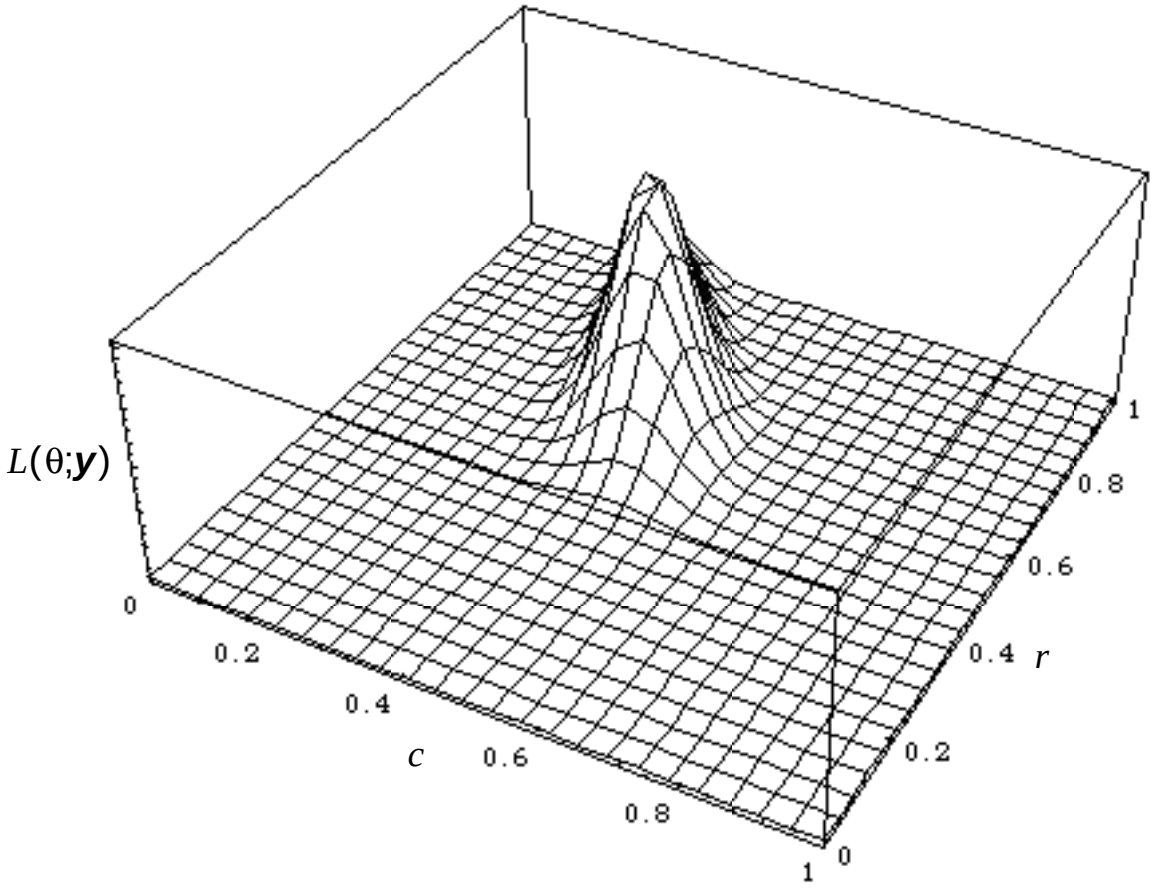

Abbildung 4.2:
Graph einer exemplarische Likelihood-Funktion zum vereinfachten EA-Modell von Batchelder und Riefer (1980, 1986).

Abschließend soll das EM-Verfahren für MVB-Modelle an einem Beispiel illustriert werden, das einem Vortrag von Erdfelder, Brandt und Bröder (1995) entnommen ist[37]. Es basiert auf dem Enkodierung-Abruf-(EA)-Modell von Batchelder

[37]Ich danke Martin Brandt für die numerische und graphische Aufbereitung des Beispiels.

und Riefer (1980, 1986; vgl. Abschnitt 3.5.1). Das Modell wurde der Einfachheit halber auf nur eine Versuchsbedingung (Itempaare) und einen zweidimensionalen Parameterraum reduziert, indem der Parameter u = 0.37 (Abrufwahrscheinlichkeit nicht geclusterter Items) als bereits bekannt unterstellt wurde. Das Parameterschätzproblem reduziert sich demzufolge auf die Bestimmung der ML-Schätzer für den Cluster-Speicherparameter c und den Cluster-Abrufparameter r.

Die Abbildung 4.2 veranschaulicht eine Likelihood-Funktion zu dem vereinfachten EA-Modell. Die dargestellte Likelihood-Funktion ergibt sich, wenn man die vier Kategoriehäufigkeiten für Itempaare wie folgt festlegt: $y_{1,1}$ = 27 (Items nacheinander reproduziert), $y_{1,2}$ = 12 (Items nicht nacheinander reproduziert), $y_{1,3}$ = 24 (nur ein Item reproduziert) und $y_{1,4}$ = 37 (kein Item reproduziert). Offensichtlich gibt es nur ein Maximum, und zwar an der durch $\hat{c}$ = 0.4031 und $\hat{r}$ = 0.6698 definierten Stelle. Folglich muß der EM-Algorithmus bei mehreren Wiederholung immer auf dieses Maximum hin konvergieren. Abbildung 4.3 veranschaulicht für fünf verschiedene Startwertpaare $\hat{c}^{(0)}$ und $\hat{r}^{(0)}$, welche Pfade im Parameterraum diesem Konvergenzprozeß entsprechen.

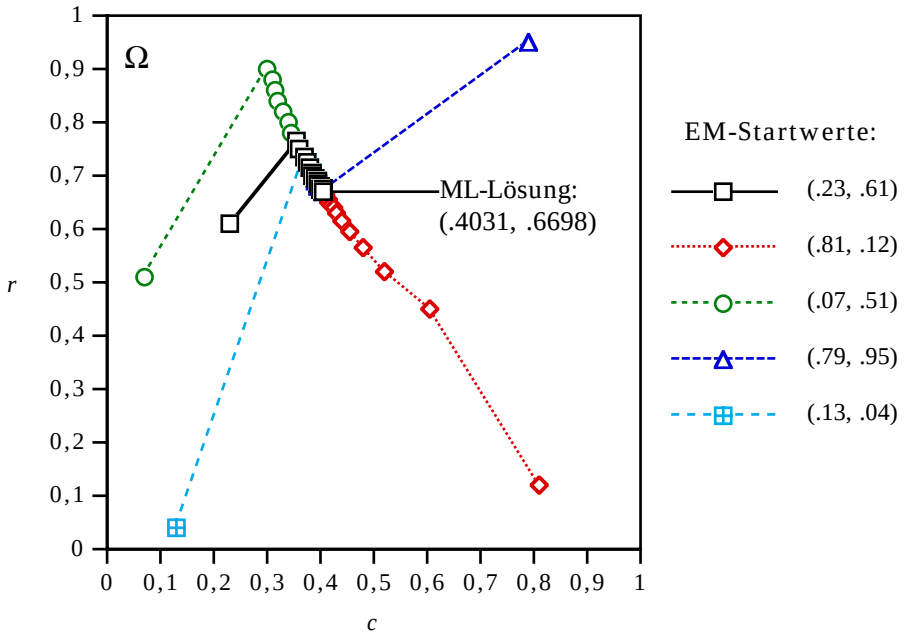

Abbildung 4.3:
Konvergenzprozesse des EM-Algorithmus für fünf verschiedene Startwertpaare, wenn die Likelihoodfunktion aus Abbildung 4.2 zugrunde liegt.

Als Fazit bleibt festzuhalten, daß der EM-Algorithmus zumindest bei MVB-Modellen die Methode der Wahl ist, wenn man ML-Schätzer benötigt. Die Implementation für nichtbinäre MVB-Modelle ergibt sich aus einer kleinen Modifikation des M-Schritts, die Hu (1999a) dargestellt hat. Auch für parametrisierte multinomiale Modelle, die keine MVB-Modelle sind, sollte man zunächst versuchen,

ein EM-Verfahren zu konstruieren. Ist dies zu aufwendig oder nicht möglich, so sind numerische Verfahren zu empfehlen, die keine partiellen Ableitungen der Likelihood-Funktion benötigen. In meiner praktischen Arbeit hat sich die Minimierung der $G^2(\boldsymbol{\theta}; \boldsymbol{y})$-Statistik mit Hilfe des Programms „PRAXIS" von Gegenfurtner (1992) als einfach implementierbar und relativ robust erwiesen. Bei Verwendung dieses oder ähnlicher Verfahren ist es aber noch wichtiger als im Falle des EM-Algorithmus, die letztlich akzeptierte ML-Lösung durch multiple Wiederholungen des Rechenverfahrens abzusichern.

4.2.2 Minimum-Potenz-Divergenz-Schätzer

Der Maximum-Likelihood-Schätzer ist sicher der bekannteste, aber keineswegs der einzige BAN-Schätzer für parametrisierte multinomiale Modelle. Read und Cressie (1988, Anhang A5) haben gezeigt, daß jeder Minimum-Potenz-Divergenz-Schätzer (minimum power-divergence estimate), d.h. jeder Schätzer $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda})}$, dessen Schätzwerte $\hat{\boldsymbol{\theta}}^{(\boldsymbol{\lambda})}$ eine sogenannte Potenz-Divergenz-Statistik (PD$^\lambda$-Statistik)

$$\mathrm{PD}^{\lambda}(\boldsymbol{\theta};\boldsymbol{y}) = \frac{2}{\lambda(\lambda+1)} \cdot \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j}\left[\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta})}\right)^{\lambda} - 1\right], \text{ für } \lambda \in \mathbb{R} \setminus \{-1, 0\} \quad (4.2.2\text{-}1)$$

minimieren, ein BAN-Schätzer ist, sofern die relativ milden Regularitätsbedingungen von Birch (1964) erfüllt sind (vgl. dazu Abschnitt 4.2.1). Eine PD$^\lambda$-Statistik ist ein Maß für die Distanz (oder Divergenz) von beobachteten Kategoriehäufigkeiten $y_{k,j}$ und erwarteten Kategoriehäufigkeiten $N_k \cdot p_{k,j}(\boldsymbol{\theta})$. Die Bezeichnung „Potenz-Divergenz-Statistik" leitet sich daraus ab, daß die Modell-Daten-Divergenz über eine gewichtete Summe von λ-Potenzen der Quotienten beobachteter und erwarteter Kategoriehäufigkeiten quantifiziert wird. Man wählt den Schätzwert $\hat{\boldsymbol{\theta}}^{(\boldsymbol{\lambda})}$ für $\boldsymbol{\theta}$ so, daß diese Divergenz möglichst gering, d.h. idealerweise null wird. Jeder λ-Wert der Statistik ($\lambda \in \mathbb{R}(\{-1, 0\})$ definiert ein anderes Distanzmaß, aber unabhängig vom gewählten Distanzmaß haben alle Minimum-PD$^\lambda$-Schätzer die gleiche asymptotische Verteilung, nämlich genau jene, die auch für ML-Schätzer gilt. Das Konzept des Minimum-Potenz-Divergenz-Schätzers ist also ein sehr allgemeines Konzept. Es ist sogar noch allgemeiner als in Gleichung (4.2.2-1) dargestellt. Dort wurden die Statistiken PD$^{\lambda=0}$ und PD$^{\lambda=-1}$ ausgeklammert, weil für diese Werte der erste Faktor auf der rechten Seite des Gleichheitszeichens undefiniert ist. Man kann die „fehlenden Statistiken" jedoch wie folgt als Grenzwerte einführen:

$$\begin{aligned} \mathrm{PD}^{\lambda=0}(\boldsymbol{\theta};\boldsymbol{y}) &:= \lim_{\lambda\to 0} \mathrm{PD}^{\lambda}(\boldsymbol{\theta};\boldsymbol{y}) \\ \mathrm{PD}^{\lambda=-1}(\boldsymbol{\theta};\boldsymbol{y}) &:= \lim_{\lambda\to -1} \mathrm{PD}^{\lambda}(\boldsymbol{\theta};\boldsymbol{y}). \end{aligned} \quad (4.2.2\text{-}2)$$

Auch diese beiden Statistiken zählen zur PD$^{\lambda}$-Familie; die entsprechenden Schätzer haben die gleichen asymptotischen Eigenschaften wie alle anderen Minimum-PD$^{\lambda}$-Schätzer.

Praktisch alle bislang für Häufigkeitsdaten oder Kontingenztafeln vorgeschlagenen Prüfstatistiken erweisen sich als Spezialfälle der PD$^{\lambda}$-Statistik. So läßt sich beispielsweise zeigen, daß PD$^{\lambda=0}(\boldsymbol{\theta}, \boldsymbol{y})$ identisch mit $G^2(\boldsymbol{\theta}, \boldsymbol{y})$ ist (Read & Cressie, 1988, S. 15). Die Log-Likelihood-Quotienten-Statistik (4.2.1-2) gehört also ebenfalls zur PD$^{\lambda}$-Familie, so daß alle ML-Schätzer zugleich Minimum-PD$^{\lambda=0}$-Schätzer sind. Weitere bekannte Prüfstatistiken sind z.B. die Pearson-χ^2-Statistik $X^2(\boldsymbol{\theta}, \boldsymbol{y})$, die — wie die Äquivalenzumformung (4.2.2-3) zeigt — identisch mit PD$^{\lambda=1}(\boldsymbol{\theta}, \boldsymbol{y})$ ist,

$$
\begin{aligned}
X^2(\boldsymbol{\theta};\boldsymbol{y}) &:= \sum_{k=1}^{K}\sum_{j=1}^{J_k} \frac{\left(y_{k,j} - N_k \cdot p_{k,j}(\boldsymbol{\theta})\right)^2}{N_k \cdot p_{k,j}(\boldsymbol{\theta})} = \\
&= \sum_{k=1}^{K}\sum_{j=1}^{J_k} \frac{\left(y_{k,j}\right)^2}{N_k \cdot p_{k,j}(\boldsymbol{\theta})} - 2 \cdot \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} + \sum_{k=1}^{K}\sum_{j=1}^{J_k} N_k \cdot p_{k,j}(\boldsymbol{\theta}) = \\
&= \sum_{k=1}^{K}\sum_{j=1}^{J_k} \frac{\left(y_{k,j}\right)^2}{N_k \cdot p_{k,j}(\boldsymbol{\theta})} - \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} = \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \left(\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta})}\right) - 1\right) = \\
&= \mathrm{PD}^{\lambda=1}(\boldsymbol{\theta};\boldsymbol{y}) \,,
\end{aligned}
\tag{4.2.2-3}
$$

oder die Neyman-modifizierte χ^2-Statistik $NM^2(\boldsymbol{\theta}, \boldsymbol{y})$ (Neyman, 1949), welche PD$^{\lambda=-2}(\boldsymbol{\theta}, \boldsymbol{y})$ entspricht.[38] Im Gegensatz zu $X^2(\boldsymbol{\theta}, \boldsymbol{y})$ wird bei dieser Statistik im Nenner durch die beobachtete statt durch die erwartete Kategoriehäufigkeit dividiert. Die Äquivalenz von $NM^2(\boldsymbol{\theta}, \boldsymbol{y})$ mit PD$^{\lambda=-2}(\boldsymbol{\theta}, \boldsymbol{y})$ zeigt die folgende Umformung:

$$
\begin{aligned}
NM^2(\boldsymbol{\theta};\boldsymbol{y}) &= \sum_{k=1}^{K}\sum_{j=1}^{J_k} \frac{\left(y_{k,j} - N_k \cdot p_{k,j}(\boldsymbol{\theta})\right)^2}{y_{k,j}} \quad = \\
&= \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \; - 2 \cdot \sum_{k=1}^{K}\sum_{j=1}^{J_k} N_k \cdot p_{k,j}(\boldsymbol{\theta}) \; + \sum_{k=1}^{K}\sum_{j=1}^{J_k} \frac{\left(N_k \cdot p_{k,j}(\boldsymbol{\theta})\right)^2}{y_{k,j}} \quad = \\
&= \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \left(\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta})}\right)^{-2} - 1\right) = \mathrm{PD}^{\lambda=-2}(\boldsymbol{\theta};\boldsymbol{y}) \,.
\end{aligned}
\tag{4.2.2-4}
$$

Minimum-χ^2-Schätzer — im Sinne der Pearson-Statistik oder der Neyman-modifizierten Statistik — sind also ebenfalls spezielle Minimum-PD$^{\lambda}$-Schätzer. Weitere

[38] Bei der Herleitung der Beziehungen zwischen den Statistiken ist zu beachten, daß die Summe der beobachteten Häufigkeiten $y_{k,j}$ über alle Kategorien per definitionem der Summe der erwarteten Häufigkeiten $N_k \cdot p_{k,j}(\boldsymbol{\theta})$ über alle Kategorien gleichen muß.

bekannte Spezialfälle der PD^{λ}-Statistik erwähnen und diskutieren Read und Cressie (1988, S. 11, 30 & 135 f.) sowie Garcia-Pérez (1994, S. 254).

Obwohl nur im Falle $\lambda = 0$ — bei Wahl der Log-Likelihood-Quotienten-Statistik also — ein Schätzer resultiert, der Effizienz zweiter Ordnung im Sinne von Rao (1962) aufweist, kann es statistische Gründe geben, sich anderen Schätzern aus der PD^{λ}-Familie zuzuwenden. Das Konzept der Effizienz zweiter Ordnung ist umstritten, weil die Varianz eines Schätzers $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda})}$ von Interesse ist und weniger die Varianz eines biaskorrigierten Schätzers, der möglicherweise nicht einmal berechnet werden kann (Read & Cressie, 1988, S. 155). Ein plausibleres Konzept zum Vergleich von BAN-Schätzern ist das (relative) Hodges-Lehmann (HL) Defizit. Das HL-Defizit eines Schätzers $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}$, der PD^{λ_1} minimiert, relativ zu einem Schätzer $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}$, der PD^{λ_2} minimiert, sei mit $D_{\mathrm{HL}}[\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}, \hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}]$ abgekürzt. Sind nun N und $M(N)$ zwei Stichprobenumfänge, die bei den Schätzern $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}$ und $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}$ zum gleichen Standardschätzfehler führen, so kann man zeigen, daß das HL-Defizit $D_{\mathrm{HL}}[\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}, \hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}]$ gerade der asymptotischen Differenz $N - M(N)$ entspricht (Read & Cressie, 1988, S. 157):

$$D_{\mathrm{HL}}[\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}, \hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}] = \lim_{N \to \infty} [N - M(N)] \qquad (4.2.2\text{-}5)$$

Je größer also $D_{\mathrm{HL}}[\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}, \hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}]$ wird, desto schlechter ist der Schätzer $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}$ relativ zu $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}$, und zwar in dem Sinne, daß bei $\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}$ ein um den Betrag $D_{\mathrm{HL}}[\hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_1})}, \hat{\boldsymbol{\Theta}}^{(\boldsymbol{\lambda_2})}]$ größerer Stichprobenumfang erforderlich ist, um den größeren Standardschätzfehler, der bei gleichem Stichprobenumfang gegeben wäre, zu kompensieren. (Man beachte, daß ein von null verschiedenes HL-Defizit durchaus damit vereinbar ist, daß beide Schätzer die BAN-Eigenschaft besitzen. Die BAN-Eigenschaften impliziert nämlich lediglich, daß der asymptotische *Quotient* $M(N) / N$ für $N \to \infty$ gegen eins strebt.)

Interessanterweise gibt es keinen generell besten BAN-Schätzer im Sinne der HL-Defizit-Minimierung. Der optimale Schätzer hängt vielmehr vom unbekannten Parameter $\boldsymbol{\theta}$ ab. Für einen breiten Bereich von Parameterwerten liegen die optimalen λ-Werte typischerweise im Intervall [0, 1], so daß man unter dem Aspekt der HL-Defizit-Reduzierung vorwiegend Minimum-PD^{λ}-Schätzer in diesem Intervall in Betracht ziehen sollte (Read & Cressie, 1988, S. 157-159). Read und Cressie (1988) haben eine besondere Präferenz für den Schätzer, der sich aus dem Wert $\lambda = 2/3 = 0.66$ ergibt. Es handelt sich hierbei also gewissermaßen um einen Kompromiß zwischen dem ML-Schätzer und dem Minimum-Pearson-χ^2-Schätzer. Ich werde in Abschnitt 4.2.4 auf die Frage zurückkommen, ob diese Präferenz durch Simulationsstudien gestützt wird.

Wie können Minimum-PD^{λ}-Schätzer berechnet werden? Im Prinzip stehen alle Wege offen, die im letzten Abschnitt bereits angesprochen wurden. Numeri-

sche Verfahren können zur Minimierung eines beliebigen $PD^{\lambda}(\boldsymbol{\theta}, \boldsymbol{y})$ genauso eingesetzt werden wie zur Minimierung von $G^2(\boldsymbol{\theta}, \boldsymbol{y})$. Verwendet man die im PRAXIS-Programm (Gegenfurtner, 1992) oder EXCELs SOLVER-Funktion (Dodson et al., 1998) implementierten Gradientenverfahren, so ist als zu minimierende Zielfunktion einfach das gewünschte $PD^{\lambda}(\boldsymbol{\theta}, \boldsymbol{y})$ einzusetzen. Im übrigen haben Hu und Batchelder (1994) eine Erweiterung des EM-Algorithmus zu einem EM^{λ}-Algorithmus vorgestellt, der für binäre MVB-Modelle PD^{λ}-Statistiken mit $\lambda \neq 0$ minimieren kann. Der E-Schritt bleibt derselbe wie beim EM-Algorithmus; lediglich im M-Schritt wird statt Gleichung (4.2.1-8) die Gleichung

$$\theta_s^{(l)} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J_k}\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta}^{(l-1)})}\right)^{\lambda} \cdot \sum_{i=1}^{I_{k,j}} a_{k,j,i,s} \cdot E(\#(\mathrm{A}_{k,j,i}))^{(l-1)}}{\sum_{k=1}^{K}\sum_{j=1}^{J_k}\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta}^{(l-1)})}\right)^{\lambda} \cdot \sum_{i=1}^{I_{k,j}} (a_{k,j,i,s} + b_{k,j,i,s}) \cdot E(\#(\mathrm{A}_{k,j,i}))^{(l-1)}} \qquad (4.2.2\text{-}6)$$

verwendet. Für $\lambda = 0$ reduziert sich Gleichung (4.2.2-6) auf Gleichung (4.2.1-8) und der EM^{λ}-Algorithmus damit auf den EM-Algorithmus. Ist $\lambda \neq 0$, so konvergiert der EM^{λ}-Algorithmus leider nicht immer, und wenn er konvergiert, dann nicht immer monoton. Es kann also durchaus vorkommen, daß eine $PD^{\lambda}(\boldsymbol{\theta}, \boldsymbol{y})$-Statistik von einem Iterationszyklus zum nächsten wieder ansteigt oder daß der Algorithmus nie zu Schätzwerten führt, die dem Konvergenzkriterium genügen. Mit etwas Geduld läßt sich jedoch für $\lambda \neq -1$ unter der Nebenbedingung $\varepsilon \cdot (1 + \lambda) > 0$ immer ein reellwertiger Schrittweitenkorrekturfaktor ε finden, der die vom M-Schritt gelieferten Schätzwerte so korrigiert, daß Konvergenz entsteht (Hu & Batchelder, 1994, S. 38, Beobachtung 9)[39]:

$$\text{Korrigiertes } \theta_s^{(l)} = \theta_s^{(l-1)} - \varepsilon \cdot \left(\theta_s^{(l-1)} - \theta_s^{(l)}\right). \qquad (4.2.2\text{-}7)$$

Wenn der ε-korrigierte EM^{λ}-Algorithmus konvergiert, dann ist sichergestellt, daß es sich um ein lokales oder globales Minimum der jeweiligen $PD^{\lambda}(\boldsymbol{\theta}, \boldsymbol{y})$-Statistik handelt, egal welcher ε-Wert gewählt wurde (Hu & Batchelder, 1994, S. 37, Beobachtung 8). Wiederholungen des Algorithmus für die gleichen Daten können dann klären, ob wirklich ein globales Minimum gefunden wurden. Die Ausführungen zur Bedeutung des Ranges der Jacobischen Matrix an der Stelle der Lösung gelten hier analog (vgl. Abschnitt 4.2.1).

[39] Beobachtung 9 bei Hu und Batchelder (1994, S. 38) ist eigentlich eine stärkere Behauptung, da sie den Fall $\lambda = -1$ nicht explizit ausschließt. Man kann sich jedoch leicht davon überzeugen, daß auch der ε-korrigierte EM^{λ}-Algorithmus für $\lambda = -1$ niemals konvergieren kann, weil sich bei diesem λ-Wert die Parameterschätzungen von einem M-Schritt zum nächsten nicht ändern (vgl. Gleichung (4.2.2-6)).

Obwohl der ε-korrigierte EM$^{\lambda}$-Algorithmus offenkundig nicht ganz so narrensicher wie der EM-Algorithmus der ML-Parameterschätzung verwendbar ist, scheint er im Regelfall dennoch die Methode der Wahl zu sein, wenn man Minimum-PD$^{\lambda}$-Schätzer für binäre (Hu & Batchelder, 1994) oder nichtbinäre MVB-Modelle (Hu, 1999a) berechnen möchte.

4.2.3 Andere Schätzer

Die Minimum-Potenz-Divergenz-Familie deckt sehr viele Schätzer für multinomiale Modelle ab, umfaßt jedoch nicht alle Methoden, die im Zusammenhang mit Häufigkeitsdaten verwendet wurden (vgl. Agresti, 1990, Kapitel 13). Krauth (1982, S. 292) und Yonelinas (z.B. 1999) haben beispielsweise den aus dem Allgemeinen Linearen Modell (ALM) bekannten Kleinst-Quadrat-Schätzer (least-squares estimator) herangezogen, d.h. die euklidische Distanz zwischen beobachteten und erwarteten Häufigkeiten minimiert. Dieser Schätzer hat im ALM viele positive Eigenschaften; insbesondere ist er in diesem Modellrahmen ein BAN- und zugleich ein ML-Schätzer. Beides gilt jedoch in parametrisierten multinomialen Modellen nicht, so daß eine statistische Begründung für die Wahl dieses Schätzers nicht gegeben werden kann. Allenfalls pragmatische Argumente (z.B. Einfachheit der Berechnung) könnten genannt werden, aber auch diese Argumente greifen in Anbetracht der derzeit verfügbaren Computerprogramme, welche BAN-Schätzer verwenden oder deren Implementation zumindest sehr einfach machen, nicht mehr.

Dem Kleinst-Quadrat-Schätzer relativ eng verwandt ist der gewichtete Kleinst-Quadrat-Schätzer (weighted least-squares estimator), den Grizzle, Starmer und Koch (1969) zur Analyse linearer Modelle für Häufigkeitsdaten vorgeschlagen haben (GSK-Schätzer). Diese Modifikation des Kleinst-Quadrat-Schätzers führt jedoch — zumindest für den Fall linearer Modellgleichungen — zurück in die Familie der Potenz-Divergenz-Statistiken. Der GSK-Schätzer ist nämlich in diesem Fall äquivalent mit dem Minimum-PD$^{\lambda=-2}$-Schätzer (vgl. Bishop et al., 1975; Read & Cressie, 1988, S. 31).

Der Hauptgrund, andere Schätzer als ML- und Minimum-PD$^{\lambda}$-Schätzer in Erwägung zu ziehen, dürfte die Unklarheit über das Verhalten dieser Standardschätzer bei kleinen Stichprobenumfängen sein. Ein weiterer Grund, an „robuste" Alternativen zu den Standardschätzern zu denken, ist dann gegeben, wenn die Annahmen des multinomialen Rahmenmodells in einem konkreten Anwendungsfall fraglich sind. Weir (1990, S. 66f) schlägt als eine Option zur Lösung dieser Probleme die Momentmethode der Parameterschätzung vor. Der Grundgedanke besteht hier darin, *S* funktional unabhängige Statistiken zu finden — z.B. Kategoriehäufigkeiten oder Summen bzw. Differenzen von Kategoriehäufigkeiten —, die bestimmte Funktionen der *S* Modellparameter erwartungstreu schätzen. Das Glei-

chungssystem löst man dann nach den Modellparametern auf und bestimmt durch Einsetzen der Stichprobendaten Momentschätzer der Parameter (vgl. auch Bishop et al., 1975, S. 505 ff). Allerdings wird die Unsicherheit über die Eigenschaft der Schätzer dadurch nicht beseitigt: „There is no general theory giving properties of moment estimators, and numerical methods may have to be used to find properties such as mean and variance" (Weir, 1990, S. 68). Insbesondere darf natürlich aus der Tatsache, daß durch die Momentmethode bestimmte Funktionen der Modellparameter erwartungstreu geschätzt werden, nicht abgeleitet werden, daß auch die Modellparameter selbst erwartungstreu geschätzt werden.

Nicht unerwähnt bleiben sollten Bayessche Methoden der Parameterschätzung (einführend: Molenaar & Lewis, 1996). Der Grundgedanke ist hier, daß man die Modellparameter nicht als reellwertige Konstanten, sondern als Zufallsvariablen auffaßt, die selbst eine Verteilung aufweisen. Für die Parameter von binären MVB-Modellen wählt man dabei typischerweise die Beta-Verteilung, für nichtbinäre MVB-Modelle die Dirichlet-Verteilung (vgl. dazu Abschnitt 5.4.2.1), denn diese Verteilungen garantieren Werte im Intervall [0, 1], die sich über disjunkte und exhaustive Kategorien hinweg zu eins aufaddieren (vgl. Batchelder, 1998; Chechile, 1998; Chechile & Meyer, 1976). Bei der Anwendung der Methode muß zunächst aufgrund von Plausibilitätsargumenten eine Apriori-Verteilung für die Parameter festgelegt werden. Unter Benutzung des Bayesschen Theorems wird dann aus dieser Apriori-Verteilung und den Stichprobendaten die Aposteriori-Verteilung errechnet. Die Mittelwerte der Aposteriori-Verteilung dienen als Punktschätzer der Parameter und deren Streuungen als Bereichsschätzer.

Der Vorteil der Bayesschen Parameterschätzmethode ist, daß bei gut begründeten Vorannahmen, die in die Apriori-Verteilungen einfließen, auch schon sehr geringe Stichprobenumfänge ausreichen können, um zu gut interpretierbaren Ergebnissen der Parameterschätzung zu gelangen. Aus diesem Grunde bietet sich die Methode insbesondere zur Analyse versuchspersonenspezifischer Datensätze an, die typischerweise einen geringen Stichprobenumfang aufweisen (vgl. dazu Abschnitt 5.4.1.2). Der Nachteil der Methode liegt aber ebenfalls auf der Hand: Sie ist — wie alle Bayesschen Methoden — sehr subjektiv insofern, als die Festlegung der Apriori-Verteilung natürlich das Ergebnis der Parameterschätzung beeinflußt. Einen Ansatz zur Überwindung dieses Problems, der allerdings sehr rechenaufwendig ist, wurde jüngst von Chechile (1998) vorgestellt. Es liegen bislang noch zu wenig Ergebnisse vor, um diese Methode abschließend zu beurteilen.

4.2.4 Welchen Schätzer sollte man wählen?

Insgesamt scheinen keine starken Argumente für die Verwendung von Schätzern außerhalb der PD^{λ}-Familie vorzuliegen. Bei gesicherten Verteilungsannahmen

und großen Stichproben ist der Fall klar: Nur BAN-Schätzer sollten in diesem Fall in Erwägung gezogen werden. Die Probleme, die PD^{λ}-Minimierer bei kleinen Stichproben und Verletzungen der Verteilungsannahmen eventuell aufweisen, lösen andere Schätzer nur möglicherweise. Ohne extensive Simulationsstudien, welche die Überlegenheit alternativer Schätzer in konkreten Anwendungsfällen klar aufzeigen, ist zunächst einmal kein Grund gegeben, auf die Alternativen zurückzugreifen.

Innerhalb der PD^{λ}-Familie sprechen die analytischen Argumente (Effizienz zweiter Ordnung und geringes Hodges-Lehmann-Defizit) für einen λ-Parameter im Intervall [0, 1]. Generell sollte man Minimum-PD^{λ}-Schätzer mit $\lambda < -2$ und $\lambda > 2$ meiden, da diese extrem sensitiv auf einzelne Kategorien reagieren, in denen die empirischen Häufigkeiten deutlich unter bzw. deutlich über den erwarteten Frequenzen liegen. Derartige Schätzer minimieren folglich in erster Linie das größte negative bzw. positive Residuum und ignorieren dabei weitgehend die restlichen Kategorien des Modells. Das macht sie sehr anfällig gegenüber „Ausreißerkategorien".

Speziell bei kleinen Stichproben sollte $\lambda \leq -1$ unbedingt vermieden werden, weil die entsprechenden PD^{λ}-Statistiken für Kategorien mit Nullhäufigkeiten nicht definiert sind. Eine Monte-Carlo-Studie von Garcia-Pérez (1994), bei der jeweils 400 Zufallsstichproben zu einem einparametrigen binären MVB-Modell mit $J = 3$ Kategorien und einem Gesamtstichprobenumfang von $N = 40$ für verschiedene Parameterwerte simuliert wurden, zeigte — wie aus dem genannten Grund zu erwarten — einen starken mittleren Schätzfehler (bias) aller Minimum-PD^{λ}-Schätzer mit $\lambda \leq -1$. Die gleiche Studie erbrachte für $\lambda \in \{-1/2, 0, 2/3, 1\}$ zwar keine perfekten, aber doch sehr zufriedenstellende Ergebnisse. Alle diese Schätzer waren trotz des geringen Stichprobenumfanges annähernd erwartungstreu und hatten ungefähr gleiche Varianz. Nur bei extremen Parameterwerten in der Nähe von null und eins gab es Ausnahmen. Die Unterschiede zwischen den genannten Schätzern waren minimal, mit geringfügigen Vorteilen für den Minimum-$PD^{\lambda=2/3}$-Schätzer (Cressie-Read-Schätzer; vgl. Garcia-Peréz, 1994, S. 255-258).

Riefer und Batchelder (1991b, S. 321 ff.) haben eine ähnliche Monte-Carlo-Studie zum Enkodierung-Abruf-Modell von Batchelder und Riefer (1980, 1986) durchgeführt, in der allerdings nur der ML-Schätzer evaluiert wurde. Simuliert wurden jeweils 1000 Stichproben mit $N_1 \in \{30, 60, 90, 120, 150\}$ Itempaaren und $N_2 \in \{15, 30, 45, 60, 75\}$ Einzelitems. Auch hier zeigten sich selbst im Falle kleiner Stichproben nur geringe mittlere Verzerrungen der Schätzer, wenn die Verteilungsannahmen des multinomialen Rahmenmodells erfüllt waren.[40] Allerdings

[40] Riefer und Batchelder (1991) haben zusätzlich noch die Robustheit der Ergebnisse im Falle von Verletzungen der Verteilungsannahmen untersucht. Ich komme in Kapitel 5 hierauf zurück.

lieferten die Ergebnisse auch erste Hinweise darauf, daß die Schätzer verschiedener Parameter eines Modells durchaus unterschiedlich abschneiden können: Während sich der ML-Schätzer des Speicherparameters *c* selbst beim kleinsten Stichprobenumfang perfekt verhielt, ergab sich beim ML-Schätzer des Abrufparameters *r* im Falle sehr kleiner Stichprobenumfänge ein deutlicher positiver Bias bis zur Größenordnung $\hat{E}(\hat{r}) - r = .1$, wenn der *c*-Parameter klein ist (z.B. $c = .2$). Das Symbol $\hat{E}(\hat{r})$ steht hierbei für den empirischen Mittelwert der *r*-Schätzer über die 1000 simulierten Stichproben. Auch diese Verzerrung verschwindet allerdings, wenn man zu größeren Stichprobenumfängen übergeht. Falls $N > 200$ ist — was für die meisten Anwendungen parametrisierter multinomialer Modelle erfüllt sein dürfte — ist beim Enkodierung-Abruf-Modell kein Grund zur Sorge gegeben. Analytische Argumente sprechen dafür, daß *alle* PD^{λ}-Schätzer des *r*-Parameters im Bereich sehr kleiner Stichprobenumfänge auf Probleme stoßen, wenn *c* klein ist (Riefer & Batchelder, 1991b, S. 320).

Pragmatische Argumente sprechen eindeutig für die Verwendung des ML-Schätzers. Dies gilt zumindest für MVB-Modelle und verwandte Modelle, die mit dem beschriebenen EM-Algorithmus analysierbar sind. Bedenkt man, daß der ML-Schätzer in den Simulationsstudien von Garcia-Peréz (1994) sowie Riefer und Batchelder (1991b) insgesamt sehr zufriedenstellend abgeschnitten hat, spricht eigentlich nichts gegen dessen Verwendung auch im Falle kleiner Stichproben. Allerdings sollte die endgültige Wahl des Schätzers nicht unabhängig von der Wahl der Prüfstatistik zur Evaluation der Modellgeltungshypothese erfolgen (vgl. Abschnitt 4.3). Damit für ein bestimmtes λ in jedem Anwendungsfall das im Sinne der Modellgeltungshypothese günstigste Ergebnis resultiert, sollte man Schätzer und Prüfstatistik aufeinander abstimmen (vgl. Read & Cressie, 1988, S. 32). Ich werde deshalb in Abschnitt 4.3.5 noch einmal auf die gemeinsame Wahl von Schätzer und Prüfstatistik zurückkommen.

Alle Empfehlungen in diesem Abschnitt stehen unter dem Vorbehalt, daß das (verbundene) multinomiale Rahmenmodell als angemessen gelten kann. Wenn starke Zweifel daran bestehen, daß vorliegende Kategoriehäufigkeiten als Realisationen multinomialverteilter Zufallsvariablen aufgefaßt werden können, muß man sich mit der Robustheit der Schätzer auseinandersetzen. Ich komme in Kapitel 5 hierauf zurück.

4.2.5 Konfidenzintervalle und Konfidenzbereiche

Bislang wurden ausschließlich Punktschätzer für Parameter behandelt. Oft besteht aber auch Interesse an Bereichsschätzern, d.h. Konfidenzintervallen und — bei mehreren zu schätzenden Parametern — Konfidenzbereichen. Konfidenzintervalle und Konfidenzbereiche geben Auskunft über die Zuverlässigkeit der Pa-

rameterschätzung und erlauben damit eine Bewertung der Abweichungen der Punktschätzungen von a priori spezifizierten Parameterhypothesen.

Ist $I_S \subseteq \mathbb{R}$ das zulässige Intervall für den Parameter θ_S, $s = 1, ..., S$, so ist ein Konfidenzintervall für den Parameter θ_S zum Niveau α eine Funktion φ, die allen möglichen Kategoriehäufigkeiten $\boldsymbol{Y} = \boldsymbol{y}$ reelle Intervalle $[L_u(\boldsymbol{Y}), L_o(\boldsymbol{Y})] \subseteq I_S$ zuordnet, derart, daß der Parameter θ_S mit einer Wahrscheinlichkeit von $1-\alpha$ in dem Intervall $[L_u(\boldsymbol{Y}), L_o(\boldsymbol{Y})]$ liegt, d.h.

$$Pr(L_u(\boldsymbol{Y}) \leq \theta_S \leq L_o(\boldsymbol{Y})) \approx 1-\alpha\,. \qquad (4.2.5\text{-}1)$$

Die Intervallgrenzen $L_u(\boldsymbol{Y})$ und $L_o(\boldsymbol{Y})$ sind Funktionen des Zufallsvektors $\boldsymbol{Y}$ und damit selbst Zufallsvariablen. In Formel (4.2.5-1) wird das Zeichen „≈" (lies: „gleich oder minimal größer als") verwendet, weil man bei kategorialen Daten exakte Gleichheit in der Regel nicht erzielen kann; die gleiche Notation verwenden auch Bishop et al. (1975, S. 499).[41] Konfidenzbereiche sind multivariate Verallgemeinerungen von Konfidenzintervallen: Ein Konfidenzbereich zum Niveau $1-\alpha$ ist also eine Funktion φ', die Kategoriehäufigkeiten $\boldsymbol{Y}$ bestimmte Teilmengen des Parameterraumes zuordnen, in denen der Parametervektor $\boldsymbol{\theta}$ mit Wahrscheinlichkeit $1-\alpha$ liegt (vgl. Elandt-Johnson, 1971, Abschnitte 11.6 - 11.9).

Um Konfidenzintervalle für Parameter in parametrisierten multinomialen Modellen bestimmen zu können, benötigt man die Varianz von Schätzern, für Konfidenzbereiche zusätzlich auch noch die Kovarianz von Schätzern. In diesem Zusammenhang sind die asymptotischen Eigenschaften von BAN-Schätzern äußerst nützlich. BAN-Schätzer folgen — wie bereits erwähnt — im Falle eines einzelnen zu schätzenden Parameters asymptotisch einer Normalverteilung, im Falle mehrerer zu schätzender Parameter asymptotisch einer multivariaten Normalverteilung mit Erwartungswertvektor $\boldsymbol{\theta}$ und Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Da BAN-Schätzer asymptotisch effizient sind, entspricht $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ der Cramer-Rao-Schranke, also gerade dem Inversen der sogenannten Fisher-Informationsmatrix $\boldsymbol{I}(\boldsymbol{\theta})$ (vgl. Elandt-Johnson, 1971; Hu & Batchelder, 1994; Klauer, 1996). Die Fisher-Informationsmatrix ist im allgemeinen nicht bekannt, da sie typischerweise vom unbekannten Parametervektor $\boldsymbol{\theta}$ abhängt; sie kann aber bei hinreichend großem Stichprobenumfang entweder durch die geschätzte erwartete oder durch die beobachtete Fisher-Informationsmatrix approximiert werden, wobei verschiedene Argumente für die Verwendung der beobachteten Matrix sprechen (Efron & Hinkley, 1978). Bei Verwendung des ML-Schätzers entspricht die beobachtete Informationsmatrix $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ immer der mit -1 multiplizierten $S \times S$-Matrix der zweiten partiellen Ableitungen der logarithmierten Likelihoodfunktion nach den S Modellpara-

[41] Kriterien zur Differenzierung zwischen „guten" und „schlechten" Konfidenzintervallen mit der Eigenschaft (4.2.5-1) sollen hier ausgespart bleibt. Eine kurze Einführung gibt Klauer (1996, S. 106f).

metern, evaluiert an der Stelle des ML-Schätzwertes $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\theta}}$ (Efron & Hinkley, 1978). Diese Matrix läßt sich für MVB-Modelle relativ problemlos errechnen. Hu und Batchelder (1994, S. 29) haben die Formel für eine beliebige Zelle (r, s), $r = 1, ..., S$, $s = 1, ..., S$, der beobachteten Fisher-Informationsmatrix $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ bei binären MVB-Modellen angegeben:

$$-\left.\frac{\partial^2 \ln(L(\boldsymbol{\theta};\boldsymbol{y}))}{\partial\theta_s\,\partial\theta_r}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} =$$

$$= -\begin{cases} \dfrac{\partial\,\alpha_s(\hat{\boldsymbol{\theta}})/\partial\,\theta_r}{\hat{\theta}_s} - \dfrac{\partial\,\beta_s(\hat{\boldsymbol{\theta}})/\partial\,\theta_r}{1-\hat{\theta}_s} & \text{, falls } r \neq s. \\ \dfrac{\partial\,\alpha_s(\hat{\boldsymbol{\theta}})/\partial\,\theta_s}{\hat{\theta}_s} - \dfrac{\partial\,\beta_s(\hat{\boldsymbol{\theta}})/\partial\,\theta_s}{1-\hat{\theta}_s} - \left[\dfrac{\alpha_s(\hat{\boldsymbol{\theta}})}{\hat{\theta}_s{}^2} + \dfrac{\beta_s(\hat{\boldsymbol{\theta}})}{(1-\hat{\theta}_s)^2}\right], & \text{falls } r = s. \end{cases} \quad (4.2.5\text{-}2)$$

Hierbei sind $\alpha_s(\boldsymbol{\theta})$ bzw. $\beta_s(\boldsymbol{\theta})$ Hilfsfunktionen und $\partial\,\alpha_s(\boldsymbol{\theta}) / \partial\,\theta_r$ bzw. $\partial\,\beta_s(\boldsymbol{\theta}) / \partial\,\theta_r$ deren partielle Ableitungen. Setzt man ML-Schätzwerte $\hat{\boldsymbol{\theta}}$ ein, so ergibt sich:[42]

$$\alpha_s(\hat{\boldsymbol{\theta}}) = \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \sum_{i=1}^{I_{k,j}} a_{k,j,i,s} \cdot \frac{p_{k,j,i}(\hat{\boldsymbol{\theta}})}{p_{k,j}(\hat{\boldsymbol{\theta}})} \;,\; \beta_s(\hat{\boldsymbol{\theta}}) = \sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \sum_{i=1}^{I_{k,j}} b_{k,j,i,s} \cdot \frac{p_{k,j,i}(\hat{\boldsymbol{\theta}})}{p_{k,j}(\hat{\boldsymbol{\theta}})} \,,$$

$$\partial\,\alpha_s(\hat{\boldsymbol{\theta}})/\partial\,\theta_r = -\sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \left[\sum_{i=1}^{I_{k,j}} a_{k,j,i,s} \frac{p_{k,j,i}(\hat{\boldsymbol{\theta}}) \cdot \sum_{l=1}^{I_{k,j}} \left[\frac{a_{k,j,l,r}}{\hat{\theta}_r} - \frac{b_{k,j,l,r}}{1-\hat{\theta}_r}\right] \cdot p_{k,j,l}(\hat{\boldsymbol{\theta}})}{(p_{k,j}(\hat{\boldsymbol{\theta}}))^2} - \sum_{i=1}^{I_{k,j}} a_{k,j,i,s} \frac{\left[\frac{a_{k,j,i,r}}{\hat{\theta}_r} - \frac{b_{k,j,i,r}}{1-\hat{\theta}_r}\right] \cdot p_{k,j,i}(\hat{\boldsymbol{\theta}})}{p_{k,j}(\hat{\boldsymbol{\theta}})} \right]$$

$$\partial\,\beta_s(\hat{\boldsymbol{\theta}})/\partial\,\theta_r = -\sum_{k=1}^{K}\sum_{j=1}^{J_k} y_{k,j} \cdot \left[\sum_{i=1}^{I_{k,j}} b_{k,j,i,s} \frac{p_{k,j,i}(\hat{\boldsymbol{\theta}}) \cdot \sum_{l=1}^{I_{k,j}} \left[\frac{a_{k,j,l,r}}{\hat{\theta}_r} - \frac{b_{k,j,l,r}}{1-\hat{\theta}_r}\right] \cdot p_{k,j,l}(\hat{\boldsymbol{\theta}})}{(p_{k,j}(\hat{\boldsymbol{\theta}}))^2} - \sum_{i=1}^{I_{k,j}} b_{k,j,i,s} \frac{\left[\frac{a_{k,j,i,r}}{\hat{\theta}_r} - \frac{b_{k,j,i,r}}{1-\hat{\theta}_r}\right] \cdot p_{k,j,i}(\hat{\boldsymbol{\theta}})}{p_{k,j}(\hat{\boldsymbol{\theta}})} \right].$$

[42] Hier ist eine korrigierte Version der Formeln von Hu und Batchelder (1994, S. 28f) wiedergegeben. In der von Hu und Batchelder publizierten Version ist das Minuszeichen bei der Definition der beobachteten Fisher-Informationsmatrix vergessen worden (linke Seite von Gleichung (4.2.5-2)). Außerdem hat sich in die Ableitungen zu den Hilfsfunktionen α_s und β_s im Zähler des ersten Summanden jeweils ein Auslassungsfehler eingeschlichen. Ich danke Christoph Klauer und Martin Brandt für Hinweise auf diesen Fehler.

In obigen Gleichungen steht $p_{k,j}(\hat{\boldsymbol{\theta}})$ für die aus $\hat{\boldsymbol{\theta}}$ vorhergesagte Wahrscheinlichkeit für die Kategorie $C_{k,j}$ und $p_{k,j,i}(\hat{\boldsymbol{\theta}})$ für die vorhergesagte Wahrscheinlichkeit des i-ten Astes in die Kategorie $C_{k,j}$ (vgl. dazu Gleichung 3.5-1). Die Erweiterung auf nichtbinäre MVB-Modelle hat Hu (1999a) behandelt. Für andere parametrisierte multinomiale Modelle muß die geschätzte erwartete oder die beobachtete Fisher-Informationsmatrix im Regelfall erst hergeleitet werden (vgl. dazu Elandt-Johnson, 1971, Kapitel 12; Riefer & Batchelder, 1988, S. 321-323), es sei denn, ein solches Modell entpuppt sich als Spezialfall eines der Standardmodelle für kategoriale Daten, für die entsprechende Formeln der einschlägigen Literatur entnommen werden können (Andersen, 1990; Bishop et al., 1975; Haberman, 1978; 1979).

Der nächste Schritt besteht darin, das Inverse von $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ zu bestimmen. Dies ist ein Schätzer für die asymptotische Varianz-Kovarianz-Matrix:

$$\Sigma(\hat{\boldsymbol{\theta}}) = \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1} \tag{4.2.5-3}$$

Das Inverse $\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}$ existiert, wann immer $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ nichtsingulär ist, und das ist genau dann der Fall, wenn das zugrundeliegende Modell an der Stelle der Lösung lokal identifizierbar ist. Ist dies z.B. dadurch sichergestellt, daß die Jacobische Matrix an der Stelle der ML-Lösung vollen Rang S hat, dann muß auch das Inverse $\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}$ existieren; umgekehrt kann die Existenz von $\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}$ als Beleg für die lokale Identifizierbarkeit angesehen werden (vgl. Formann, 1984, S. 24-26; McHugh, 1956, S. 337).

In der Hauptdiagonale von $\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}$ stehen die geschätzten Varianzen $Var(\hat{\Theta}_s)$, $s = 1, ..., S$, der ML-Schätzer; entsprechend erhält man die Standardschätzfehler $Std(\hat{\Theta}_s)$ einfach durch Wurzelziehen. Die Statistik $(\hat{\Theta}_s - \theta_s)/Std(\hat{\Theta}_s)$ ist folglich asymptotisch standardnormalverteilt, so daß bei hinreichend großem Stichprobenumfang

$$Pr(-z_{\alpha/2} \le \frac{\hat{\Theta}_s - \theta_s}{\sqrt{Var(\hat{\Theta}_s)}} \le z_{\alpha/2}) \approx 1-\alpha \tag{4.2.5-4}$$

gelten muß, wobei $z_{\alpha/2}$ den kritischen z-Wert bezeichnet, oberhalb dessen $\alpha/2 \cdot 100\%$ der Standardnormalverteilung liegen. Durch einfache Termumformung erhält man daraus ein Konfidenzintervall zum Niveau α:

$$Pr(\theta_s \in [\hat{\Theta}_s - z_{\alpha/2} \cdot \sqrt{Var(\hat{\Theta}_s)},\ \hat{\Theta}_s + z_{\alpha/2} \cdot \sqrt{Var(\hat{\Theta}_s)}]) \approx 1-\alpha\,. \tag{4.2.5-5}$$

Da die Kovarianzen der Schätzer verschiedener Parameter im Regelfall nicht null sind, können auch Konfidenzregionen von Interesse sein. Hierbei handelt es sich im Falle zweier bivariat normalverteilter Schätzer um Konfidenzellipsen bzw. im höherdimensionalen Falle um Konfidenzellipsoide, innerhalb derer der Anteil $(1-\alpha)$ der gemeinsamen Verteilung der Schätzer mit den höchsten Wahrschein-

lichkeitsdichten liegt (Johnson & Kotz, 1972, S. 86f). Ein konkretes Beispiel für Konfidenzregionen beim Enkodierung-Abruf-Modell geben Batchelder und Riefer (1986, S. 143f).

Die Verläßlichkeit der geschätzten Konfidenzintervalle und Konfidenzbereiche hängt davon ab, ob (a) die Verteilungsvoraussetzungen parametrisierter Modelle erfüllt sind, (b) BAN-Schätzer der Parameter Verwendung finden und (c) der Stichprobenumfang hinreichend groß ist, so daß die beobachtete Fisher-Informationsmatrix eine gute Approximation an die faktische Informationsmatrix darstellt. Sind alle Bedingungen erfüllt, so kann man den geschätzten Konfidenzintervallen unabhängig vom verwendeten BAN-Schätzer trauen. Wenig bekannt ist über die Güte der Approximation bei kleinen Stichprobenumfängen. Meines Wissens haben sich bislang lediglich Riefer und Batchelder (1991b, S. 324f) in der bereits erwähnten Monte-Carlo-Studie zum Enkodierung-Abruf-Modell mit der Approximationsgüte faktischer Konfidenzintervalle für Parameter aus MVB-Modellen durch geschätzte asymptotische Konfidenzintervalle beschäftigt. Die Ergebnisse fielen ähnlich wie die zur finiten Erwartungstreue der Schätzer aus. Die asymptotischen Konfidenzintervalle zum *c*-Parameter stimmten im Durchschnitt nahezu perfekt mit den faktischen Konfidenzintervallen überein. Dagegen überschätzten die asymptotischen Konfidenzintervalle zum *r*-Parameter die faktischen Konfidenzintervalle, wenn der *c*-Parameter sowie der Stichprobenumfang klein waren. Auch hier zeigte sich jedoch, daß — wie zu erwarten — der Fehler mit wachsendem Stichprobenumfang immer kleiner wird; bereits bei $N_1 = 150$ Itempaaren und $N_2 = 75$ Einzelitems kann man ihn vernachlässigen.

Zur Frage, welcher BAN-Schätzer bei kleinen Stichprobenumfängen die verläßlichsten Konfidenzintervalle und -bereiche liefert, wurde bislang meines Wissens noch keine gezielte Simulationsstudie durchgeführt. Garcia-Pérez (1994, S. 256f) hat allerdings die Varianzen der verschiedenen BAN-Schätzer für das von ihm untersuchte einparametrige binäre MVB-Modell graphisch veranschaulicht. Sieht man einmal von den PD^{λ}-Minimierern mit $\lambda \leq -1$ ab, die man ohnehin vermeiden sollte (vgl. Abschnitt 4.2.4), so lieferten alle Schätzer bei $N = 40$ ungefähr gleiche Varianzen. Dies läßt vermuten, daß die Wahl des Schätzers für PD^{λ}-Minimierer mit $-0.5 \leq \lambda \leq 1$ die Güte der Konfidenzintervallschätzung nicht wesentlich beeinflußt. Aber diese Vermutung ist mit Vorsicht zu genießen, da sie bislang nur für ein einziges binäres MVB-Modell und nur für einen Stichprobenumfang geprüft wurde. Sicher ist nur, daß große Vorsicht bei der Interpretation der Varianzen und Kovarianzen von Schätzern geboten ist, wenn einige der geschätzten Varianzen negativ werden. Dies weist auf erhebliche Stichprobenfehler hin, die möglicherweise auch die restlichen Elemente der geschätzten Varianz-Kovarianz-Matrix belasten.

Bei kleinen Stichprobenumfängen (z.B. bei Gesamtstichprobenumfängen unter 100) sollte man auf die Anwendung asymptotischer Methoden sicherheitshalber besser verzichten. Dies gilt auch, wenn der Stichprobenumfang zwar größer als 100 ist, die geschätzte Varianz-Kovarianz-Matrix jedoch Auffälligkeiten wie z.B. negative Varianzschätzungen zeigt oder wenn Parameterschätzungen auf oder in der Nähe des Randes des Parameterraumes liegen, so daß die asymptotischen Konfidenzintervallschätzungen nicht mehr Teilmengen des Intervalls [0, 1] sind (vgl. auch Hu & Phillips, 1999). Stattdessen sollte man an Methoden der Bestimmung von Bereichsschätzern denken, die weniger Annahmen voraussetzen als asymptotische Verfahren. Insbesondere ist hier der Bootstrap zu nennen (Efron, 1982; Efron & Tibshirani, 1993; McLachlan & Krishnan, 1997, S. 139f.; Shao & Tu, 1995; Sievers, 1998). Beim gebräuchlichsten Bootstrap-Verfahren — dem sogenannten nonparametrischen Bootstrap — werden keine Apriori-Annahmen über die Verteilung in der Grundgesamtheit gemacht; stattdessen versucht man, eine möglichst optimale Schätzung der Verteilung in der zugrundeliegenden Population aus einer vorliegenden Stichprobe abzuleiten. Aus der geschätzten zugrundeliegenden Verteilung generiert man dann die gemeinsame Stichprobenverteilung der Parameterschätzer mittels Computersimulation und berechnet aus dieser empirisch erzeugten Stichprobenverteilung Varianzen und Kovarianzen der Schätzer sowie Konfidenzintervalle oder Konfidenzbereiche.

Wird der nonparametrische Bootstrap auf kategoriale Daten angewendet, so verwendet man die beobachtete Häufigkeitsverteilung in der Stichprobe als Schätzung für die Populationsverteilung. Die Stichprobenverteilung erhält man dann durch Ziehen von Zufallsstichproben aus der Originalstichprobe, wobei jede Beobachtung nach der Ziehung zurückgelegt wird („Ziehen mit Zurücklegen”). Diese sogenannten Bootstrap-Stichproben müssen jeweils den gleichen Umfang N wie die Originalstichprobe haben (bzw. den gleichen Stichprobenumfangsvektor $\boldsymbol{N} = (N_1, ..., N_K)$, sofern es sich um ein verbundenes multinomiales Modell handelt). Hat man eine hinreichende Anzahl von Bootstrap-Stichproben (z.B. 50-100, vgl. Efron & Tibshirani, 1993, S. 50ff) aus der Originalstichprobe generiert, so berechnet man die Minimum-PD$^{\lambda}$-Schätzer der Modellparameter für jede dieser Bootstrap-Stichproben. Für die resultierenden 50 bis 100 Vektoren berechnet man dann die Varianz-Kovarianz-Matrix der Schätzer, wobei biaskorrigierte Formeln verwendet werden sollten (Efron & Tibshirani, 1993, Kapitel 6; McLachlan & Krishnan, 1997, S. 139f.) So erhält man eine erwartungstreue Schätzung der exakten Varianz-Kovarianz-Matrix bei dem analysierten Stichprobenumfang N bzw. bei gegebenem Stichprobenumfangsvektor $\boldsymbol{N}$.

Aus den geschätzten Varianzen kann man durch Wurzelziehen Standardschätzfehler ableiten und im Prinzip auch Konfidenzintervalle, indem man die oben geschilderten Schritte zur Herleitung asymptotischer Konfidenzintervalle

analog anwendet (vgl. Hu & Phillips, 1999). Allerdings gibt es bessere Methoden der Gewinnung von Konfidenzintervallen aus Bootstrap-Stichproben, „besser" in dem Sinne, daß nominelle und faktische Abdeckungswahrscheinlichkeiten genauer übereinstimmen. Hierzu zählt das intuitiv naheliegende Perzentilverfahren, bei dem man die beiden Parameterschätzungen, die das $\alpha/2 \cdot 100$-ste und das $(1-\alpha/2) \cdot 100$-ste Perzentil der Bootstrap-Stichprobenverteilung des betreffenden Schätzers markieren, als Grenzwerte des Konfidenzintervalls zum Niveau α wählt. Auch das Perzentilverfahren läßt sich jedoch noch optimieren. Das BC_a-Verfahren („bias-corrected and accelerated") und das ABC-Verfahren („approximate bootstrap confidence intervals") sind Beispiele hierfür (vgl. Efron & Tibshirani, 1993, Kapitel 14). Anders als beim Bootstrap zur Bestimmung des Standardschätzfehlers benötigt man deutlich mehr als 100 Bootstrap-Stichproben für eine vernünftige Konfidenzintervallabschätzung. Efron und Tibshirani (1993, Kapitel 14) empfehlen 1000, teilweise sogar 2000 Stichproben.

Der nonparametrische Bootstrap ist offensichtlich sehr rechenaufwendig. Außerdem sollte man seine Leistungsfähigkeit nicht unkritisch überbewerten. Die Bezeichnung „nonparametrischer Bootstrap" darf beispielsweise nicht fälschlich dahingehend interpretiert werden, daß das Verfahren keinerlei Verteilungsannahmen macht und damit automatisch auch robust gegen Verletzungen von Verteilungsannahmen multinomialer Modelle ist (vgl. auch Sievers, 1998, Vorwort). Da die Ziehungen aus der Originalstichprobe jeweils mit Zurücklegen erfolgen, hat jede Beobachtung bei jeder Ziehung die gleiche Chance, in die Bootstrap-Stichprobe zu gelangen. Die Ziehungschance entspricht gerade der relativen Häufigkeit der Beobachtungskategorie in der Originalstichprobe und ist von den vorherigen Ziehungsergebnissen unabhängig. Somit ist der auf kategoriale Daten angewandte nonparametrische Bootstrap einer Monte-Carlo-Studie äquivalent, bei der Zufallsstichproben aus einer (ggf. verbundenen) Multinomialverteilung generiert werden, deren Kategoriewahrscheinlichkeiten mit den relativen Kategoriehäufigkeiten in einer beobachteten Stichprobe übereinstimmen. Im Grunde genommen gehen also die Verteilungsannahmen eines (verbundenen) multinomialen Modells (vgl. Kapitel 3.1 und 3.2) in das nonparametrische Bootstrap-Verfahren ebenfalls ein. Alle diese Annahmen werden bei der Erzeugung der Bootstrap-Stichproben benutzt. Der Hauptunterschied zur Standardanalyse multinomialer Modelle ist der Verzicht auf asymptotische Approximationen: nicht die asymptotische Stichprobenverteilung wird beim Bootstrap approximiert, sondern die exakte Stichprobenverteilung für einen gegebenen Stichprobenumfang. Das Bootstrap-Verfahren ist also zunächst einmal nur eine Absicherung gegen mögliche Diskrepanzen zwischen exakten und asymptotischen Stichprobenverteilungen; das Robustheitsproblem bei eventuellen Verletzungen der Verteilungsannahmen — z.B. als Folge in-

terindividueller Variabilität in den Parametern oder voneinander abhängiger Beobachtungen — muß gesondert betrachtet werden (vgl. Kapitel 5).

Eine weitere Frage ist, ob der nonparametrische Bootstrap wirklich das geeignetste Bootstrap-Verfahren zur Konfidenzintervallbestimmung bei parametrisierten multinomialen Modellen ist. Die Parameter sind nur innerhalb eines bestimmten Modells definiert und bei Ungültigkeit dieses Modells nicht interpretierbar. Folglich interessieren die Konfidenzintervalle der Parameterschätzungen auch nur unter der Voraussetzung, daß das betreffende Modell gilt. Genau diese Annahme geht aber in den nonparametrischen Bootstrap nicht ein. Hierbei wird lediglich eine (einfache oder verbundene) Multinomialverteilung vorausgesetzt, die aber nicht zwangsläufig durch die Modellgleichungen des betreffenden Modells erklärbar sein muß.

Wird die Gültigkeit eines bestimmten Modells unterstellt, so ist der sogenannte parametrische Bootstrap ein effizienteres (d.h. präziseres) Verfahren als der nonparametrische Bootstrap (Efron & Tibshirani, 1993, Kapitel 6.5 und 21; McLachlan & Krishnan, 1997). Beim parametrischen Bootstrap wird die Verteilung in der zugrundeliegenden Population unter der Voraussetzung geschätzt, daß das betreffende Modell gilt. Praktisch bedeutet dies, daß die parametrischen Bootstrap-Stichproben nicht aus der Originalstichprobe gezogen werden, sondern aus einer „korrigierten Originalstichprobe", die keinerlei Abweichungen vom Modell zeigt. Anstelle der beobachteten Häufigkeiten werden also die erwarteten Häufigkeiten bei Modellgültigkeit zugrunde gelegt. Äquivalent hierzu wäre wiederum eine Monte-Carlo-Studie zu einer Multinomialverteilung, die man aus den Modellgleichungen des betreffenden Modells erhält, indem man die Parameterschätzungen der Originalstichprobe für die Parameter einsetzt. Standardschätzfehler und Konfidenzintervalle ergeben sich aus den parametrischen Bootstrap-Stichproben genauso wie aus nonparametrischen Bootstrap-Stichproben.

Sicherlich kann der parametrische Bootstrap nicht robuster als der nonparametrische Bootstrap sein, da die mit der Multinomialverteilung verbundenen Annahmen beiden Verfahren gemeinsam sind. Die Methoden werden daher auf Verletzungen dieser Annahmen ähnlich reagieren. Auf der anderen Seite gibt es aber auch keine zusätzliche Robustheitsproblematik beim parametrischen Bootstrap, die man gern ausschließen möchte: Es macht keinen Sinn, Konfidenzintervalle für Parameterschätzungen zu fordern, die auch bei Ungültigkeit der Modellgleichungen ihre Gültigkeit behalten. Die Modellgeltungshypothese muß immer vorausgesetzt werden, wenn Konfidenzintervalle sinnvoll interpretierbar sein sollen, egal ob sie nonparametrisch oder parametrisch geschätzt werden. Wenn dem schon so ist, dann sollte man in den Fällen, in denen asymptotische Methoden der Konfidenzintervallbestimmung nicht vertretbar erscheinen, dem effizienteren

Verfahren — nämlich dem parametrischen Bootstrap — den Vorzug geben (vgl. auch McLachlan & Krishnan, 1997, S. 140).

4.3 Statistische Tests

Bislang wurde vorausgesetzt, daß ein bestimmtes parametrisiertes multinomiales Modell in einem konkreten Anwendungsfall gilt, d.h. daß ein Parametervektor $\boldsymbol{\theta}^{\mathbf{o}} \in \boldsymbol{\Omega}$ existiert, der auf die konkret gegebenen Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ abgebildet wird, so daß $\boldsymbol{\pi} = f(\boldsymbol{\theta}^{\mathbf{o}})$ gilt. Trivialerweise erfüllt ist diese Voraussetzung aber nur dann, wenn zu jedem prinzipiell denkbaren $\boldsymbol{\pi}$ ein entsprechender Parametervektor $\boldsymbol{\theta}$ mit der Eigenschaft $\boldsymbol{\pi} = f(\boldsymbol{\theta})$ existiert. Das ist dann der Fall, wenn f surjektiv ist, so daß $f(\boldsymbol{\Omega}) = \boldsymbol{\Gamma}$ gilt. Man nennt Modelle mit dieser Eigenschaft, die zudem auch noch identifizierbar sind, saturierte Modelle. Liegt ein saturiertes Modell vor, wie beispielsweise das Standard-SDT-Modell aus Abschnitt 3.3, so kann man direkt zur Parameterschätzung übergehen.

Ist $f(\boldsymbol{\Omega})$ dagegen eine echte Teilmenge von $\boldsymbol{\Gamma}$, dann steht das Modell möglicherweise im Widerspruch zu den konkret gegebenen Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ und ist damit testbar. Global identifizierbare Modelle sind genau dann testbar, wenn sie unterparametrisiert sind. Allerdings ist Unterparametrisiertheit keine notwendige Bedingung für Testbarkeit: Auch voll parametrisierte und sogar überparametrisierte Modelle können potentiell testbar sein, sind dann allerdings nicht identifizierbar (vgl. Bamber & van Santen, 1985; 2000). Ich komme in Abschnitt 4.6 auf die Überprüfung derartiger Modelle zurück.

Vorerst sei vorausgesetzt, daß ein wenigstens lokal (in der näheren Umgebung eines Schätzwertes $\hat{\boldsymbol{\theta}}$) identifizierbares Modell mit $S < M$ Parametern vorliegt. Die Parameterschätzungen für ein solches Modell sind nur dann sinnvoll interpretierbar, wenn davon ausgegangen werden kann, daß die gegebenen Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ den Modellrestriktionen folgen. Folglich gilt es, zunächst die Modellgeltungshypothese

$$H_0: \boldsymbol{\pi} \in f(\boldsymbol{\Omega}) \tag{4.3-1}$$

gegen die Alternativhypothese

$$H_1: \boldsymbol{\pi} \notin f(\boldsymbol{\Omega}) \tag{4.3-2}$$

zu testen. Wenn die H_0 für gegebene Stichprobendaten $\boldsymbol{Y} = \boldsymbol{y}$ beibehalten werden kann, so darf das Modell als bewährt gelten. Andernfalls muß es verworfen werden bzw. es muß nach geeigneten Modellrevisionen gesucht werden.

Bei diesem Modellgeltungstest (goodness-of-fit test) ist nicht nur die Kontrolle der Fehlerwahrscheinlichkeit 1. Art (α), sondern zusätzlich auch die Kontrolle der Fehlerwahrscheinlichkeit 2. Art (β) von zentraler Bedeutung: Weder darf das Risiko hoch sein, eine gültiges Modell fälschlich zu verwerfen (Alpha-Fehler),

noch sollte das Risiko hoch sein, ein erheblich verletztes Modell fälschlich als gültig anzusehen (Beta-Fehler, vgl. Bredenkamp, 1972; 1980; Cohen, 1988; Erdfelder, 1984; 1985). Um dies entscheiden zu können, bedarf es natürlich einer genaueren Spezifizierung dessen, was unter einem „erheblich verletzten Modell" zu verstehen ist.

Ich werde auf diesen zentralen Punkt in Abschnitt 4.3.3 zurückkommen. Zuvor soll jedoch dargestellt werden, wie eine Prüfung der Nullhypothese (4.3-1) unter Kontrolle des α-Fehlers erfolgen kann (Abschnitt 4.3.1) und wie man im Falle ihrer Beibehaltung bestimmte Parameterhypothesen zu diesem Modell bzw. Submodelle dieses Modells testen kann (Abschnitt 4.3.2). Erst im Anschluß werden Methoden zur Kontrolle des β-Fehlers vorgestellt (Abschnitt 4.3.3). In Abschnitt 4.3.4 folgt eine Diskussion der Frage, welche Prüfstatistiken am besten verwendet werden sollten. In Abschnitt 4.3.5 schließlich werden informationstheoretische Verfahren zur vergleichenden Evaluation von parametrisierten multinomialen Modellen vorgestellt, die weder gleiche Parameterzahl noch hierarchisch organisierte Modellklassen voraussetzen.

4.3.1 Modellgeltungstests

In Abschnitt 4.2.2 wurde die Familie der Minimum-PD^{λ}-Statistiken als Kriterium zur Gewinnung von BAN-Schätzern für die Parameter multinomialer Modelle eingeführt. Hierbei richtete sich das Interesse auf die Parameterwerte, die eine PD^{λ}-Statistik minimieren, nicht auf die numerischen Werte, welche die Minimum-PD^{λ}-Statistiken selbst annehmen. Diese sind jedoch aus einem anderen Grund von Interesse: Unabhängig vom gewählten λ-Wert kann man sie zur Prüfung der Modellgeltungshypothese (4.3-1) einsetzen. Wenn die sechs Regularitätsbedingungen von Birch (1964) erfüllt sind, ist die asymptotische Verteilung jeder Minimum-PD^{λ}-Statistik bei Gültigkeit der H_0 (4.3-1) eine zentrale $\chi^2_{(df)}$-Verteilung mit $df = M-S$ Freiheitsgraden (Read & Cressie, 1988, S. 50 und Anhang A6). Man erhält also bei hinreichend großem Stichprobenumfang einen approximativen Modellgeltungstest zum Niveau α, wenn man zu einem gegebenen Modell und gegebenen Daten *irgendeine* PD^{λ}-Statistik minimiert und die so erhaltene Prüfstatistik mit dem kritischen Wert der $\chi^2_{(M-S)}$-Verteilung vergleicht, der die oberen $\alpha \cdot 100\%$ der Verteilung abtrennen. Ist der Wert der Prüfstatistik größer als dieser kritische Wert, wird die H_0 (4.3-1) verworfen und sonst beibehalten.

Die vorauszusetzenden Regularitätsbedingungen von Birch (1964) sind dieselben, die auch für die Existenz von BAN-Schätzern im allgemeinen und von ML-Schätzern im besonderen vorausgesetzt werden müssen (vgl. dazu die Abschnitte 4.2.1 und 4.2.2). Wie zum Teil bereits erwähnt, verlangen sie — von oftmals unproblematischen Stetigkeits- und Differenzierbarkeitsforderungen an die

Modellgleichungen und deren Umkehrabbildung abgesehen — im wesentlichen folgendes:

(1) Gilt $\boldsymbol{\pi} = f(\boldsymbol{\theta}^{\mathbf{o}})$, ist also $\boldsymbol{\theta}^{\mathbf{o}} \in \boldsymbol{\Omega}$ der Parametervektor, der die gegebenen Kategoriewahrscheinlichkeiten „erzeugt", so darf $\boldsymbol{\theta}^{\mathbf{o}}$ nicht auf dem Rand des Parameterraumes liegen. Für MVB-Modelle impliziert dies $0 < \theta^{o}_{s} < 1$, $s = 1, ..., S$.

(2) $\boldsymbol{\theta}^{\mathbf{o}}$ wird auf Kategoriewahrscheinlichkeiten abgebildet, die sämtlich größer als null sind: $p_{k,j}(\boldsymbol{\theta}^{\mathbf{o}}) > 0$, $k = 1, ..., K$; $j = 1, ..., J_k$.

(3) Die Jacobische Matrix hat an der Stelle $\boldsymbol{\theta}^{\mathbf{o}}$ vollen Rang S.

Kritisch für die statistische Rechtfertigung des asymptotischen Tests ist also vor allem die lokale Identifizierbarkeit des Nullhypothesenmodells an der Stelle des wahren Parameters $\boldsymbol{\theta}^{\mathbf{o}}$ und die Forderung, daß — grob gesprochen — bei Modellgültigkeit die Ränder von $\boldsymbol{\Omega}$ und $\boldsymbol{\Gamma}$ ausgespart bleiben müssen. Alle diese Forderungen sind für parametrisierte multinomiale Modelle — insbesondere für MVB-Modelle — keineswegs trivialerweise erfüllt. Ist die lokale Identifizierbarkeit verletzt oder konvergieren Schätzwerte der Parameter auf Grenzwerte am Rand des Parameterraumes, so gibt es im allgemeinen keine statistische Begründung für die Verwendung der $\chi^2_{(M-S)}$-Verteilung als Referenzverteilung. Wie im Falle von Identifizierbarkeitsproblemen alternativ verfahren werden kann, wird in Abschnitt 4.6 besprochen. Liegt ein Schätzwert exakt auf dem Rand des Parameterraumes, so legt das den Verdacht nahe, daß das Modell entweder ungültig ist (H_1) oder daß der entsprechende Parameter in der zugrundeliegenden Population ebenfalls auf dem Rand bzw. in der unmittelbaren Nähe des Randes von $\boldsymbol{\Omega}$ liegt (H_0). Auch wenn kein Schätzwert grenzwertig ist, so kann $\boldsymbol{\theta}^{\mathbf{o}}$ selbstverständlich dennoch auf dem Rand von $\boldsymbol{\Omega}$ liegen.

Die asymptotische Verteilung von Modellanpassungsstatistiken kann prinzipiell auch unter einer Nullhypothese mit grenzwertigen Parametern bestimmt werden. Es handelt sich dann allerdings nicht mehr um χ^2-Verteilungen, sondern um Mischungen von χ^2-Verteilungen (für multinomiale Modelle vgl. z.B. Lee, 1987, Theorem 3.2; für Strukturgleichungsmodelle vgl. Shapiro, 1985). Insbesondere die korrekte Bestimmung der Gewichte der einzelnen Komponenten der Mischung kann in der Praxis Schwierigkeiten bereiten. Einen konzeptuell sehr einfachen, wenn auch rechnerisch aufwendigen Ausweg aus diesem Problem haben Klauer und Oberauer (1995, S. 681) vorgeschlagen. Sie generierten die H_0-Stichprobenverteilung der Modellanpassungsstatistik (in ihrem Fall G^2) durch Monte-Carlo-Simulation zahlreicher Stichproben aus einer Population, in der der Parametervektor durch den ML-Schätzwertvektor $\hat{\boldsymbol{\theta}}$ der beobachteten Stichprobe — darunter auch grenzwertige Parameterschätzungen — ersetzt wurde. Die simulierten Stichproben hatten dabei den gleichen Umfang wie die beobachtete Stichprobe. Für jede der simulierten Stichproben berechneten sie die Anpassungsstatistik für das zu testende Modell. Man erhält dann eine Schätzung der Überschreitungs-

wahrscheinlichkeit für die beobachtete Anpassungsstatistik unter der H_0 der Modellgültigkeit mit Parametervektor $\boldsymbol{\theta}^{o} = \hat{\boldsymbol{\theta}}$, indem man auszählt, welcher Anteil der simulierten Verteilung größer als die ursprünglich beobachtete Anpassungsstatistik ist. Ist dieser Anteil kleiner als ein zuvor festgelegter Wert α, so wird die H_0 verworfen und sonst beibehalten.

Das vorgeschlagenen Verfahren ist nichts anderes als der schon in Abschnitt 4.2.5 erwähnte, hier jedoch auf Modellanpassungsstatistiken angewandte parametrische Bootstrap (vgl. Efron, 1982; Efron & Tibshirani, 1993; Hall & Wilson, 1991; Klauer & Batchelder, 1996, S. 210; Langeheine, Pannekoek & van de Pol, 1996; Romano, 1988; Shao & Tu, 1995, Kapitel 4.5; Sievers, 1998; Simonoff, 1986; von Davier, 1997). Die Besonderheit des Ansatzes von Klauer und Oberauer (1995) liegt darin, daß der Bootstrap zur Schätzung von Überschreitungswahrscheinlichkeiten eingesetzt wird, wenn der H_0-Parametervektor $\boldsymbol{\theta}^{o}$ auf dem Rand von $\boldsymbol{\Omega}$ liegt. Man schlägt dabei gewissermaßen zwei Fliegen mit einer Klappe: Einerseits wird die Referenzverteilung automatisch korrigiert um den Effekt grenzwertiger Parameter. Anders als beim klassischen asymptotischen Anpassungstest muß also beim Bootstrap nicht unterschieden werden zwischen Fällen mit und ohne grenzwertige Parameter. Darüber hinaus werden Abweichungen der exakten Verteilung von der asymptotischen Verteilung beim Bootstrap zusätzlich berücksichtigt. Bedenkt man, daß das von Klauer und Oberauer vorgeschlagene Verfahren zwar sehr rechenaufwendig, aber auf einem Computer relativ leicht implementierbar ist, so scheint es auch aus pragmatischer Perspektive die Methode der Wahl zu sein, wenn eine oder mehrere Parameterschätzungen auf dem Rand des Parameterraums liegen.[43]

---

[43] Aitkin, Anderson und Hinde (1981) haben das gleiche Verfahren zur Prüfung von Latent-Class-Modellen und anderen finiten Mischverteilungsmodellen vorgeschlagen (vgl. Abschnitt 6.3). Wenn die H_0, daß ein c-Klassen-Modell gilt, gegen die H_1 getestet werden soll, daß ein $(c+1)$-Klassen-Modell gilt, hat man es ebenfalls mit dem Problem grenzwertiger Parameter zu tun: Das H_0-Modell liegt am Rande des Parameterraums des H_1-Modells, weil eine Klassenwahrscheinlichkeit des H_1-Modells gemäß H_0 den Wert Null annehmen muß. Die Konsequenz aus diesem Sachverhalt wird bei Routineanwendungen der LCA oftmals übersehen: „Thus in the two-component mixture model the distribution of $-2 \cdot \log L$ (d.h. G^2, Ergänzung des Verfassers der vorliegenden Arbeit) under the null hypothesis of a single population is unknown" (Aitkin et al., 1981, S. 424). Analog zu Klauer und Oberauer (1995) generierten Aitkin et al. (1981) deshalb Monte-Carlo-Stichproben aus einer Population, deren Parameterwerte den ML-Parameterschätzungen zum H_0-Modell für die empirische Stichprobe entsprachen. Für jede dieser Monte-Carlo-Stichproben sowie die empirische Stichprobe wurden dann die G^2-Modellanpassungsstatistiken zum H_0-Modell berechnet. Wenn die empirische Statistik in den oberen $\alpha \cdot 100\%$ der Verteilung lag, wurde H_0 verworfen und sonst beibehalten. Aitkin et al. (1981, S. 425) wiesen allerdings auch auf ein noch offenes Problem des parametrischen Bootstrap-Verfahrens hin: „It is possible that the true asymptotic distribution may depend on the parameter values under the null hypotheses". Wenn die asymptotische Verteilung von den konkreten Parameterwerten unabhängig ist, ist das Verfahren unproblematisch. Liegt jedoch eine Abhängigkeit vor, kann der parametrische Bootstrap u.U. verfälscht sein, d.h. faktische α–Niveaus liefern, die von den nominellen abweichen. Obwohl mathematisch nicht besonders elegant, kann dieses Problem prinzipiell mit Monte-Carlo-Mitteln exploriert werden: Man generiert einfach mehrere parametrische Bootstrap-Verteilungen zu G^2 für verschiedene Parameterwerte unter H_0, nicht nur diejenigen, die den ML-Schätzern entsprechen. Wenn sich diese Bootstrap-Verteilungen signifikant unterscheiden, ist der parametrische Bootstrap problematisch. Andernfalls kann er als begründet gelten. Dies bedeutet aber auch,

Sind die Regularitätsbedingungen nach Birch (1964) dagegen mutmaßlich erfüllt und ist der gegebene Stichprobenumfang hinreichend groß, so sprechen pragmatische Argumente eindeutig für die Anwendung des asymptotischen χ^2-Tests mit $\chi^2_{(M-S)}$ als Referenzverteilung. Unter dem Aspekt der α-Fehlerkontrolle gibt es dabei keinen Grund, eine der PD^λ-Statistiken zu bevorzugen, da sie bei Modellgültigkeit alle die gleichen asymptotischen Eigenschaften haben. Im Falle kleiner Stichprobenumfänge kann es dagegen durchaus Unterschiede geben, denn die exakten Verteilungen der PD^λ-Statistiken können unter Umständen unterschiedlich gut durch die $\chi^2_{(M-S)}$-Verteilung approximierbar sein. Vorliegende Simulationsstudien und auch analytische Betrachtungen zeigen ziemlich klar, daß die Approximation für $\lambda = 0$ — also die Log-Likelihood-Quotienten-Statistik — schlechter ausfällt als für $\lambda = 2/3$ (Cressie-Read-Statistik) und vor allem $\lambda = 1$ (Pearson-Statistik). Generell funktioniert die χ^2-Approximation für die Pearson-Statistik auch bei sehr kleinem N überraschend gut, insbesondere bei gleichwahrscheinlichen Kategorien (Erklärungen geben Read & Cressie, 1988; Yarnold, 1972). Je mehr λ jedoch von 1 abweicht, desto schlechter wird die Approximation (Larntz, 1978; Read, 1984).

Bei sehr kleinen Stichprobenumfängen und Modellen ohne zu schätzende Parameter ($S = 0$) kann das Problem prinzipiell durch exakte Tests umgangen werden: Man berechnet für alle möglichen Häufigkeitsverteilungen, die (a) gemessen an einem beliebigen PD^λ-Kriterium mindestens so stark wie die empirischen Daten vom zu testenden Modell abweichen und (b) mit dem gegeben Stichprobenumfangsvektor $\boldsymbol{N}$ konsistent sind, die Multinomialwahrscheinlichkeiten unter der H_0-Punkthypothese und summiert diese auf. Ist diese Summenwahrscheinlichkeit kleiner als α, wird H_0 verworfen und sonst beibehalten (z.B. Wickens, 1998a, S. 541). Dieser Test hält das α-Niveau immer ein, ist aber auch mit Hilfe schneller Computer nur bei relativ kleinen Stichprobenumfängen mit vertretbarem Zeitaufwand durchführbar (Berry & Mielke, 1995). Außerdem ist es nicht leicht, die erforderlichen (oftmals sehr kleinen) Multinomialwahrscheinlichkeiten hinreichend genau zu berechnen. Hierfür existiert jedoch ein brauchbarer Lösungsvorschlag von Garcia-Pérez (1999; in press).

Read (1984) sowie Read und Cressie (1988, Kapitel 5) haben rechnerisch einfachere Korrekturverfahren zu den asymptotischen Tests vorgeschlagen, die transformierte PD^λ-Statistiken liefern, deren exakte Verteilungen unter H_0 auch bei kleinen Stichprobenumfängen noch gut durch die χ^2-Verteilung approximierbar sind. Insbesondere die momentkorrigierte Statistik $T_c(PD^\lambda)$, die für $\lambda > -1$ den kor-

---

daß die Verwendung der empirischen ML-Schätzer beim parametrischen Bootstrap von Modellanpassungsstatistiken prinzipiell verzichtbar ist: Für die Modellprüfung sollte man auch andere H_0-Parameterwerte verwenden können, ohne daß das an den Ergebnissen etwas ändert. Für die Gewinnung von Bootstrap-Schätzern der Parameter und Bootstrap-Konfidenzintervallen (vgl. Abschnitt 4.2.5) gilt das natürlich nicht.

rekten Erwartungswert und die korrekte Varianz der entsprechenden χ^2-Referenzverteilung aufweist, läßt sich mit wenig Aufwand berechnen und hält dennoch das nominelle α-Niveau auch bei kleinem N ziemlich genau ein (Read & Cressie, 1988, S. 67-75). Es handelt sich um eine einfache lineare Transformation der PD^{λ}-Statistik. Read und Cressie (1988, S. 67) demonstrieren sie für den Fall eines voll spezifizierten einfachen multinomialen Modells mit K Kategorien und $S = 0$ zu schätzenden Parametern. Die Formeln bei Read und Cressie (1988, S. 67) lassen sich jedoch auch für den allgemeinen Fall eines verbundenen multinomialen Modells mit M funktional unabhängigen Kategoriewahrscheinlichkeiten und S zu schätzenden Parametern reformulieren, indem man statt der Kategorienzahl, die Read und Cressie mit dem Kleinbuchstaben k notieren, den Term $M-S+1$ einsetzt. Allerdings ist über die Approximationsgenauigkeit für den Fall $S > 0$ nichts bekannt.

Vorliegende Simulationsstudien zeigen, daß bei Verwendung der Read-Cressie-, der Pearson- und auch anderen Minimum-PD^{λ}-Statistiken mit $1/3 \leq \lambda \leq 3/2$ selbst im allgemeinen Fall eines verbundenen Modells mit S zu schätzenden Parametern auf die Anwendung eines Korrekturverfahrens verzichtet werden kann, sofern die erwarteten Kategoriehäufigkeiten sämtlich größer als 1 sind (Drost, Kallenberg, Moore & Oosterhoff, 1989b, S. 130; Read & Cressie, 1988, S. 75). Sogar bei erwarteten Häufigkeiten unter 1 kann die χ^2-Approximation für unkorrigierte Cressie-Read- und Pearson-Statistiken noch verwendet werden, wenn die erwarteten Häufigkeiten nicht zu stark divergieren (Hosmane, 1987; Rudas, 1986). Die χ^2-Approximation für die unkorrigierte Log-Likelihood-Quotienten-Statistik verlangt dagegen im allgemeinen deutlich größere Stichprobenumfänge, wenn eine hinreichende Approximationsgüte sichergestellt sein soll.

Für die praktische Anwendung multinomialer Modelle in der kognitiven Psychologie dürften diese Probleme von untergeordneter Bedeutung sein, da man typischerweise wenige Beobachtungskategorien und einen vergleichsweise großen Gesamtstichprobenumfang vorliegen hat, so daß die erwarteten Kategoriehäufigkeiten bei Modellgültigkeit recht groß sind. In diesem Falle sind die Unterschiede in der Approximationsgüte für die primär interessierenden Potenz-Divergenzstatistiken ($\lambda = 0$, $\lambda = 2/3$ und $\lambda = 1$) nicht gravierend; dies zeigt auch die Simulationsstudie, über die in Abschnitt 4.3.3 berichtet wird. Selbst wenn — z.B. bei komplexen, hochdimensionalen Kontingenztafeln aus Fragebogendaten — einmal der Fall kleiner erwarteter Kategoriehäufigkeiten eintreten sollte, muß das noch kein zwingender Grund sein, auf die Log-Likelihood-Quotienten-Statistik zu verzichten, da man gegebenenfalls eines der Korrekturverfahren von Read und Cressie (1988) einsetzen kann.

4.3.2 Tests von Parameterhypothesen und Submodellen

Ist ein Modellgeltungstest relativ zum a priori festgelegten Niveau α insignifikant ausgegangen, so daß die Modellgeltungshypothese beibehalten werden kann, dann richtet sich das substanzwissenschaftliche Interesse häufig auf bestimmte Null- oder Alternativhypothesen über die Parameter des betreffenden Modells oder auf bestimmte Spezialfälle (Submodelle) des Ausgangsmodells. Wenn das Ausgangsmodell ein MVB-Modell ist, so fallen die Parameter- und Submodellhypothesen typischerweise in eine der folgenden fünf Kategorien:

(1) *Hypothesen über die Gleichheit (H_0) versus Ungleichheit (H_1) von zwei oder mehr Parametern*. Wird beispielsweise ein Modell simultan für mehrere Probandengruppen formuliert, so interessiert in der Regel, ob sich alle oder zumindest einige dieser Parameter zwischen den Gruppen unterscheiden. Wenn z.B. das Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986) auf ein Zwei-Gruppen-Modell für ältere und jüngere Erwachsene erweitert wird, liegt der Test der beiden Nullhypothesen nahe, daß sich die Cluster-Speicherungsparameter c_{alt} und c_{jung} bzw. die Cluster-Abrufparameter r_{alt} und r_{jung} jeweils nicht unterscheiden (vgl. Erdfelder & Bayen, 1991; Riefer & Batchelder, 1991a).

(2) *Hypothesen über konkrete Parameterwerte (H_0) bzw. Abweichungen von diesen Werten (H_1)*. Gelegentlich interessiert, ob einer oder mehrere Parameter eines Modells Werte annehmen, die eine bestimmte Theorie über die zugrundeliegenden kognitiven Prozesse vorschreibt. Betrachten wir ein Beispiel: Unter der Annahme, daß Probanden zwar den Anteil zweier Klassen zuvor dargebotener Items gelernt haben (z.B. fragmentarisch versus vollständig dargebotene Items), sich aber beim anschließenden Gedächtnistest nicht mehr an die Klassenzugehörigkeit eines konkreten Testitems erinnern können, müssen sie — falls der Gedächtnistest eine entsprechende Quellenangabe verlangt — die Klassenzugehörigkeit raten. Folgt dieser Rateprozeß dem Prinzip des „probability matching” (vgl. z.B. Gallistel, 1990, Kapitel 11), so sollte die Ratewahrscheinlichkeit für das Urteil „fragmentarisch” beispielsweise $g = .125$ betragen, wenn ursprünglich 12.5% der Items fragmentarisch dargeboten wurden (vgl. Erdfelder & Bredenkamp, 1998, Experiment 1).

(3) *Hypothesen über die stochastische Unabhängigkeit (H_0) versus Abhängigkeit (H_1) von latenten Ereignissen in MVB-Modellen*. Gegenwärtig wird eine Debatte über die Unabhängigkeit oder Abhängigkeit bewußter und unbewußter Gedächtnisprozesse geführt, die als stochastische Unabhängigkeitshypothese im Rahmen von MVB-Modellen zur sogenannten Prozeßdissoziationsprozedur formulierbar (Buchner, Erdfelder & Vaterrodt-Plünnecke, 1995; Erdfelder & Buchner, 1995; 1998b) und unter bestimmten Zusatzannahmen

auch empirisch testbar ist (Krüger & Vaterrodt-Plünnecke, 1997). Ein weiteres Beispiel für eine stochastische Unabhängigkeitshypothese im Rahmen eines MVB-Modells — die Annahme unabhängigen Abrufs nichtgeclusterter Items eines Itempaares — hatten wir bereits in Abschnitt 3.5.1 kennengelernt.

(4) *Hypothesen über die Additivität (H_0) versus Interaktion (H_1) der Effekte von unabhängigen Variablen.* Wenn man etwa einen Rekognitionstest für ältere und jüngere Probanden unter Bereitstellung von Abrufhilfen oder ohne solche Abrufhilfen durchführt, kann von Interesse sein, ob der Rekognitionsparameter r_{2ht} des 2HT-Modells (vgl. Abschnitt 3.5.2) durch die Bereitstellung von Abrufhilfen bei jüngeren Probanden genauso stark ansteigt wie bei älteren Probanden (H_0) oder ob eine Interaktion in dem Sinne vorliegt, daß ältere Probanden von Abrufhilfen stärker profitieren als jüngere bzw. umgekehrt jüngere Vpn stärker profitieren als ältere (H_1).

(5) *Hypothesen über die Rangordnung bestimmter Parameterwerte (H_0) versus Abweichungen von der prognostizierten Rangordnung (H_1).* Wird beispielsweise ein Gedächtnistraining mit den Intensitätsstufen „einfach", „mittel" und „intensiv" bei drei experimentellen Gruppen realisiert, so ist im Rahmen eines auf alle drei Trainingsgruppen angewandten 2HT-Modells möglicherweise die H_0 $r_{2ht(einfach)} \leq r_{2ht(mittel)} \leq r_{2ht(intensiv)}$ — bezogen auf die im Anschluß an das Training erhobene Rekognitionsleistung — von Interesse.

Alle oben aufgeführten Nullhypothesen und noch weitere, die in Beobachtung 5 bei Hu und Batchelder (1994, S. 30) zusammengefaßt sind, lassen sich durch bestimmte Parameterrestriktionen aus dem Ausgangsmodell gewinnen. Ist das Ausgangsmodell ein MVB-Modell, so ist auch das restringierte Modell ein MVB-Modell, allerdings mit reduziertem Parameterraum $\boldsymbol{\Omega}$ und reduziertem Bild des Parameterraumes in $\boldsymbol{\Gamma}$. Im folgenden wird zunächst besprochen, wie sich geeignete restringierte Modelle aus einem gegeben Ausgangsmodell gewinnen lassen. Anschließend wird auf statistische Tests eingegangen, mit denen man Modellvergleiche zwischen restringiertem Modell und Ausgangsmodell vornehmen kann.

4.3.2.1 Konstruktion geeignet restringierter Modelle

Für die Hypothesenklassen (1) und (2) dürfte offensichtlich sein, wie man ein restringiertes Modell, das der zu testenden Nullhypothese entspricht, aus dem Ausgangsmodell herleitet. Im Fall (1) ersetzt man einfach alle Parameter, die gemäß H_0 den gleichen Wert annehmen sollen, durch einen Parameter, z.B. den „ersten" Parameter der betreffenden Parametergruppe. Die Anzahl der Parameter reduziert sich damit pro Gleichheitsrestriktion um 1 und in der Modellgleichung (3.4-7) für allgemeine MVB-Modelle (bzw. Gleichung (3.5-1) für binäre MVB-Modelle) än-

dern sich die Werte der *a*- und *b*-Zählvariablen für den Parameter, der nach der Gleichheitsrestriktion noch erhalten bleibt. Ansonsten sind die Modellgleichungen des restringierten Modells gegenüber dem Ausgangsmodell unverändert. Analog fällt im Fall (2) derjenige Parameter fort, der gemäß H_0 eine bestimmten Wert annehmen muß. Die Parameterzahl reduziert sich pro Parameterfixierung um 1 und die multiplikativen *c*-Konstanten in Gleichung (3.4-7) bzw. (3.5-1) ändern sich für die Äste des Modells, in denen der fixierte Parameter vorkommt, um den Faktor, der gerade dem fixierten Parameterwert entspricht. Sonst ändert sich nichts.

Nicht ganz so offensichtlich ist die Umsetzung der Nullhypothesen (3), (4) und (5). Wie sich jedoch zeigen läßt, entspricht die stochastische Unabhängigkeitshypothese (3) immer einer Parametergleichheitshypothese des Typs (1) und läßt sich folglich auch genauso behandeln. Dies wurde anhand des Enkodierung-Abruf-Modells ja auch schon exemplarisch demonstriert (vgl. Abschnitt 3.5.1). Betrachten wir ein weiteres Beispiel: Angenommen, es wird ein MVB-Modell zu einem Experiment formuliert werden, bei dem Probanden die Testitems bezüglich zweier Quellen als alt oder neu bewerten sollen. Beispielsweise werden den Probanden in der Lernphase nacheinander zwei komplexe Bilder dargeboten, die bestimmte Items enthalten oder nicht enthalten können. Für die Testitems im anschließenden Gedächtnistest ist anzugeben, ob das betreffende Item in Bild 1 vorhanden war (ja/nein) und ob es in Bild 2 vorhanden war (ja/nein). Wenn wir nun solche Testitems betrachten, die tatsächlich in beiden Quellen vorkamen, und zunächst stochastische Abhängigkeit der Erinnerungsprozesse bezüglich beider Bilder zulassen wollen, müssen wir für die Erinnerungswahrscheinlichkeiten drei Parameter vorsehen: die Wahrscheinlichkeit r_1, daß das Item als Teil von Bild 1 erinnert wird, die Wahrscheinlichkeit r_{2+}, daß das Item als Teil von Bild 2 erinnert wird, falls es schon als Teil von Bild 1 erinnert wurde, und schließlich die Wahrscheinlichkeit r_{2-}, daß das Item als Teil von Bild 2 erinnert wird, falls es *nicht* als Teil von Bild 1 erinnert wurde. Die Erinnerungsprozesse bezüglich beider Bilder sind positiv korreliert, wenn $r_{2+} > r_{2-}$ gilt; entsprechend sind sie negativ korreliert, falls $r_{2+} < r_{2-}$ gilt. Wenn man also die Hypothese der stochastischen Unabhängigkeit beider Erinnerungsprozesse testen will, muß man folglich die H_0 $r_{2+} = r_{2-}$ prüfen. Dies kann auf die gleiche Weise wie bei Parametergleichheitshypothesen des Typs (1) geschehen.

Interaktions- versus Additivitätshypothesen des Typs (4) lassen sich ebenfalls über Parametergleichheitsrestriktionen gegeneinander testen. Interessiert beispielsweise, ob das Alter der Versuchspersonen (ältere versus jüngere Probanden) einerseits und Abrufhilfen (Tests mit versus ohne Erinnerungshilfen) andererseits bezüglich der Rekognitionsleistung r_{2ht} im 2HT-Modell interagieren oder additiv wirken, so formuliert man für die beiden Altersgruppen ohne Abrufhilfe zunächst zwei Standard-2HT-Modelle mit unrestringierten altersabhängigen Rekognitions-

parametern $r_{2ht(alt)}$ und $r_{2ht(jung)}$ (vgl. Abschnitt 3.5.2). Für die beiden Gruppen mit Abrufhilfe verwendet man dagegen ein erweitertes Modell, das auf der ersten Ebene in bezug auf alte Items drei Ereignisse unterscheidet: Rekognition mit der gleichen Wahrscheinlichkeit wie ohne Abrufhilfe ($r_{2ht(alt)}$ bzw. $r_{2ht(jung)}$), Rekognition aufgrund der zusätzlichen Bereitstellung der Abrufhilfe ($a_{2ht(alt)}$ bzw. $a_{2ht(jung)}$) und keine Rekognition. Analog verfährt man im Modell für neue Items (vgl. Abbildung 4.4).

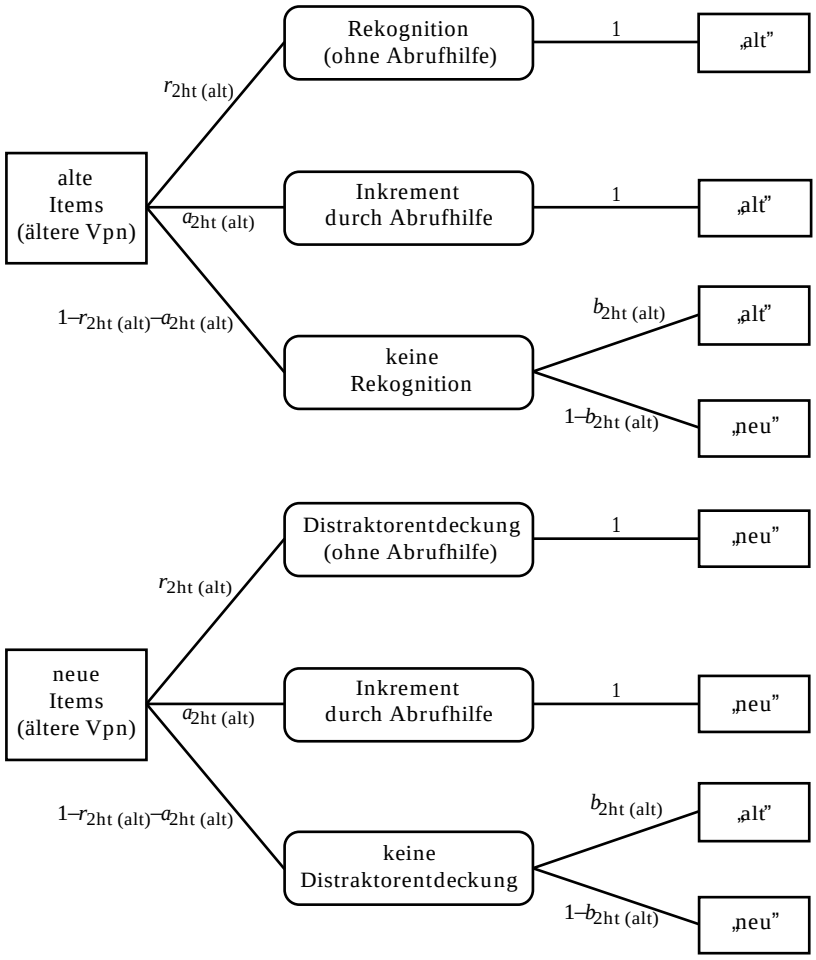

Abbildung 4.4:
Ein erweitertes 2HT-Modell zur Prüfung der Interaktionshypothese von Probandenalter und Abrufhilfen bezüglich des 2HT-Rekognitionsparameters r_{2ht}. Dargestellt ist nur das erweiterte Teilmodell für ältere Probanden in der Bedingung mit Abrufhilfe. Das entsprechende Teilmodell für jüngere Probanden ist analog, wobei zur Prüfung der Nullhypothese einer fehlenden Interaktion $a_{2ht(alt)} = a_{2ht(jung)}$ zu setzen ist. Für die Bedingung ohne Abrufhilfen werden jeweils Standard-2HT-Modelle für ältere und junge Probanden zugrunde gelegt.

Ein solches Modell ist saturiert und identifizierbar, wenn die Parameter $r_{2ht(alt)}$ bzw. $r_{2ht(jung)}$ durch Gleichheitsrestriktionen an die korrespondierenden Rekognitionsparameter für die Bedingung ohne Abrufhilfe gebunden werden. Wie man leicht sieht, ist das erweiterte Modell einem gewöhnlichen 2HT-Modell

mit Rekognitionsparameter $r_{2ht(alt)} + a_{2ht(alt)}$ bzw. $r_{2ht(jung)} + a_{2ht(jung)}$ äquivalent. Um dies zu zeigen, muß man nur jeweils die oberen beiden Äste, die beide in die gleiche Kategorie „alt" (bei alten Items) bzw. „neu" (bei neuen Items) einmünden, zu einem Ast zusammenfassen. Hierbei addieren sich die Astwahrscheinlichkeiten und man erhält das gewünschte Resultat. Will man nun die Nullhypothese einer fehlenden Interaktion prüfen, so setzt man einfach $a_{2ht(alt)} = a_{2ht(jung)}$. Diese Parametergleichheit bedeutet ja gerade, daß der Anstieg in der Rekognitionsleistung durch Bereitstellung von Abrufhilfen bei alten Probanden genauso stark ausfällt wie bei jüngeren.

Man beachte, daß die Überprüfung der Additivitätsnullhypothese nur mit Hilfe eines restringierten nichtbinären MVB-Modells gelingt. Eine Überführung des Modells aus Abbildung 4.4 in ein binäres MVB-Modell ist nicht möglich, da zwecks Prüfung der Additivitätsnullhypothese jeweils *alle* Parameter nichtbinärer Parametergruppen durch Gleichheitsrestriktionen an andere Parameter des Gesamtmodells zu binden sind. Die Zusammenfassung von Ästen, die für die Überführung in ein binäres Modell erforderlich wäre, würde die entsprechenden Gleichheitsrestriktionsmuster zerstören (vgl. Abschnitt 3.5).

Nullhypothesen des Typs (5) über die Rangordnung bestimmter Parameter nehmen eine gewisse Sonderstellung ein. Bei den oben behandelten Hypothesen (1) bis (4) hat das Ausgangsmodell immer einen S-dimensionalen Parameterraum $\mathbf{\Omega}_S$, und das restringierte Nullhypothesenmodell einen S'-dimensionalen Parameterraum $\mathbf{\Omega}_{S'}$, wobei $S' < S$ gilt. Die zugeordneten Parameterräume stehen also nicht nur in einer Teilmengenrelation ($\mathbf{\Omega}_{S'} \subset \mathbf{\Omega}_S$), sondern sie sind auch von unterschiedlicher Dimensionalität. Graphisch veranschaulicht entspricht $\mathbf{\Omega}_{S'}$ einem „infinitesimally thin film" (Bamber & van Santen, 1985, S. 445) in $\mathbf{\Omega}_{S'}$. Gleiches gilt für die Bilder dieser Parameterräume unter den Abbildungen f bzw. f', die durch die Modellgleichungen des Ausgangsmodells und des restringierten Modells definiert werden ($f(\mathbf{\Omega}_{S'}) \subset f'(\mathbf{\Omega}_S)$). Modelle mit diesen Eigenschaften bilden eine typische hierarchische Modellfamilie, wobei die restringierten Modelle dimensionsreduzierte Spezialfälle oder „Submodelle" („Untermodelle") des allgemeineren Modells sind. Man sagt in solchen Fällen auch, das restringierte Modell sei in das Ausgangsmodell „geschachtelt" (oder „nested"; vgl. Bishop et al., 1975, Abschnitt 14.9.6).

Mit einer Rangordnungsnullhypothese des Typs (5) ist zwar ebenfalls ein gegenüber dem Ausgangsmodell reduzierter Parameterraum und auch ein entsprechend reduziertes Bild des Parameterraumes verbunden, jedoch keine Dimensionsreduktion, d.h. es gilt $S = S'$, die Parameterzahl bleibt also gegenüber dem Ausgangsmodell unverändert. $\mathbf{\Omega}_{S'}$ entspricht folglich keinem infinitesimal dünnen Film in $\mathbf{\Omega}_S$, sondern einem erheblichen Teil von $\mathbf{\Omega}_S$, so daß bei Zufallsauswahl eines Parameters aus $\mathbf{\Omega}_S$ die Wahrscheinlichkeit größer als null ist, daß die-

ser Zufallswert auch in $\Omega_{S'}$ liegt (vgl. die Unterscheidung von qualitativer und quantitativer Testbarkeit bei Bamber & van Santen, 1985; 2000).

Die Konstruktion restringierter Modelle, welche Hypothesen des Typs (5) entsprechen, kann auf unterschiedliche Weise erfolgen. Soll beispielsweise die Hypothese $r_{2ht(einfach)} \leq r_{2ht(mittel)} \leq r_{2ht(intensiv)}$ bezüglich der Rekognitionsparameter des 2HT-Modells getestet werden, so kann man u.a. das erweiterte 2HT-Modell aus Abbildung 4.4 heranziehen. Zunächst wird ein Standard-2HT-Modell mit Rekognitionsparameter $r_{2ht(einfach)}$ für die Versuchsbedingung mit einfachem Gedächtnistraining formuliert. Für die Bedingung mit mittelintensivem Gedächtnistraining wird ein erweitertes 2HT-Modell gewählt, wobei der erste Ast erneut den Rekognitionsparameter $r_{2ht(einfach)}$ und der zweite Ast das additive Inkrement $a_{2ht(mittel)}$ relativ zum einfachen Gedächtnistraining repräsentiert. Die Summe dieser beiden Parameter entspricht also dem Rekognitionsparameter $r_{2ht(mittel)}$ des mittelintensiven Trainings. Da $a_{2ht(mittel)} \geq 0$ gelten muß, erzwingt diese Form der Modellierung zunächst die Rangordnung $r_{2ht(einfach)} \leq r_{2ht(mittel)}$. Um auch die zweite Ungleichheitsrestriktion in das Modell zu integrieren, benötigt man ein nochmals erweitertes Modell für die Bedingung mit intensivem Gedächtnistraining. Dieses umfaßt vier Äste auf der ersten Ebene, nämlich einen Ast für die Rekognitionswahrscheinlichkeit $r_{2ht(einfach)}$ im einfachen Training, einen für das Inkrement $a_{2ht(mittel)}$ durch das mittelintensive Training, einen weiteren für das Inkrement $a_{2ht(intensiv)}$) durch das intensive Training und schließlich einen letzten Ast für den Fall „keine Rekognition". Die Summe der ersten drei Astwahrscheinlichkeiten entspricht dem Rekognitionsparameter $r_{2ht(intensiv)}$ für das intensive Training. Wiederum sind $r_{2ht(einfach)}$ und $a_{2ht(mittel)}$ durch Gleichheitsrestriktionen an die entsprechenden Parameter der anderen Ereignisbäume zu binden. Da auch hier $a_{2ht(intensiv)} \geq 0$ gelten muß, repräsentiert das Gesamtmodell die gewünschte Rangordnung $r_{2ht(einfach)} \leq r_{2ht(mittel)} \leq r_{2ht(intensiv)}$. Man beachte, daß sich durch diese Ordnungsrestriktionen die Gesamtparameterzahl gegenüber dem unrestringierten 3-Gruppen-Ausgangsmodell (mit drei Rekognitions- und drei Antworttendenzparametern) nicht geändert hat. Auch das restringierte Modell umfaßt sechs Parameter; die drei Rekognitionsparameter des unrestringierten Modells werden lediglich in einen Rekognitionsparameter $r_{2ht(einfach)}$ und zwei additive Inkrementparameter $a_{2ht(mittel)}$ und $a_{2ht(intensiv)}$ dekomponiert. Die Rekognitionsparameter $r_{2ht(mittel)}$ und $r_{2ht(intensiv)}$ sind nicht mehr direkt als Parameter im restringierten Modell repräsentiert, lassen sich jedoch aus $r_{2ht(einfach)}$ und den additiven Inkrementparametern leicht errechnen.

Die geschilderte Methode der Umsetzung von Ungleichheitsrestriktionen setzt die Benutzung nichtbinärer multinomialer Verarbeitungsbaummodelle voraus. Man kann das gleiche Ziel aber auch im Rahmen binärer MVB-Modelle erreichen, und zwar auch hier wieder auf unterschiedliche Weise (z.B. Batchelder,

1999). Will man zunächst die Restriktion $r_{2ht(einfach)} \leq r_{2ht(mittel)}$ realisieren, so kann man ein Standard-2HT-Modell für das leichte Gedächtnistraining und ein gemäß Abbildung 4.5 (oben) modifiziertes Modell für das mittelintensive Training vorsehen. Hierbei wird der Rekognitionsparameter $r_{2ht(mittel)}$ erneut durch die Summe der oberen beiden Astwahrscheinlichkeiten repräsentiert, d.h. $r_{2ht(mittel)} = r_{2ht(einfach)} + (1-r_{2ht(einfach)})\cdot x_{2ht(mittel)}$, wobei $x_{2ht(mittel)}$ ein multiplikativer Inkrementparameter ist. Das Modell für neue Items ist in Abbildung 4.5 nicht dargestellt, kann jedoch völlig analog gewonnen werden, indem man einfach die oberen beiden Äste in die Beobachtungskategorie „neu" einmünden läßt. Da $0 \leq x_{2ht(mittel)} \leq 1$ gilt, wird durch die Repräsentation in Abbildung 4.5 (oben) die Ordnungsrestriktion $r_{2ht(einfach)} \leq r_{2ht(mittel)}$ erzwungen.

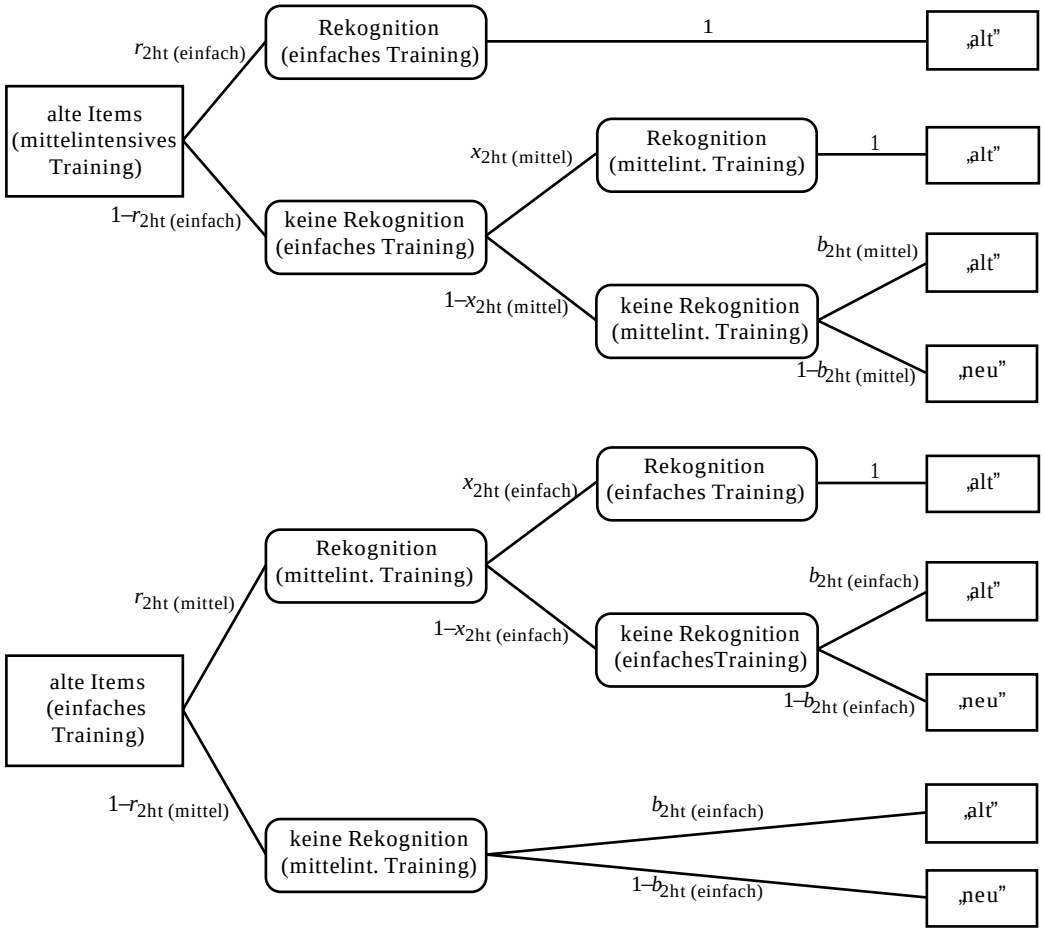

Abbildung 4.5:
Modifizierte 2HT-Modelle zur Prüfung von Ordnungsrestriktionen bezüglich des 2HT-Rekognitionsparameters r_{2ht}. Dargestellt sind nur die Teilmodelle für alte Items. Realisiert wird die Restriktion, daß der Rekognitionsparameter nicht kleiner (oberes Modell) bzw. nicht größer als in einer Vergleichsbedingung ausfallen darf (unteres Modell). Das obere Modell entspricht einem 2HT-Modell mit Rekognitionsparameter $r_{2ht(mittel)} = r_{2ht(einfach)} + (1-r_{2ht(einfach)})\cdot x_{2ht(mittel)}$, wobei $r_{2ht(einfach)}$ der Rekognitionsparameter der Vergleichsbedingung ist. Das untere Modell ist ein 2HT-Modell mit Rekognitionsparameter $r_{2ht(einfach)} = r_{2ht(mittel)} \cdot x_{2ht(einfach)}$, wobei $r_{2ht(mittel)}$ der Rekognitionsparameter der Vergleichsbedingung ist. Die dargestellten Teilmodelle für alte Items müssen jeweils analog auf neue Items übertragen und mit Standard-2HT-Modellen für die Vergleichsbedingung kontrastiert werden, um vollständige Modelle zu erhalten.

Ein alternativer Weg besteht darin, ein Standard-2HT-Modell für das mittelintensive Training und ein gemäß Abbildung 4.5 (unten) modifiziertes Modell für das einfache Training vorzusehen. Dieses Modell entspricht einem 2HT-Modell mit Rekognitionsparameter $r_{2ht(einfach)} = r_{2ht(mittel)} \cdot x_{2ht(einfach)}$, wobei $r_{2ht(mittel)}$ der Rekognitionsparameter für das mittelintensive Training und $x_{2ht(einfach)}$ ein multiplikativer Dekrementparameter ist. Da $0 \leq x_{2ht(einfach)} \leq 1$ gilt, wird auch durch dieses Teilmodell die Ungleichheitsrestriktion $r_{2ht(einfach)} \leq r_{2ht(mittel)}$ erzwungen.

Die Erweiterung der dargestellten Ansätze auf mehrere Ungleichheitsrestriktionen folgt immer dem gleichen Prinzip und erzeugt — von den zunehmend komplexer werdenden Ereignisbaumstrukturen abgesehen — keine neuen Probleme. Die Erweiterung braucht deshalb nicht separat besprochen werden. Im Prinzip ist es gleichgültig, welche Methode der Realisierung von Ungleichheitsrestriktionen man präferiert. Selbstverständlich sind alle drei dargestellten Ansätze äquivalent in dem Sinne, daß bei Anwendung auf empirische Daten immer die gleichen Parameterschätzungen resultieren, sofern man die angegebenen Formeln, welche die Beziehungen zwischen den verschiedenen Modellparametern spezifizieren, auf die Schätzer anwendet.

Die bislang behandelten Ungleichheitsrestriktionen betreffen Ordnungsrelationen zwischen Modellparametern in verschiedenen Ereignisbäumen eines Mehr-Gruppen-Gesamtmodells. Analoge Techniken lassen sich anwenden, wenn Ordnungshypothesen zu verschiedenen Parametern innerhalb eines Ereignisbaums zu prüfen sind, sofern sie sich ebenfalls in der Form „Parameter 2 = Parameter 1 + positives Inkrement" oder „Parameter 2 = Parameter 1 × Dekrementfaktor" schreiben lassen. Gelegentlich kann es jedoch vorkommen, daß Ordnungshypothesen *alle* Parameter einer Parametergruppe in einem MVB-Modell betreffen. Im Latent-Class-Modell zum Levine-Experiment (vgl. Abschnitt 2.5) sind beispielsweise acht latente Klassenwahrscheinlichkeiten $p(h_i)$, $i = 1, ..., 8$, für die Hypothesen des Hypothesenraums vorgesehen. Man stelle sich nun vor, daß der Vl vor dem Versuch das von den Vpn zu identifizierende Konzept aus dem Raum möglicher Konzepte so auswählt, daß die Konzepte nach ihrer Wahrscheinlichkeit geordnet werden können. Diese Wahrscheinlichkeiten werden den Vpn vor dem Versuch bekannt gemacht. Folgen die Vpn dem Prinzip der Wahrscheinlichkeitsangleichung (Gallistel, 1990, Kapitel 11) — was auf dem Hintergrund vorliegender Befunde plausibel ist —, so sollten die Klassenwahrscheinlichkeiten für die acht Hypothesen ebenfalls in der vorgegebenen Rangordnung stehen, beispielsweise $p(h_j) \leq p(h_{j+1})$, $j = 1, ...,7$.

Die Prüfung dieser speziellen Ordnungshypothese gelingt mit den o.g. Mitteln nicht. Dies liegt daran, daß nicht nur jeder Modellparameter der Parametergruppe — $p(h_1)$ bis $p(h_7)$ —, sondern zusätzlich auch deren Komplement $p(h_8)$ von

der Restriktion betroffen ist: Das Komplement muß größer als alle anderen Parameter der Gruppe sein, wenn die Ordnungshypothese stimmt. Auch in der Literatur zur Latent-Class-Analyse wurde dieses Problem m.W. bislang ausgeklammert; publizierte Ansätze behandeln lediglich Ordnungsrestriktion zu latenten Itemparametern, welche die lokalen Verteilungen innerhalb der latenten Klassen beschreiben, nicht jedoch die Klassengrößenparameter selbst (z.B. Croon, 1990; Formann, 1992).

Trotz dieser Probleme lassen sich Fragestellungen der geschilderten Art im Rahmen restringierter multinomialer MVB-Modelle — prinzipiell sogar im Rahmen binärer MVB-Modelle — adäquat behandeln. Abbildung 4.6 illustriert dies für den einfachsten nichttrivialen Fall einer Parametergruppe mit zwei freien Parameter p_1 und p_2 und einem Komplementparameter $p_3 = 1 - p_1 - p_2$. Der binäre Fall mit nur einem freien Parameter p_1 ist deshalb trivial, weil die Hypothese $p_1 \leq 1\text{-}\, p_1$ der Hypothese $p_1 \leq .5$ äquivalent ist, so daß man sie über das bereits besprochene Restriktionsmuster „Parameter 2 = Parameter 1 × Dekrementfaktor" testen kann. Man muß nur Parameter 2 mit p_1 identifizieren und Parameter 1 auf .5 fixieren.

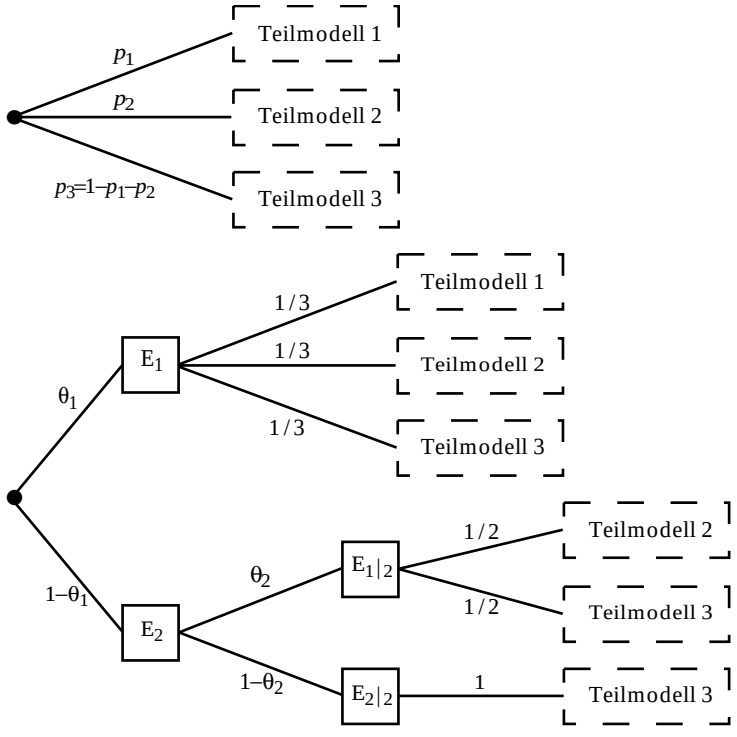

Abbildung 4.6:
Überführung der Ordnungshypothese $p_1 \leq p_2 \leq p_3$ zu den Parametern einer Parametergruppe (oben) in ein restringiertes MVB-Modell (unten). Das restringierte Modell ist aus Gründen der Platzersparnis als nichtbinäres Modell dargestellt, kann aber problemlos in ein äquivalentes binäres Modell überführt werden.

Aus dem in Abbildung 4.6 unten dargestellten restringierten Modell erhält man die drei interessierenden Parameter p_1, p_2 und p_3 des oben dargestellten Modells durch Einsetzen der Hilfsparameter θ_1 und θ_2 in die Gleichungen

$$\begin{aligned} p_1 &= \tfrac{1}{3} \cdot \theta_1 \\ p_2 &= \tfrac{1}{3} \cdot \theta_1 + \tfrac{1}{2} \cdot (1-\theta_1) \cdot \theta_2 = p_1 + \tfrac{1}{2} \cdot (1-\theta_1) \cdot \theta_2 \\ p_3 &= \tfrac{1}{3} \cdot \theta_1 + \tfrac{1}{2} \cdot (1-\theta_1) \cdot \theta_2 + (1-\theta_1) \cdot (1-\theta_2) = p_2 + (1-\theta_1) \cdot (1-\theta_2) \,. \end{aligned} \qquad (4.3.2.1\text{-}1)$$

Man kann sich leicht überzeugen, daß die interessierenden Parameter die gewünschte Ordnungsrelation für alle möglichen Werte der Hilfsparameter einhalten. Die Übertragung des Restriktionsprinzips auf Parametergruppen mit mehr als zwei freien Parametern liegt auf der Hand und braucht deshalb nicht separat demonstriert werden. Bei drei statt zwei freien Parametern folgen dem Ereignis E_1 beispielsweise nicht drei, sondern vier Teilbäume mit fixierten Wahrscheinlichkeiten von jeweils 1/4. Dem Ereignis E_2 würde ein Teilmodell folgen, das komplett analog zu dem in Abbildung 4.6 (unten) dargestellten Modell ist.

Ein interessantes Nebenprodukt des vorgestellten Restriktionsschemas ist dessen Bedeutung für die Identifizierbarkeitsfrage bei Latent-Class-Modellen. LCA-Modelle sind im allgemeinen niemals global, sondern allenfalls lokal identifizierbar. Dies liegt daran, daß es bei komplett freien Klassenwahrscheinlichkeiten oft möglich ist, durch Permutation der Klassenindizes der Modellparameter alternative Parametervektoren zu erzeugen, die exakt die gleichen Kategoriewahrscheinlichkeiten implizieren (vgl. Abschnitt 4.1). Diese Mehrdeutigkeit kann eliminiert werden, indem man den Klassen eine Ordnung aufzwingt, beispielsweise eine an der Größe der Klassenwahrscheinlichkeiten orientierte Ordnung. Dies kann mit dem dargestellten Ordnungsrestriktionsschema erreicht werden. Natürlich funktioniert dieses Prinzip nicht sicher, sondern nur „fast sicher". Wenn nämlich mindestens zwei Klassenwahrscheinlichkeiten exakt gleich groß sind, entstehen trotz Ordnungsrestriktionen Mehrdeutigkeiten. Für die Anwendungspraxis kann dieser zwar denkbare, aber extrem unwahrscheinliche Fall jedoch getrost ignoriert werden. Lokal identifizierbare Latent-Class-Modelle werden durch Ordnungsrestriktionen auf den Klassenwahrscheinlichkeiten global identifizierbar, wenn man die geschilderten Sonderfälle aus dem Raum zulässiger Parameterwerte ausschließt.

Abschließend sei noch einmal betont, daß dieser Abschnitt nicht als Versuch einer erschöpfenden Darstellung denkbarer Parameter- und Submodellhypothesen mißverstanden werden darf. Es ging lediglich um die in Anwendungskontexten häufigsten Hypothesen und deren Überführung in geeignet restringierte Modelle. Binäre und nichtbinäre MVB-Modelle bieten weitaus mehr Möglichkeiten für Parameterrestriktionen als hier angesprochen wurden. Erweitert man die Perspektive darüber hinaus auf parametrisierte multinomiale Modelle, die keine MVB-

Modelle sind, so erscheint der Spielraum für die Gewinnung restringierter Modelle nahezu unerschöpflich.

4.3.2.2 Prüfung restringierter Modelle

Wenn ein geeignet restringiertes multinomiales Modell formuliert und dessen Parameter durch Anwendung auf empirische Daten geschätzt wurden, gilt das Hauptinteresse natürlich der statistischen Prüfung der betreffenden Nullhypothese. Ein naheliegender Gedanke ist es, ebenso wie in Abschnitt 4.3.1 besprochen zu verfahren und das restringierte Modell einfach über eine Minimum-PD$^{\lambda}$-Modellanpassungsstatistik anhand einer zentralen χ^2-Verteilung mit $df = M-S'$ zu evaluieren. Man behandelt das restringierte Modell also wie das Ausgangsmodell und fragt global nach dessen Adäquatheit für gegebene Daten.

Was ist von einem solchen Vorgehen zu halten? Statistisch gibt es keine prinzipiellen Einwände, sofern die in Abschnitt 4.3.1 erwähnten Regularitätsbedingungen von Birch (1964) beachtet werden. Allerdings muß man sich darüber im klaren sein, daß auf diese Weise nicht die interessierende Parameter- oder Submodellhypothese isoliert geprüft wird, sondern vielmehr die *Konjunktion* zweier Nullhypothesen, nämlich die Modellgeltungshypothese zum Ausgangsmodell in Verbindung mit der interessierenden Parameter- oder Submodellnullhypothese. Wenn das Ausgangsmodell schon sehr schlecht auf vorliegende Daten paßt, wird die Prüfstatistik also zwangsläufig sehr groß. Dies mag durchaus erwünscht sein, denn eine Parameter- oder Submodellhypothese ist nur dann von Interesse, wenn das Ausgangsmodell zu gegebenen Daten paßt. Umgekehrt kann jedoch der Fall eintreten, daß das Ausgangsmodell sehr gut zu den Daten paßt, so daß die Prüfstatistik für das Ausgangsmodell nahe bei null liegt. In diesem Fall fällt auch die globale Prüfstatistik für das restringierte Modell tendenziell niedrig aus. Anders ausgedrückt: Die zu prüfende Parameter- oder Submodellhypothese im (sehr gut passenden) Ausgangsmodell muß schon gravierend verletzt sein, damit die globale Prüfstatistik des restringierten Modells die Signifikanzgrenze erreicht. Dies gilt um so mehr, als die Freiheitsgradzahl $M-S'$ beim Test des restringierten Modells im Regelfall größer als beim Test des Ausgangsmodells ist. Praktisch läuft dies auf ein Teststärkeproblem hinaus: Paßt das Ausgangsmodell sehr gut auf die Daten, hat der globale Modellanpassungstest für das restringierte Modell kaum eine Chance, vorliegende Verletzungen der zu prüfenden Parameter- oder Submodellnullhypothese aufzudecken. Wir werden auf diesen Punkt im nächsten Abschnitt zurückkommen.

Glücklicherweise gibt es bei hierarchischen Modellen mit S' Parametern im restringierten Modell und $S > S'$ Parametern im Ausgangsmodell eine deutlich teststärkere Alternative zu dem geschilderten Vorgehen: den Differenzen-χ^2-Test.

Dieser Test basiert auf den Log-Likelihood-Quotienten-Statistiken (bzw. den Minimum-PD$^{\lambda=0}$-Statistiken) zum Ausgangsmodell ($G^2(M_1)$) und zum restringierten Modell ($G^2(M_0)$). Die Teststatistik des Differenzentests ergibt sich einfach durch Differenzbildung:

$$\Delta G^2(M_0|M_1) := G^2(M_0) - G^2(M_1) \qquad (4.3.2.2\text{-}1)$$

Da M_0 ein parameterreduziertes Submodell von M_1 ist, kann es nicht besser zu gegebenen Daten passen als M_1. Es muß also zwangsläufig $G^2(M_0) \geq G^2(M_1)$ und damit $\Delta G^2(M_0 \mid M_1) \geq 0$ gelten. Können die Regularitätsbedingungen von Birch (1964) vorausgesetzt werden, so läßt sich ferner zeigen, daß bei Gültigkeit des restringierten Modells M_0 (und folglich auch bei Gültigkeit des Ausgangsmodells M_1) nicht nur $G^2(M_0)$ zentral $\chi^2_{(M-S')}$- und $G^2(M_1)$ zentral $\chi^2_{(M-S)}$-verteilt ist — dies folgt bereits aus den Ausführungen in Abschnitt 4.3.1 —, sondern daß zusätzlich auch noch $\Delta G^2(M_0 \mid M_1)$ zentral χ^2-verteilt ist mit $df = S - S'$ (Bishop et al., 1975, Kapitel 4).

Der Differenzen-χ^2-Test (4.3.2.2-1) und der gewöhnliche, auf G^2 bzw. PD$^{(\lambda=0)}$ basierende globale Modellanpassungstest (4.2.1-2) folgen demselben Grundprinzip, auch wenn dies nicht auf den ersten Blick ersichtlich ist: Sie sind Spezialfälle des allgemeinen Log-Likelihood-Quotienten-Tests. Die Teststatistiken lassen sich in beiden Fällen als transformierte logarithmierte Quotienten zweier Likelihoods darstellen, nämlich der Likelihood unter einem allgemeinen Modell M_a und einem hierarchisch darin geschachtelten Modell M_u (Wilks, 1938):

$$G^2(M_u|M_a) := -2 \cdot \ln(L(M_u) / L(M_a)) \,. \qquad (4.3.2.2\text{-}2)$$

Bereits Wilks (1938) hat gezeigt, daß $\Delta G^2(M_u \mid M_a)$ bei Gültigkeit von M_u unter sehr allgemeinen Randbedingungen — nicht beschränkt auf multinomiale Modelle — zentral χ^2-verteilt ist, wobei die Freiheitsgradzahl gerade der Differenz freier Parameter in M_a und M_u entspricht. Die zuvor besprochene globale Log-Likelihood-Quotienten-Anpassungsstatistik (4.2.1-2) zu einem multinomialen Modell mit S Parametern ergibt sich als Spezialfall der allgemeinen Formel (4.3.2.2-2), wenn man einerseits die spezielle Form der Likelihood-Funktion (4.2.1-1) bei parametrisierten multinomialen Modellen berücksichtigt und andererseits M_a als saturiertes Modell (mit M zu schätzenden Parametern) sowie M_u als zu prüfendes Modell (mit S zu schätzenden Parametern) wählt. In Gleichung (4.3.2.2-1) sind sowohl $G^2(M_0)$ als auch $G^2(M_1)$ globale Log-Likelihood-Quotienten-Statistiken in diesem Sinne, d.h. sie vergleichen jeweils die Likelihood eines Modells (M_0 bzw. M_1) mit der des saturierten Modells M_a. Folglich gilt für die Differenzen-χ^2-Statistik gemäß Definition (4.3.2.2-1):

$$\begin{aligned}\Delta G^2(M_0|M_1) &:= G^2(M_0) - G^2(M_1)\\ &= (-2\cdot\ln(L(M_0)\,/\,L(M_a)) - (-2\cdot\ln(L(M_1)\,/\,L(M_a))\\ &= -2\cdot\ln(L(M_0)\,/\,L(M_1))\,.\end{aligned} \qquad (4.3.2.2\text{-}3)$$

Die Differenzenstatistik $\Delta G^2(M_0 \mid M_1)$ ist also ebenfalls ein Spezialfall der allgemeinen Log-Likelihood-Quotienten-Statistik (4.3.2.2-2), wobei M_1 als allgemeineres Modell und M_0 als hierarchisch darin geschachteltes Modell fungiert. Man könnte $\Delta G^2(M_0 \mid M_1)$ — obwohl umständlicher — auch so berechnen, daß man eine modifizierte globale Log-Likelihood-Quotienten-Statistik zum restringierten Modell M_0 bestimmt, bei der statt der beobachteten Häufigkeiten $y_{k,j}$ die unter M_1 erwarteten Häufigkeiten in Gleichung (4.2.1-2) eingesetzt werden. Eine solchermaßen modifizierte Statistik mißt dann die $PD^{\lambda=0}$-Divergenz zwischen den erwarten Häufigkeiten unter M_0 und M_1 und nicht — wie die globale Modellanpassungsstatistik $G^2(M_0)$ — die Divergenz zwischen M_0 und dem saturierten Modell, dessen erwartete Häufigkeiten gerade den beobachteten Häufigkeiten entsprechen.

Zusammenfassend kann festgehalten werden, daß sich die globale Log-Likelihood-Quotienten-Teststatistik $G^2(M_0)$ für ein restringiertes Modell M_0 additiv dekomponieren läßt in zwei unabhängige, bei Gültigkeit von M_0 zentral χ^2-verteilte Prüfstatistiken $G^2(M_1)$ und $\Delta G^2(M_0 \mid M_1)$. Ist das restringierte Modell verletzt, so ist $G^2(M_1)$ sensitiv für Modellverletzungen, die bereits das Ausgangsmodell betreffen, $\Delta G^2(M_0 \mid M_1)$ dagegen für Verletzungen der Parameter- oder Submodellhypothese, die vom Ausgangsmodell zum Submodell führt. Im nächsten Abschnitt wird gezeigt, daß ein auf $\Delta G^2(M_0 \mid M_1)$ basierender Test bei Ungültigkeit von M_0 und Gültigkeit von M_1 im Regelfall erheblich teststärker als ein auf $G^2(M_0)$ basierender globaler Modellgeltungstest ist. Wenn man also — z.B. aufgrund eines vorgeschalteten Tests von M_1 — das Ausgangsmodell bereits angenommen hat, ist eine isolierte Prüfung von Parameter- oder Submodellhypothesen über den Differenzen-χ^2-Test $\Delta G^2(M_0 \mid M_1)$ klar indiziert.

In Anbetracht der Tatsache, daß bei Modellgültigkeit alle globalen Minimum-PD^λ-Statistiken unabhängig von λ die gleiche asymptotische χ^2-Verteilung haben, liegt die Frage nahe, ob dieses Ergebnis auch auf die zugehörigen Differenzenstatistiken $\Delta PD^\lambda(M_0 \mid M_1) := PD^\lambda(M_0) - PD^\lambda(M_1)$ für $\lambda \neq 0$ übertragbar ist. Interessanterweise ist das nicht der Fall (vgl. Bishop et al., 1975, S. 125-131). Natürlich lassen sich Differenzen zwischen Minimum-PD^λ-Statistiken auch im Falle $\lambda \neq 0$ prinzipiell berechnen; sie unterscheiden sich bei großen Stichproben auch meistens nicht wesentlich von $\Delta G^2(M_0 \mid M_1)$. Allerdings besteht im Falle $\lambda \neq 0$ anders als im Fall $\lambda = 0$ keine Äquivalenz mehr zwischen der Differenzenstatistik $\Delta PD^\lambda(M_0 \mid M_1) = PD^\lambda(M_0) - PD^\lambda(M_1)$ und der PD^λ-Modellvergleichsstatistik, die man erhält, wenn man in die allgemeine PD^λ-Formel (4.2.2.2-1) zu Modell M_0 an-

stelle der beobachteten Häufigkeiten die erwarteten Häufigkeiten unter M_1 einsetzt.

Als Fazit bleibt festzuhalten, daß über die Verteilungseigenschaften von PD^{λ}-Divergenzmaßen für den Vergleich hierarchisch ineinander geschachtelter Modelle gegenwärtig noch wenig bekannt ist, sofern $\lambda \neq 0$. Klar ist jedoch, daß der Differenzen-χ^2-Test nur im Falle $\lambda = 0$ zum korrekten Divergenzmaß führt, und hierfür ist die asymptotische Verteilung auch bekannt. Die Pearson-Statistik und andere Minimum-PD^{λ}-Statistiken lassen sich nicht in der Weise additiv dekomponieren, wie dies bei $G^2(M_0)$ der Fall ist. Dieser Sachverhalt und auch die Tatsache, daß sich die auf Log-Likelihood-Quotienten-Statistiken basierende Differenzenstatistik leicht berechnen läßt und relativ teststark ist (vgl. Abschnitt 4.3.3), führt zu der Empfehlung, sie zu verwenden, wann immer Parameter- oder Submodellnullhypothesen zu einem bereits akzeptierten Ausgangsmodell zu testen sind. Sofern die Regularitätsbedingungen von Birch (1964) erfüllt sind und der Stichprobenumfang hinreichend groß ist, kann die Differenzenstatistik anhand der zentralen $\chi^2_{(S-S')}$-Verteilung auf Signifikanz beurteilt werden.

Ein Sonderfall ergibt sich, wenn das restringierte Modell ebenso viele Parameter wie das Ausgangsmodell hat (d.h. $S = S'$), beispielsweise beim Test von Ordnungsrestriktionen des Typs (5). In solchen Fällen kann die Differenzenstatistik nur dann größer als null werden, wenn mindestens ein Parameter des restringierten Modells gegen einen Grenzwert des Parameterraumes konvergiert. Die Regularitätsbedingungen von Birch (1964) sind dann nicht erfüllt, und die statistische Rechtfertigung für die Verwendung der χ^2-Verteilung als Referenzverteilung entfällt. Man kann in solchen Fällen einerseits versuchen, die korrekte asymptotische Verteilung bei grenzwertigen Parametern herzuleiten, die unter gewissen Regularitätsbedingungen eine Mischung von χ^2-Verteilung ist (Lee, 1987). Eine einfachere Möglichkeit ist die Übertragung des bereits erwähnten parametrischen Bootstrap-Verfahrens (vgl. Abschnitt 4.3.1) auf χ^2-Differenzenstatistiken (vgl. z.B. Klauer, Wegener & Ehrenberg, 1999). Es wird also eine Monte-Carlo-Studie zu dem Modell M_0 durchgeführt, das sich durch Einsetzen der Parameterschätzungen — darunter eventuell auch grenzwertige — in die Modellgleichungen ergibt. Aus der entsprechenden Population werden viele Stichproben des Umfangs der beobachteten Stichprobe generiert, und für jede dieser Stichproben wird die Differenzenstatistik bezüglich M_0 und M_1 berechnet. Die resultierende Bootstrap-Stichprobenverteilung dient als Referenzverteilung für die Bewertung der Differenzenstatistik zur beobachteten Stichprobe.[44] Die H_0 der Gültigkeit des restringierten Modells M_0

[44]In das Verfahren geht die Annahme ein, daß die Bootstrap-Stichprobenverteilung der Differenzenstatistik unabhängig von den Werten nichtgrenzwertiger Parameter ist. Es empfiehlt sich deshalb, verschiedene Bootstrap-Stichprobenverteilungen für eine kleine Auswahl von H_0-Parametervektoren zu generieren, um anschließend z.B. mittels eines Kolmogoroff-Smirnoff-Anpassungstests

wird verworfen, wenn die empirisch beobachtete Differenzenstatistik zu den oberen $\alpha \cdot 100\%$ der Referenzverteilung gehört, sonst wird sie beibehalten. Dieses Verfahren hält das α-Niveau auch dann ein, wenn grenzwertige Parameter vorliegen. Vorauszusetzen ist allerdings auch hier, daß das statistische Rahmenmodell erfüllt ist (unabhängige Ziehungen aus der gleichen Multinomialverteilung).

4.3.3 Teststärkekontrolle

Ist $\hat{\boldsymbol{\theta}}^{(\boldsymbol{\lambda})}$ der Minimum-PD$^{\lambda}$-Schätzwert für die Parameter eines Modells bei gegebenen Daten $\boldsymbol{Y} = \boldsymbol{y}$ und c_{α} der kritische χ^2-Wert der zugehörigen Modellanpassungsstatistik PD$^{\lambda}(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\lambda})}, \boldsymbol{y})$, d.h. der kleinste Wert, für den bei modellkonsistenten Kategoriewahrscheinlichkeiten $\boldsymbol{\pi} \in f(\boldsymbol{\Omega})$ asymptotisch

$$Pr(\mathrm{PD}^{\lambda}(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\lambda})}, \boldsymbol{y}) \geq c_{\alpha} | \boldsymbol{\pi} \in f(\boldsymbol{\Omega})) \leq \alpha \,, \qquad (4.3.3\text{-}1)$$

gilt, dann ist die Teststärke (synonym: Power, Effizienz) des auf PD$^{\lambda}$ basierenden Modellgeltungstests bei modellinkonsistenten Kategoriewahrscheinlichkeiten $\boldsymbol{\pi} \notin f(\boldsymbol{\Omega})$ definiert als

$$1 - ß(\boldsymbol{\pi}) = Pr(\mathrm{PD}^{\lambda}(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\lambda})}, \boldsymbol{y}) \geq c_{\alpha} | \boldsymbol{\pi} \notin f(\boldsymbol{\Omega})) \,. \qquad (4.3.3\text{-}2)$$

Modellgeltungstests parametrisierter multinomialer Modelle wie auch Tests von Parameter- und Submodellhypothesen müssen so durchgeführt werden, daß Verletzungen der zu prüfenden Nullhypothese mit großer Teststärke $1-\beta(\boldsymbol{\pi})$ aufgedeckt werden, sofern diese Verletzungen bedeutsam sind. Ist die Teststärke unzureichend, so bestehen lediglich stark eingeschränkte Möglichkeiten, substanzwissenschaftliche Theorien oder Hypothesen, die in das zu testende Modell einfließen oder durch das Modell repräsentiert werden, empirisch zu widerlegen. Die prinzipielle Widerlegungsmöglichkeit von Theorien ist jedoch ein Kernmerkmal vieler Wissenschaftsauffassungen, so daß das Problem der Teststärke- bzw. β-Fehler-Kontrolle nicht ignoriert werden kann (Bredenkamp, 1972; 1980; Erdfelder, 1984; Erdfelder & Bredenkamp, 1994; Hager, 1987; Herrmann, 1979; Lakatos, 1970). Um so erstaunlicher ist es, daß Teststärkekontrollen im Zusammenhang mit MVB-Modellen lange Zeit ausgeklammert bzw. zu den zukünftig zu lösenden statistischen Problemen gezählt wurden (vgl. Batchelder & Riefer, 1990, S. 561).

Selbstverständlich kommt es nicht darauf an, jede minimale Verletzung einer Modellgeltungshypothese aufzudecken, da gewisse Vorhersagefehler häufig schon bei der Modellkonstruktion einkalkuliert werden, um die Modelle hinreichend einfach und vor allem identifizierbar zu halten (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). Im Enkodierung-Abruf-Modell von Batchelder und

---

zu zeigen, daß sich die verschiedenen Bootstrap-Stichprobenverteilungen nicht signifikant unterscheiden (vgl. auch Fußnote 43).

Riefer (1980; 1986) (Enkodierung-Abruf-Modell, vgl. Abschnitt 3.5.1) wird beispielsweise die Möglichkeit ausgeschlossen, daß nichtgeclusterte, aber dennoch einzeln abrufbare Elemente eines Itempaares direkt nacheinander reproduziert werden ($C_{1,1}$-Ereignis). Natürlich kann dieses Ereignis aber dennoch zufällig einmal vorkommen, wenn auch — jedenfalls bei langen Itemlisten — mit geringer Wahrscheinlichkeit. Wenn man also den Modellgeltungstest zum Enkodierung-Abruf-Modell durchführt, möchte man eigentlich nur solche Modellverletzungen mit hoher Teststärke $1-\beta(\boldsymbol{\pi})$ aufdecken, die nicht auf zufällige $C_{1,1}$-Ereignisse der beschriebenen Art zurückzuführen sind. Es geht ausschließlich um „bedeutsame Modellverletzungen", die auf prinzipiellen, nicht a priori einkalkulierten Fehlspezifikationen im Modell beruhen und zu Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ deutlich außerhalb des modellkonsistenten Bereichs $f(\boldsymbol{\Omega})$ führen.

Die praktische Umsetzung dieser Art von Teststärkekontrolle setzt zunächst einmal voraus, daß man Maße für die Verletzung von Nullhypothesen festlegt — sogenannte Effektstärkemaße —, die genauer spezifizieren, was mit einer „bedeutsamen Verletzung der H_0" gemeint ist. Im nächsten Schritt ist dann die Beziehung zwischen diesen Maßen und der Teststärke zu analysieren.

Der historisch älteste Ansatz zur Teststärkebestimmung bei χ^2-Modellanpassungstests beruht auf Resultaten von Cochran (1952) und Mitra (1958) zu einfachen parametrisierten multinomialen Modelle (d.h. $K = 1$). Sie betrachteten Folgen sogenannter lokaler Alternativhypothesen $H_{1,N}$, die gegen H_0 konvergieren, wenn der Gesamtstichprobenumfang N gegen unendlich strebt. Sind die J Kategoriewahrscheinlichkeiten bei Modellgültigkeit durch den (im allgemeinen unbekannten) Vektor $\boldsymbol{p} = f(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), ..., p_J(\boldsymbol{\theta}))$ gegeben, die tatsächlichen Kategoriewahrscheinlichkeiten unter H_1 dagegen durch den Vektor $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$, und ist $\boldsymbol{\delta} = (\delta_1, ..., \delta_J)$ ein Vektor beliebiger reellwertiger Konstanten, die sich zu null aufsummieren, so kann man zeigen, daß jede Minimum-PD$^\lambda$-Statistik unter der Sequenz von Alternativhypothesen

$$H_{1,N}: \boldsymbol{\pi} = \boldsymbol{p} + \boldsymbol{\delta} / \sqrt{N} \qquad (4.3.3\text{-}3)$$

für $N \to \infty$ asymptotisch nonzentral $\chi(\gamma)^2_{(df)}$-verteilt ist mit Nonzentralitätsparameter

$$\gamma = \sum_{j=1}^{J} \frac{\delta_j^2}{p_j(\boldsymbol{\theta})} = \sum_{j=1}^{J} \frac{\left((\pi_j - p_j(\boldsymbol{\theta})) \cdot \sqrt{N}\right)^2}{p_j(\boldsymbol{\theta})} = N \cdot \sum_{j=1}^{J} \frac{\left(\pi_j - p_j(\boldsymbol{\theta})\right)^2}{p_j(\boldsymbol{\theta})} \; . \qquad (4.3.3\text{-}4)$$

Vorauszusetzen sind hierbei — wie bereits im Falle der zentralen χ^2-Verteilung unter H_0 — die Regularitätsbedingungen von Birch (vgl. Mitra, 1958; Read & Cressie, 1988, Anhang A8). Die Freiheitsgrade berechnen sich wie in den letzten beiden

Abschnitten besprochen ($df = M-S$ bei globalen Modellgeltungstests und $df = S-S'$ bei χ^2-Differenzentests basierend auf ΔG^2).

In Anbetracht dieses asymptotischen Resultats liegt es nahe, bei endlichen Stichprobenumfängen für jede Minimum-PD$^\lambda$-Statistik die Teststärke-Approximation

$$1-\beta(\boldsymbol{\pi}) \cong Pr(\chi(\gamma)^2_{(df)} \geq c_\alpha) \qquad (4.3.3\text{-}5)$$

zu verwenden, wobei der Nonzentralitätsparameter γ gemäß (4.3.3-4) zu bestimmen ist. Man beachte, daß dieser Nonzentralitätsparameter auch als Produkt von N und einem vom Stichprobenumfang unabhängigen Effektstärkeparameter geschrieben werden kann, welcher rein rechnerisch der Pearson-χ^2-Statistik entspricht, die allerdings nicht auf beobachtete und erwartete Stichprobenhäufigkeiten, sondern auf Kategoriewahrscheinlichkeiten unter H_0 und H_1 anzuwenden ist (s. rechte Seite von Gleichung (4.3.3-4)). Cohen (1988, Kapitel 7) hat vorgeschlagen, die Quadratwurzel dieses Parameters als Effektstärkemaß

$$w = \sqrt{\sum_{j=1}^{J} \frac{\left(\pi_j - p_j(\boldsymbol{\theta})\right)^2}{p_j(\boldsymbol{\theta})}} \qquad (4.3.3\text{-}6)$$

zu verwenden, so daß man für den Nonzentralitätsparameter auch kurz $\gamma = N \cdot w^2$ schreiben kann. Darüber hinaus hat Cohen (1988) Vorschläge für Effektstärkekonventionen unterbreitet. Demnach ist eine H_0-H_1-Diskrepanz der Größe $w = .10$ ein „kleiner Effekt", eine Diskrepanz von $w = .30$ ein „mittlerer Effekt" und schließlich eine Diskrepanz der Größe $w = .50$ ein „großer Effekt" (vgl. Cohen, 1988, S. 224-227). Soweit bislang überhaupt Teststärkebetrachtungen bei statistischen Tests im Rahmen multinomialer Modelle durchgeführt wurden — was eher der Ausnahmefall war — wurde der auf den Cohenschen Konventionen basierende Ansatz zugrunde gelegt: Stichprobenumfänge und Signifikanzniveaus wurden typischerweise so gewählt, daß „kleine" oder „moderate" Verletzungen der H_0 (d.h. $.10 \leq w \leq .20$) mit einer gemäß Gleichung (4.3.3-5) approximierten Teststärke $1-\beta(\boldsymbol{\pi})$ aufgedeckt werden, die in etwa $1-\alpha$ entspricht (z.B. Buchner, Steffens et al.1997; Buchner et al., 1995; Buchner et al., 1997; Erdfelder & Bredenkamp, 1998; Erdfelder & Buchner, 1998a; 1998b; Klauer & Batchelder, 1996; Klauer & Oberauer, 1995). Auf diese Weise wurde versucht, eine angemessene Balance zwischen den beiden Fehlerwahrscheinlichkeiten erster und zweiter Art auf möglichst niedrigem Niveau herzustellen („Kompromiß-Poweranalyse", vgl. Erdfelder, 1984; Erdfelder et al., 1996).

Allerdings läßt diese Form der Teststärkebetrachtung — so begrüßenswert sie vom Ansatz her ist — eine Reihe von Fragen offen, die in diesem Abschnitt näher untersucht werden sollen:

1) Läßt sich der Cohensche Ansatz überhaupt auf verbundene multinomiale Modelle ($K > 1$) sinnvoll anwenden? Wenn ja, wie hängt in diesem Fall der Nonzentralitätsparameter von den Effektstärken in den einzelnen Gruppen bzw. Versuchsbedingungen des verbundenen Modells ab?
2) Die Effektstärkekonventionen von Cohen (1988) können in verschiedenen Anwendungskontexten unterschiedliche Bedeutung haben und reichen deshalb als alleinige Grundlage für Teststärkebetrachtungen kaum aus. Um es mit Cohens (1988, S. 224) eigenen Worten auszudrücken: „Their use requires particular caution, since, apart from their possible inaptness in a given substantive context, what is subjectively the same degree of departure or degree of association may yield varying *w* (...), and conversely. The investigator is best advised to use the conventional definitions as a general frame of reference (...) and not to take them too literally". Wünschenswert wäre es, Modellverletzungen bzw. Abweichungen von Parameter- und Submodellnullhypothesen direkt als Funktion von Modellparametern unter H_1 ausdrücken zu können. Beim Modellgeltungstest für das Enkodierung-Abruf-Modell würde man z.B. gerne wissen, was eine „kleine" Modellverletzung ($w = .10$) bedeutet, wenn man sie als Differenz der kritischen Modellparameter *u* and *a* ausdrückt. Dazu muß geklärt werden, wie sich *w* bzw. der Nonzentralitätsparameter γ als Funktion des Modellparametervektors unter H_1 errechnen läßt.
3) Ist die Teststärkeapproximation (4.3.3-5) eigentlich eine gute Approximation für diejenigen Modellklassen und Stichprobenumfänge, die bei Hypothesentests zu parametrisierten multinomialen Modellen typischerweise zugrunde liegen?
4) Gibt es Alternativen zur Teststärkeapproximation (4.3.3-5), die insgesamt genauere Abschätzungen der Teststärke für praktisch relevante Stichprobenumfänge erlauben?
5) Sieht man einmal von den wesentlichen Determinanten der Teststärke ab — Grad der vorliegenden Modellverletzung (Effektstärke), Stichprobenumfang und α-Niveau —, gibt es eventuell weitere versuchsplanerische und datenanalytische Möglichkeiten, die Teststärke positiv zu beeinflussen?

Glücklicherweise gibt es auf alle diese Fragen Antworten, die sich recht leicht gewinnbringend einsetzen lassen.

4.3.3.1 Teststärke bei verbundenen multinomialen Modellen

Die Erweiterung des Ansatzes von Cochran, Mitra und Cohen auf verbundene multinomiale Modelle bereitet keine prinzipiellen Probleme (z.B. Meng & Chapman, 1966; Taneichi & Sekiya, 1995, S. 98/99). Man betrachtet die gleiche, asymptotisch gegen H_0 konvergierende Klasse von Alternativhypothesen (4.3.3-3), wobei

die Vektoren $\boldsymbol{p}$, $\boldsymbol{\pi}$ und $\boldsymbol{\delta}$ allerdings entsprechend auf ein K-Gruppen Design zu erweitern sind. Für den H_0-H_1-Diskrepanzvektor $\boldsymbol{\delta}$ bedeutet dies, daß sich die Vektorelemente $\delta_{k,j}$ pro Gruppe k zu null aufaddieren müssen ($\forall k$: $\Sigma_j\, \delta_{k,j} = 0$). N ist nach wie vor der Gesamtstichprobenumfang über die K Gruppen hinweg und $\tau_k := N_k / N$ ist der (konstante) Anteil des Gesamtstichprobenumfangs, der in Gruppe k fällt. Gelten die Regularitätsbedingungen von Birch (1964), ist die asymptotische Verteilung jeder Minimum-PD$^\lambda$-Statistik für $N \to \infty$ eine nonzentrale $\chi(\gamma)^2_{(df)}$-Verteilung mit Nonzentralitätsparameter

$$\gamma = \sum_{k=1}^{K}\sum_{j=1}^{J_k}\frac{(\delta_{k,j})^2}{p_{k,j}(\boldsymbol{\theta})} = \sum_{k=1}^{K} N_k \cdot \sum_{j=1}^{J_k}\frac{\left(\pi_{k,j} - p_{k,j}(\boldsymbol{\theta})\right)^2}{p_{k,j}(\boldsymbol{\theta})} = \\ = N \cdot \sum_{k=1}^{K}\tau_k \cdot \sum_{j=1}^{J_k}\frac{\left(\pi_{k,j} - p_{k,j}(\boldsymbol{\theta})\right)^2}{p_{k,j}(\boldsymbol{\theta})}\ . \qquad (4.3.3.1\text{-}1)$$

Definiert man also

$$w_k = \sqrt{\sum_{j=1}^{J_k}\frac{\left(\pi_{k,j} - p_{k,j}(\boldsymbol{\theta})\right)^2}{p_{k,j}(\boldsymbol{\theta})}} \qquad (4.3.3.1\text{-}2)$$

als spezifisches Effektstärke- bzw. Modellverletzungsmaß für die Gruppe k und

$$w = \sqrt{\sum_{k=1}^{K}\tau_k \cdot {w_k}^2} \qquad (4.3.3.1\text{-}3)$$

als gruppenübergreifendes Gesamteffektmaß, dann kann man den Nonzentralitätsparameter auch bei verbundenen multinomialen Modellen in der Form $\gamma = N \cdot w^2$ schreiben (vgl. Erdfelder & Bredenkamp, 1998, S. 937, Fußnote 3).

Die Herleitung dieses Resultats ergibt sich aus der Tatsache, daß die PD$^\lambda$-Anpassungsstatistik für ein verbundenes multinomiales Modell mit K Gruppen als Summe von K unabhängigen Teststatistiken für die einzelnen Gruppen geschrieben werden kann (vgl. Abschnitt 4.3.4). Unter H_1 ist nach Cochran und Mitra jede dieser K Teststatistiken asymptotisch nonzentral χ^2-verteilt mit Nonzentralitätsparameter $\gamma_k = N_k \cdot {w_k}^2$, $k = 1, ..., K$. Nun ist aber die Summe unabhängiger nonzentral χ^2-verteilter Zufallsvariablen selbst wieder nonzentral χ^2-verteilt, wobei sich sowohl die Freiheitsgrade als auch die Nonzentralitätsparameter addieren (Johnson & Kotz, 1970b, S. 135). Die Summenstatistik hat also den Nonzentralitätsparameter

$$\gamma = \sum_{k=1}^{K}\gamma_k = \sum_{k=1}^{K} N_k \cdot {w_k}^2 = N \cdot \sum_{k=1}^{K}\tau_k \cdot {w_k}^2 = N \cdot w^2\ , \qquad (4.3.3.1\text{-}4)$$

was zu beweisen war.

Leider ist die Anwendung der Cohenschen Effektstärkekonventionen auf den Mehr-Gruppen-Fall im allgemeinen problematisch. Man beachte insbesondere folgende Komplikation: Im Falle $K = 1$ ist mit einem Anstieg des Gesamtstichprobenumfangs automatisch ein Anstieg des Nonzentralitätsparameters und damit auch der Teststärke verbunden, sofern $w > 0$. Gleiches gilt im Fall $K > 1$ nicht mehr zwangsläufig. Ist zwar $w > 0$, dennoch aber ein bestimmtes $w_k = 0$, so daß die Häufigkeiten in dieser Gruppe modellbedingt immer residuenfrei angepaßt werden, dann ändert sich durch einen größeren Stichprobenumfang N_k in dieser Gruppe weder w noch γ und folglich auch nicht die Teststärke. Die Teststärke hängt also nicht nur vom Gesamtstichprobenumfang ab, sondern ganz wesentlich auch von der Aufteilung des Gesamtstichprobenumfangs auf die verschiedenen Gruppen. Eine Ausnahme ist nur dann gegeben, wenn alle w_k gleich sind, was zur Folge hat, daß sich die τ_k-Terme aus Gleichung (4.3.3.1-3) herauskürzen. In diesem Fall erscheint eine Übertragung der Cohenschen Konventionen auf verbundene multinomiale Modelle unproblematisch. Unterscheiden sich die w_k -Effekte jedoch, so können aus der blinden Übernahme der Cohenschen Konventionen Paradoxien derart resultieren, daß man beispielsweise eine „mittlere" Gesamt-Effektstärke w bei konstanten H_0- und H_1-Kategoriewahrscheinlichkeiten allein dadurch in eine „kleine" oder „große" Gesamt-Effektstärke verwandeln kann, daß man die Aufteilung der Gesamtstichprobe auf die einzelnen Gruppen verändert. Dies ist sicherlich nicht intendiert, wenn man den Begriff „Effektstärke" verwendet.

4.3.3.2 Teststärke als Funktion der Modellparameter unter H_1

Eine naheliegende Alternative besteht darin, modellspezifische Effektstärkemaße auf der Basis der Modellparameter zu definieren. Im Falle des Enkodierung-Abruf-Modells (vgl. Abschnitt 3.5.1) kann man beispielsweise vom saturierten Modell mit vier Parametern c, r, u, a ausgehen und als Effektstärkemaß für den Test des 3-Parameter-Modells (H_0: $u = a$) die Differenz $\Delta = |u - a|$ festlegen. Nennt man nun z.B. $\Delta = .10$ eine „kleine", $\Delta = .30$ eine „mittlere" und $\Delta = .50$ eine „große" Modellabweichung, dann ist unmittelbar klar, wie diese Begriffe zu verstehen sind. Die Bedeutung von „klein", „mittel" und „groß" ändert sich auch nicht dadurch, daß man die Anteile der Gesamtstichprobe ändert, die auf Itempaare und Einzelitems entfallen.

Die praktische Umsetzung der Teststärkekontrolle bereitet allerdings insoweit Probleme, als mit einem bestimmten Δ-Effektstärkemaß noch nicht der komplette Kategoriewahrscheinlichkeitsvektor $\boldsymbol{\pi}$ festgelegt ist, der eine bestimmte Punkt-Alternativhypothese definiert. Sehr viele unterschiedliche $\boldsymbol{\pi}$-Vektoren passen zum gleichen Δ-Effektmaß, liefern aber unterschiedliche Teststärken $1-\beta(\boldsymbol{\pi})$.

Man löst dieses Problem am besten so, daß man für alle Modellparameter unter H_1 Werte festlegt, die für eine gegebene Anwendungssituation plausibel sind. Hierbei ist die Randbedingung einzuhalten, daß die kritischen Modellparameter *u* und *a* die gewünschte Differenz Δ aufweisen. Plausible Parameterwerte können z.B. aus einschlägigen Hintergrundtheorien oder auch aus Pilotstudien (die unter ähnlichen Versuchsbedingungen und mit ähnlichen Probanden durchgeführt wurden) abgeleitet werden.

Sind alle Modellparameterwerte unter H_1 festgelegt, kann man den H_1-Kategoriewahrscheinlichkeitsvektor $\boldsymbol{\pi}$ durch Einsetzen der H_1-Parameterwerte in die entsprechenden Modellgleichungen deduzieren. Manchmal ist eine Erweiterung des zu prüfenden Modells zu einem saturierten H_1-Modell nicht so leicht möglich wie im Falle des Enkodierung-Abruf-Modells. In diesem Fall muß man den Vektor $\boldsymbol{\pi}$ auf andere Weise ableiten, beispielsweise dadurch, daß man gemäß (4.3.3-3) einen H_0-H_1-Diskrepanzvektor $\boldsymbol{\delta}$ mit den o.g. Eigenschaften auf fiktive H_0-Kategoriewahrscheinlichkeiten $\boldsymbol{p} = f(\boldsymbol{\theta})$ addiert, wobei $\boldsymbol{\theta}$ und $\boldsymbol{\delta}$ erneut „plausibel" zu wählen sind.

Im nächsten Schritt ist das Effektmaß *w* bzw. der Nonzentralitätsparameter γ zu bestimmen, der zu diesem Vektor $\boldsymbol{\pi}$ gehört. Da die Elemente von $\boldsymbol{\theta}$ freie Parameter des Modells sind, d.h. beim Modelltest nicht a priori auf bestimmte Werte fixiert sind, geht dies nicht einfach durch Einsetzen von $\boldsymbol{\pi}$ und $\boldsymbol{p}(\boldsymbol{\theta})$ in die γ-Formeln (4.3.3-4) für einfache oder (4.3.3.1-1) für verbundene multinomiale Modelle. Vielmehr müssen die (unbekannten) Kategoriewahrscheinlichkeiten unter H_0 bestimmt werden, die noch am ehesten mit dem festgelegten $\boldsymbol{\pi}$ kompatibel sind. Man kann eine Untergrenze für den Nonzentralitätsparameter und die Teststärke finden, indem man $\boldsymbol{p}(\boldsymbol{\theta})$ unter H_0 so wählt, daß γ bei gegebenen $\boldsymbol{\pi}$- und $\boldsymbol{N}$-Vektoren minimal wird (Guenther, 1977). Wie man sich durch Umformung der Gleichungen (4.3.3-4) und (4.3.3.1-1) leicht klarmachen kann, ist der minimale Nonzentralitätsparameter γ im Fall $K = 1$ und auch im Fall $K > 1$ rechnerisch nichts anderes als die Minimum-$PD^{\lambda=1}$-Statistik zu dem H_0-Modell, angewandt auf die erwarteten Häufigkeiten unter H_1 zu gegebenen Stichprobenumfängen $\boldsymbol{N} = (N_1, ..., N_k)$. Man kann also den Nonzentralitätsparameter γ im Prinzip genauso wie eine Minimum-Pearson-χ^2-Modellanpassungsstatistik berechnen, wobei man allerdings statt empirischer Häufigkeiten die erwarteten Häufigkeiten unter H_1 als „Daten" verwenden muß. Da $\gamma = N \cdot w^2$ gilt, erhält man die zugehörige Effektgröße *w* einfach durch die Umkehrtransformation $w = (\gamma / N)^{0.5}$, mit *N* als Gesamtstichprobenumfang.

Obwohl der Nonzentralitätsparameter γ rechnerisch ein Minimum-$PD^{\lambda=1}$-Parameter ist, ist dessen Begründung nicht davon abhängig, welche Minimum-PD^{λ}-Statistik als Prüfstatistik Verwendung findet. Für die betrachtete Klasse lokaler Alternativhypothesen $H_{1,N}$ folgen alle Minimum-PD^{λ}-Statistiken für $N \rightarrow \infty$

asymptotisch einer nonzentralen $\chi^2(\gamma)_{(df)}$-Verteilung (vgl. Drost et al., 1989b, S. 131; Read & Cressie, 1988, S. 171f).

4.3.3.3 Güte der klassischen $\chi^2(\gamma)_{(df)}$-Teststärkeapproximation

Ungeachtet des berichteten asymptotischen Resultats stellt sich natürlich die Frage, wie gut die Teststärkeapproximation (4.3.3-5) bei Stichprobenumfängen abschneidet, die in der Anwendungspraxis realistisch sind. Ferner ist von Interesse, wie sich die Approximation im Falle fixierter Alternativhypothesen verhält, die nicht mit zunehmenden N näher an H_0 heranrücken. Der Fall fixierter Alternativhypothesen scheint für die Anwendungspraxis relevanter als der Fall lokaler Alternativhypothesen zu sein, da letzterer bei großen Stichprobenumfängen ausschließlich minimale Modellverletzungen berücksichtigt. Ob eine auf der Prämisse „asymptotisch verschwindender" Modellverletzungen aufbauende Teststärkeapproximation für einen breiten Bereich von H_0-Verletzungsgraden und Stichprobenumfängen brauchbar ist, erscheint fraglich (Drost et al., 1989b, S. 131).

Die bislang publizierten Monte-Carlo-Studien erbrachten das auf den ersten Blick überraschende Ergebnis, daß die Brauchbarkeit der Teststärkeapproximation (4.3.3-5) u.a. davon abhängt, welche Minimum-PD^{λ}-Teststatistik Verwendung findet. Für Pearsons Statistik ($\lambda = 1$) ist die Approximation durchaus brauchbar, für $\lambda = 0$ und $\lambda = 2/3$ ist sie schlechter und für extreme λ-Werte muß sie als völlig unbrauchbar gelten (Drost et al., 1989b; Kallenberg, Oosterhoff & Schriever, 1985). Darüber hinaus spielt natürlich der Stichprobenumfang eine wichtige Rolle: Auch bei Verwendung von Pearsons Statistik ist die Teststärkeapproximation (4.3.3-5) z.T. unbefriedigend, sofern $N < 100$, und sie ist schlecht, wenn $N < 50$ (z.B. Slakter, 1968). Hierbei kommt es primär auf den Gesamtstichprobenumfang an; die Anzahl der Beobachtungskategorien bzw. die erwarteten Kategoriehäufigkeiten scheinen eine untergeordnete Rolle zu spielen. Generell besteht eine Verzerrungstendenz in dem Sinne, daß die Teststärkeapproximation (4.3.3-5) bei $10 \leq N \leq 50$ die faktische Teststärke um bis zu 25% überschätzt. Dies gilt zumindest beim Test von Gleichverteilungsnullhypothesen und $K = 1$ (Slakter, 1968).

Da bei Anwendungen multinomialer Modelle in der kognitiven Psychologie Stichprobenumfänge deutlich über $N = 100$ die Regel sein dürften — jedenfalls wenn die Daten über Vpn und Items aggregiert werden (vgl. dazu Kapitel 5) — ist die mäßige Approximationsgüte bei kleinem N kein schwerwiegendes Problem. Zu beachten ist auch, daß der Sinn von Teststärkeanalysen ja gerade darin besteht, durch entsprechende Planung des Stichprobenumfangs ausreichende Teststärke für die primär interessierenden Modell- und Hypothesentests sicherzustellen. Das wird nur mit hinreichend großem N zu erreichen sein, so daß man automatisch in

einen Bereich kommt, in dem die Teststärkeapproximation zumindest für die Pearson-Statistik recht gut funktioniert.

Ein gravierenderes Problem besteht darin, daß auch ein Stichprobenumfang im realistischen Bereich $100 < N < 500$ noch unbefriedigende Approximationsgüte zur Folge haben kann, wenn $\lambda = 0$ oder $\lambda = 2/3$ statt $\lambda = 1$ in der Minimum-PD$^\lambda$-Statistik gewählt wird. Da viele Argumente für die Verwendung der Cressie-Read- und vor allem der Log-Likelihood-Quotienten-Statistik sprechen (vgl. die Abschnitte 4.2.4 und 4.3.2.2), ist es naheliegend, sich nach alternativen Teststärke-Approximationsmethoden umzusehen.

4.3.3.4 Alternative Teststärkeapproximationen

Die einfachste Modifikation der Approximation (4.3.3-5) besteht darin, den Nonzentralitätsparameter γ der gewählten PD$^\lambda$-Prüfstatistik anzupassen. Bei Verwendung von PD$^\lambda(\boldsymbol{\theta}, \boldsymbol{y})$ als Prüfstatistik wäre also statt (4.3.3.1-1) der modifizierte Nonzentralitätsparameter

$$\gamma^{(\lambda)} = \min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \mathrm{PD}^\lambda(\boldsymbol{\theta}; \boldsymbol{N}, \boldsymbol{\pi}) = \min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \frac{2}{\lambda(\lambda+1)} \cdot \sum_{k=1}^{K} \boldsymbol{N}_k \sum_{j=1}^{J_k} \pi_{k,j} \left[\left(\frac{\pi_{k,j}}{p_{k,j}(\boldsymbol{\theta})} \right)^\lambda - 1 \right] \quad (4.3.3.4\text{-}1)$$

in die Approximationsformel (4.3.3-5) einzusetzen, wobei $\gamma^{(\lambda)}$ für $\lambda = -1$ und $\lambda = 0$ wie im Falle der PD$^\lambda$-Statistik (4.2.2-1) als Grenzwerte zu definieren sind. Gleichung (4.3.3.4-1) ist nichts anderes als die Minimum-PD$^\lambda$-Statistik, angewandt auf die erwarteten Häufigkeiten unter H_1 zu gegebenen Stichprobenhäufigkeiten $\boldsymbol{N}$ und H_1-Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$. Für $\lambda = 1$ ist (4.3.3.4-1) folglich identisch mit (4.3.3.1-1), nicht aber für $\lambda \neq 1$. Der Nonzentralitätsparameter $\gamma^{(\lambda)}$ für einfache multinomiale Modelle ergibt sich als Spezialfall von (4.3.3.4-1) für $K = 1$.

Die gleiche Taylorreihenerweiterung, mit der man (4.3.3.1-1) als Nonzentralitätsparameter der nonzentralen χ^2-Verteilung bei lokalen Alternativhypothesen rechtfertigen kann, kann ebenfalls zur Begründung von (4.3.3.4-1) herangezogen werden (Drost et al., 1989b, S. 131). Der traditionelle, auf lokalen Alternativhypothesen basierende Ansatz trennt also nicht besonders gut zwischen unterschiedlichen Approximationsmethoden für die Verteilung der Minimum-PD$^\lambda$-Statistik unter H_1.

Man kann jedoch versuchen, durch Monte-Carlo-Studien Klarheit in das Verhalten verschiedener Approximationsmethoden bei realistischen Stichprobenumfängen zu bringen. Die bislang publizierten Befunde zeigen ziemlich klar, daß der modifizierte Nonzentralitätsparameters $\gamma^{(\lambda)}$ für $\lambda = 0$ tatsächlich einen erheblichen Gewinn an Approximationsgüte zur Folge hat, wenn die Log-Likelihood-Quotienten-Prüfstatistik verwendet wird. Für λ-Werte, die sich noch stärker von 1 unterscheiden, hilft allerdings auch der modifizierte Nonzentralitätspara-

meter (4.3.3.4-1) nicht viel weiter (Drost et al., 1989b, S. 131). Anzumerken ist, daß die einschlägigen Monte-Carlo-Studien bislang auf einfache multinomiale Modelle ohne freie Parameter beschränkt sind.

Ein grundsätzlich anderer Ansatz der Teststärkeapproximation geht auf eine Arbeit von Broffitt und Randles (1977) zurück. Sie betrachteten nicht lokale, sondern fixierte Alternativhypothesen, die für $N \to \infty$ nicht näher an H_0 heranrücken. Unter diesen Umständen konvergiert die Verteilung der Pearson-Statistik asymptotisch gegen eine Normalverteilung, für die Broffitt und Randles (1977) den Mittelwert und die Varianz in Abhängigkeit von $\boldsymbol{\pi}$, $\boldsymbol{p}(\boldsymbol{\theta})$ und $\boldsymbol{N}$ angegeben haben. Wie Read und Cressie (1988) gezeigt haben, läßt sich dieser Ansatz auch auf andere Minimum-PD^{λ}-Statistiken mit $\lambda > -1$ erweitern, wobei jedoch Mittelwert und Varianz der asymptotischen Normalverteilung von λ abhängen. Die entsprechende Formeln für den Mittelwert (identisch mit $\gamma^{(\lambda)}$ gemäß (4.3.3.4-1)) und die Varianz (vgl. Drost et al., 1989b, S. 135; Read & Cressie, 1988, S. 56) sind publiziert, wobei allerdings der interessante Fall $\lambda = 0$ leider ausgespart bleibt.

Insgesamt scheint kein großer Gewinn mit der Verwendung der Normalverteilungsapproximation verbunden zu sein. Broffitt und Randles (1977) berichteten Monte-Carlo-Studien zur nonzentralen Verteilung der Pearson-Statistik bei einfachen multinomialen Modellen ($K = 1$) und voll spezifizierten H_0- und H_1-Kategoriewahrscheinlichkeiten. Die Normalverteilungsapproximation war in einigen Fällen schlechter und nur bei Alternativhypothesen mit großen Effektstärken besser als die klassische nonzentrale χ^2-Approximation mit Nonzentralitätsparameter γ. Mit wachsendem Stichprobenumfang (bis $N = 50$) stieg die Approximationsgüte der klassischen nonzentralen χ^2-Approximation schneller als die der Normalverteilungsapproximation an. Asymptotische Fehlerbetrachtungen zeigen darüber hinaus, daß die Normalverteilungsapproximation bei lokalen Alternativhypothesen (für die sie ja auch nicht entwickelt wurde) sehr schlecht abschneidet (Drost, Kallenberg, Moore & Oosterhoff, 1989a; 1989b).

Eine vielversprechendere Alternative zur klassischen nonzentralen χ^2-Approximation stammt von Drost et al. (1989a, 1989b). Drost und Mitarbeiter betrachteten eine Taylorreihenerweiterung der Minimum-PD^{λ}-Statistik, wobei sie neben linearen auch quadratische Terme berücksichtigten. Ignoriert man letztere, so resultiert unter beliebigen Alternativhypothesen eine Normalverteilung, die nach Korrektur von Mittelwert und Varianz exakt auf die von Broffitt und Randles (für $\lambda = 1$) bzw. Read und Cressie (für $\lambda \neq 1$) abgeleitete Normalverteilung hinausläuft. Die Normalverteilungsapproximation kann also aus einem Spezialfall des Ansatzes von Drost und Mitarbeitern hergeleitet werden, bei dem nur die linearen Terme einer Taylorreihenerweiterung der Minimum-PD^{λ}-Statistik beachtet werden. Berücksichtigt man darüber hinaus die quadratischen Terme, resultiert als Verteilung unter beliebigen Alternativhypothesen eine Linearkombination non-

zentraler χ^2-Verteilungen. Wie Drost et al. (1989b) anhand von Monte-Carlo-Studien gezeigt haben, ist diese Approximation für einen breiten Bereich von Alternativhypothesen sehr genau, wobei die Vorteile im Vergleich zu anderen Approximationen natürlich in erster Linie bei kleinen Stichprobenumfängen deutlich werden. Der Gewinn an Approximationsgenauigkeit wird allerdings mit einem erheblichem Aufwand bei der Berechnung der Teststärkewerte erkauft, der für Routineanwendungen unangemessen erscheint. Hinzu kommt, daß — von einigen Ausnahmefällen abgesehen (Taneichi & Sekiya, 1995) — gegenwärtig noch nicht klar ist, wie sich der Ansatz auf Tests verbundener parametrisierter Modelle mit mehreren zu schätzenden Parametern übertragen läßt.

Drost et al. (1989b) haben die Anwendungsprobleme einer „accurate but computationally difficult approximation" (S. 130) durchaus gesehen und deshalb eine Vereinfachung vorgeschlagen, die daraus resultiert, daß alle quadratischen Terme der Taylorreihenerweiterung das gleiche Gewicht erhalten. Dieses Gewicht wird so gewählt, daß sich der Erwartungswert der Taylorreihenapproximation durch die Gleichsetzung der Gewichte nicht ändert (Drost et al., 1989b, S. 133).[45] Als Verteilung dieser vereinfachten Taylorreihenerweiterung unter H_1 resultiert eine linear transformierte nonzentrale χ^2-Verteilung (vgl. Drost et al., 1989b, S. 134, Gleichungen (2.11) bis (2.13)). Nach entsprechender linearer Transformation des kritischen Wertes c_α kann man also die Teststärke wie bei der klassischen Approximation (4.3.3-5) unter Bezugnahme auf eine nonzentrale χ^2-Verteilung abschätzen. Allerdings ist deren Nonzentralitätsparameter nicht identisch mit (4.3.3.1-1) oder (4.3.3.4-1) (vgl. Gleichung (2.12) bei Drost et al., 1989b).

Die von Drost et al. (1989b) publizierten Monte-Carlo-Ergebnisse zu einfachen multinomialen Modellen mit voll spezifizierten H_0- und H_1-Wahrscheinlichkeiten zeigen einen Vorteil ihrer linear transformierten nonzentralen χ^2-Approximation gegenüber der Approximation (4.3.3-5), welcher aber schon ab einem Gesamtstichprobenumfang von $N = 100$ vernachlässigbar ist, sofern man extreme λ-Werte ausklammert und für Minimum-PD^λ-Prüfstatistiken mit $\lambda \neq 1$ statt der klassischen Approximation den modifizierten Nonzentralitätsparameter $\gamma^{(\lambda)}$ gemäß Gleichung (4.3.3.4-1) verwendet (Drost et al., 1989b, S. 140, Tabelle 1).

Ein Problem der Drostschen nonzentralen χ^2-Approximation besteht darin, daß nicht klar ist, wie sie sich in auf verbundene multinomiale Modelle mit $S > 0$ zu schätzenden Parametern anwenden läßt (eine Ausnahme ist wiederum der von Taneichi & Sekiya, 1995, behandelte Fall von K-Gruppen-Homogenitätstests). Gibt es freie Parameter im Modell, so sind die H_0-Kategoriewahrscheinlichkeiten $\boldsymbol{p}(\boldsymbol{\theta})$

[45] Für die Log-Likelihood-Quotienten-Statistik G^2 ergibt sich durch diese Vereinfachung keine Änderung, denn hierbei sind die Gewichte der quadratischen Terme von vornherein gleich. Die beiden von Drost und Mitarbeitern vorgeschlagenen Approximationsmethoden sind also in bezug auf G^2 äquivalent.

nicht von vornherein bekannt, sondern müssen geeignet gewählt werden. Bei der klassischen Approximation wurde dieses Problem so gelöst, daß eine Untergrenze für die Teststärke durch Minimierung des Nonzentralitätsparameters γ für fixiertes $\boldsymbol{N}$ und $\boldsymbol{\pi}$ unter H_1 und beliebiges $\boldsymbol{p}(\boldsymbol{\theta}) \in f(\boldsymbol{\Omega})$ unter H_0 bestimmt wurde. Der gleiche Weg läßt sich bei Verwendung des Nonzentralitätsparameters $\gamma^{(\lambda)}$ beschreiten. Diesen Ansatz kann man jedoch nicht ohne weiteres auf die linear transformierte nonzentrale χ^2-Approximation von Drost et al. (1989b) übertragen, denn hierbei muß die Minimierung des Nonzentralitätsparameters wegen der erforderlichen linearen Transformation des kritischen Wertes c_α nicht automatisch eine Untergrenze für die Teststärke liefern. Ein Ansatz, der darauf hinausläuft, $\boldsymbol{p}(\boldsymbol{\theta}) \in f(\boldsymbol{\Omega})$ so zu wählen, daß die Teststärke minimal wird, kann also rechnerisch ähnlich aufwendig wie die „exakte", auf Linearkombinationen nonzentraler χ^2-Verteilungen basierende Approximation von Drost et al. (1989a, 1989b) sein.

4.3.3.5 Eine Monte-Carlo-Studie zur Power des EA-Modelltests

Um abschätzen zu können, ob der mit den Drostschen Teststärkeapproximationen verbundene Aufwand überhaupt erforderlich ist, wenn man sich mit Approximationsfehlern kleiner als .05 für typische Anwendungsfälle zufrieden gibt, wurde eine Monte-Carlo-Studie zum Enkodierung-Abruf-Modell (EA-Modell) von Batchelder und Riefer (1980, 1986) mit Gesamtstichprobenumfängen zwischen $N = 30$ und $N = 960$ durchgeführt. Das EA-Modell wurde als exemplarischer Anwendungsfall ausgewählt, weil es in bezug auf Kategorienanzahl, Gruppenanzahl, Parameteranzahl und Freiheitsgradzahl des Modelltests ($df = 4 - 3 = 1$) ein recht typisches kognitionspsychologisches MVB-Modell ist, das in der Vergangenheit zudem auch noch recht häufig eingesetzt wurde (vgl. Abschnitt 3.5.1). Evaluiert wurde die klassische nonzentrale χ^2-Approximation (4.3.3-5) in Verbindung mit dem Nonzentralitätsparameter γ sowie zusätzlich die Approximation auf der Basis des Nonzentralitätsparameters $\gamma^{(\lambda)}$ für $\lambda = 0, 2/3$. Sechs verschiedene Populationssituationen wurden zum EA-Modell konstruiert, nämlich eine H_0-konsistente und fünf H_1-konsistente Grundgesamtgesamtheiten, die unterschiedliche Abweichungen vom zu testenden H_0-Modell aufweisen. Die Werte der entsprechenden Modellparameter sind in Tabelle 4.1 zusammengefaßt. Zusätzlich sind dort verschiedene Effektstärkemaße zum Modelltest des Drei-Parameter-Modells (H_0: $u = a$) aufgeführt, die aus den jeweils angegebenen Modellparameterwerten resultieren, wenn 2/3 des Gesamtstichprobenumfangs auf Itempaare und 1/3 auf Einzelitems entfallen. Diese Aufteilung wurde in allen Simulationsstudien zugrunde gelegt, weil sie in den Arbeiten von Batchelder und Riefer (1980; 1986) zum EA-Modell präferiert wurde. Auf die Frage, ob diese Aufteilung unter Teststärkegesichtspunkten optimal ist, komme ich später zurück.

Tabelle 4.1:
Parameterwerte und Effektstärken der Monte-Carlo-Studie zum Log-Likelihood-Quotienten-, Cressie-Read- und Pearson-Modelltest des Enkodierung-Abruf-Modells von Batchelder und Riefer (1980; 1986).

Simulierte Parameterwerte				Effektstärkemaße			
c	r	u	a	Δ	$w^{(\lambda=1)}$	$w^{(\lambda=2/3)}$	$w^{(\lambda=0)}$
.50	.50	.50	.50	.0000	.0000	.0000	.0000
.50	.50	.45	.55	.1000	.0700	.0702	.0707
.50	.50	.40	.60	.2000	.1332	.1343	.1363
.50	.50	.35	.65	.3000	.1890	.1915	.1965
.50	.50	.30	.70	.4000	.2365	.2412	.2505
.50	.50	.25	.75	.5000	.2751	.2823	.2970

Das Effektmaß $w^{(\lambda=1)}$ entspricht dem in Gleichung (4.3.3.1-3) definierten Cohenschen Effektmaß w. Die Maße $w^{(\lambda=2/3)}$ und $w^{(\lambda=0)}$ sind die analogen Effektmaße für $\lambda \neq 1$, die zu den Nonzentralitätsparametern $\gamma^{(\lambda=2/3)}$ und $\gamma^{(\lambda=0)}$ führen. Errechnen kann man alle drei Effektmaße, indem man zunächst erwartete Häufigkeiten zu den Parameterwerten in den vier linken Spalten generiert, wobei ein beliebiger Gesamtstichprobenumfang N gewählt werden kann, jedoch die Aufteilung auf Itempaare ($\tau_1 = 2/3$) und Einzelitems ($\tau_2 = 1/3$) zu beachten ist. Anschließend wird die Minimum-PD$^{\lambda}$-„Statistik" des H_0-Modells zu diesen erwarteten Häufigkeiten berechnet. Diese liefert den Nonzentralitätsparameter $\gamma^{(\lambda)}$. Durch die Umkehrtransformation $w^{(\lambda)} = (\gamma^{(\lambda)} / N)^{0.5}$ erhält man dann die Effektstärke $w^{(\lambda)}$. Wie man durch Vergleich mit der Spalte Δ erkennt, ist $w^{(\lambda)}$ durchweg erheblich kleiner als die absolute Differenz von u and a. Bereits ein „mittlerer Effekt" ($w = .3$) im Sinne von Cohen (1988) korrespondiert also offensichtlich einer substantiellen Modellverletzung, die — in Termini der kritischen Parameterdifferenz ausgedrückt — größer als $\Delta = .5$ ist. Dies ist übrigens ein recht typisches Resultat, welches sich fast regelmäßig ergibt, wenn man Parameterdifferenzen unter H_1 in $w^{(\lambda)}$-Effektmaße umrechnet. Das, was man intuitiv als mittlere bis große Parameterdifferenzen in MVB-Modellen bezeichnen würde (z.B. $\Delta = .3$ bis $\Delta = .5$), entspricht oftmals eher kleinen bis mittleren Effektgrößen im Sinne Cohens (1988).

Aus jeder der sechs Grundgesamtheiten gemäß Tabelle 4.1 wurden jeweils 1000 unabhängige Stichproben des Gesamtstichprobenumfangs N = 30, 60, 120, 240, 480, 960 generiert, wobei die Annahmen des Modells einer verbundenen Multinomialverteilung eingehalten wurden (vgl. Abschnitt 3.2). Praktisch umgesetzt wurde dies mit Hilfe eines Pseudozufallszahlengenerators, der im Intervall [0, 1]

gleichverteilte Zufallszahlen lieferte.[46] Das Intervall [0, 1] wurde in vier bzw. zwei Bereiche eingeteilt, deren Größe jeweils den aus dem Populationsmodell abgeleiteten Kategoriewahrscheinlichkeiten für Itempaare bzw. Einzelitems entsprach. Pro Stichprobe eines Umfangs N wurden $2/3 \cdot N$ Zufallszahlen zu Itempaaren und $1/3 \cdot N$ Zufallszahlen zu Einzelitems erzeugt. Es wurde jeweils die Häufigkeit derjenigen Kategorie um 1 inkrementiert, in deren Bereich die betreffende Zufallszahl fiel, bis der gewünschte Gesamtstichprobenumfang erreicht war. Die so generierten 1000 Stichproben zum gleichen Populationsmodell und N wurden mittels AppleTree (Rothkegel, 1998; 1999) auf der Grundlage des 3-Parameter-EA-Modells analysiert, d.h. es wurden mit Hilfe des EM^{λ}-Algorithmus (vgl. Abschnitt 4.2.2) jeweils die Minimum-PD^{λ}-Statistiken zur H_0: $u = a$ für $\lambda = 0, 2/3$ und 1 berechnet. Ein Schrittweitenkorrekturfaktor von $\varepsilon = .5$ führte in jedem Fall zur Konvergenz; lokale Minima wurden in keinem einzigen Fall identifiziert. Die einzige Komplikation waren vereinzelte Kategoriehäufigkeiten von 0, die bei den kleineren Gesamtstichprobenumfängen gelegentlich vorkamen. In diesen Fällen inkrementiert AppleTree automatisch alle Kategoriehäufigkeiten um 1, so daß die Log-Likelihood-Quotienten-Statistik (4.2.1-2) unabhängig von der Schreibweise immer definiert ist. Diese „Datenmodifiktion" erzwingt natürlich eine gewisse Abweichung vom zugrundeliegenden Populationsmodell, dürfte aber auf die Ergebnisse der Monte-Carlo-Studie keinen wesentlichen Einfluß gehabt haben, da Nullzellen insgesamt sehr selten vorkamen.

Für jede der drei untersuchten Prüfstatistiken wurde getrennt ausgezählt, wie viele der PD^{λ}-Statistiken zu den 1000 Monte-Carlo-Stichproben die kritischen Werte der $\chi^2_{(1)}$-Verteilung für die Signifikanzniveaus $\alpha = .01, .05$ und .10 erreicht oder überschritten haben. Die relative Häufigkeit signifikanter Prüfstatistiken diente als Schätzer $1-\hat{\beta}(\boldsymbol{\pi})$ für die Teststärke relativ zum verwendeten Signifikanzniveau, Populationsmodell, Stichprobenumfang und λ-Koeffizienten der Prüfstatistik. Bei der Bewertung dieser Teststärke-Schätzer ist zu beachten, daß sie natürlich mit Schätzfehlern behaftet sind, also mit der faktischen Teststärke nicht

[46] Verwendet wurde die Standard-RND-Funktion des GWBASIC-Interpreters für das DOS-Betriebssystem, die sich in zuvor durchgeführten Qualitätstests als zuverlässig erwiesen hat. Prinzipiell lassen sich Monte-Carlo-Studien auch mit den Programmen MBT (Hu, 1995) und GPT (Hu, 1999; Hu & Phillips, 1999) durchführen. Die Monte-Carlo-Option in GPT war jedoch zum Zeitpunkt der Durchführung der Monte-Carlo-Studie noch nicht verfügbar und ist im übrigen auch derzeit noch auf die Log-Likelihood-Quotienten-Statistik beschränkt. Somit lassen sich nicht alle Aspekte der hier durchgeführten Monte-Carlo-Studie mit GPT abdecken. Versuche mit dem Programm MBT zeigten, daß die verwendete Version 1.1 vom 25.9.95 bei der Erzeugung von Zufallsstichproben zum EA-Modell mit vorgegebenen Modellparametern offenbar nicht fehlerfrei arbeitete: Auch bei sehr großen simulierten Stichprobenumfängen war eine Kategorie ($C_{1,2}$) bei $p(C_{1,2}) = .5 \cdot .3 \cdot .3 = .045$ auffallend häufig nicht besetzt, und in den restlichen Fällen waren die Kategoriehäufigkeiten der Monte-Carlo-Stichproben auffallend häufig Vielfache derselben natürlichen Zahl. Eine Überprüfung mit einer neueren Kopie von MBT 1.1 vom 26. 3. 98 bestätigte dieses Ergebnis. Deshalb wurde ein eigenes Programm zur Erzeugung von Zufallsstichproben geschrieben.

exakt übereinstimmen müssen. Allerdings ist der Standardschätzfehler bei 1000 Monte-Carlo-Stichproben schon recht gering: Liegt die faktische Teststärke bei $1-\beta(\boldsymbol{\pi}) = .5$, so nimmt der Standardschätzfehler für $1-\hat{\beta}(\boldsymbol{\pi})$ den bei 1000 Stichproben maximal möglichen Wert $\sigma = .016$ an; für $1-\beta(\boldsymbol{\pi}) = .95$ oder $1-\beta(\boldsymbol{\pi}) = .05$ beträgt er sogar nur $\sigma = .007$. Abweichungen zwischen nominellen und geschätzten Teststärkewerten im Bereich .01 bis .02 sind also durchaus noch auf Stichprobenfehler bei der Ermittlung der Teststärke-Schätzwerte zurückzuführen; erst Abweichungen ab .02 sind ein klares Indiz für Approximationsfehler bei der Bestimmung der nominellen Teststärke. Da der Standardschätzfehler für Teststärkewerte in der Nähe von $1-\beta(\boldsymbol{\pi}) = .5$ etwa doppelt so groß wie bei extremen Teststärkewerten ist, sollten auch die Diskrepanzen zwischen geschätzten und nominellen Teststärkewerten im mittleren Bereich tendenziell größer als bei extremen Werten ausfallen.

Tabelle 4.2:
Durchschnittliche (∅) und maximale absolute Differenzen (max) zwischen nominellen und faktischen Teststärkewerten in der Monte-Carlo-Studie zum Log-Likelihood-Quotienten- (G^2), Cressie-Read- (CR) und Pearson-Modelltest des EA-Modells von Batchelder und Riefer (1980; 1986); für verschiedene Gesamtstichprobenumfänge und Signifikanzniveaus.

		Prüfstatistik mit verwendetem Nonzentralitätsparameter					
		G^2 mit $\gamma^{(0)}$		C-R mit $\gamma^{(2/3)}$		Pearson mit $\gamma^{(1)}$	
N	α	∅	max	∅	max	∅	max
30	.01	.051	.134	.055	.135	.054	.128
	.05	.065	.157	.075	.173	.078	.162
	.10	.047	.107	.062	.126	.069	.146
60	.01	.017	.037	.035	.071	.047	.092
	.05	.007	.019	.010	.020	.015	.032
	.10	.023	.035	.019	.043	.017	.045
120	.01	.015	.038	.017	.062	.023	.067
	.05	.021	.040	.027	.049	.031	.066
	.10	.017	.042	.023	.046	.027	.057
240	.01	.022	.045	.026	.048	.029	.057
	.05	.017	.031	.020	.041	.021	.048
	.10	.012	.034	.016	.025	.017	.027
480	.01	.006	.025	.010	.036	.011	.038
	.05	.004	.011	.005	.010	.006	.014
	.10	.006	.018	.006	.015	.006	.014
960	.01	.005	.017	.005	.015	.006	.019
	.05	.003	.014	.003	.014	.004	.016
	.10	.001	.003	.001	.004	.001	.005
Durchschnitt:		.019	.048	.023	.051	.026	.057

Die Ergebnisse der Monte-Carlo-Studie bestätigen zunächst eine Vermutung, die man aus den bereits referierten Untersuchungen zu einfachen multinomialen Modellen mit fixierten H_0-Kategoriewahrscheinlichkeiten ableiten kann: Die klassische, auf dem Nonzentralitätsparameter γ basierende χ^2-Approximation (4.3.3-5) ist zwar für Pearsons χ^2 besser als die auf $\gamma^{(\lambda)}$ basierenden Approximationen mit $\lambda \neq 1$. Für G^2 oder die Cressie-Read-Statistik erhält man jedoch optimalere Teststärkeapproximationen, wenn man statt γ den Nonzentralitätsparameter $\gamma^{(\lambda)}$ verwendet, wobei λ der Prüfstatistik anzupassen ist (d.h. $\lambda = 0$ für G^2 und $\lambda = 2/3$ für die Cressie-Read-Statistik). Die im folgenden berichteten Ergebnisse beschränken sich daher ausschließlich auf die nonzentrale χ^2-Approximation mit Nonzentralitätsparameter $\gamma^{(\lambda)}$.

Die Ergebnisse für die in der Anwendungspraxis dominierenden Signifikanzniveaus $\alpha = .01$ und $\alpha = .05$ sind in den Abbildungen 4.7 und 4.8 graphisch dargestellt. Zusätzlich werden durchschnittliche und maximale Abweichungen zwischen nominellen und geschätzten Teststärkewerte in Tabelle 4.2 zusammengefaßt, und zwar für alle drei untersuchten Signifikanzniveaus. Wie nicht anders zu erwarten, funktionieren alle Approximationen für $N = 30$ ziemlich schlecht. Die mittlere absolute Differenz zwischen nominellen und empirisch geschätzten Teststärkewerten — im folgenden „mittlerer Approximationsfehler" genannt — liegt oberhalb von .05, der maximale Approximationsfehler oberhalb von .10, z.T. ist er sogar größer als .15 (vgl. Tabelle 4.2). Wie schon Slakter (1968) berichtete, überschätzen die nominellen Teststärkewerte die faktischen, und zwar um so mehr, je stärker das H_0-Modell verletzt ist. Doch auch bei Gültigkeit des Nullhypothesenmodells sind die Tests konservativ, d.h. das faktische α ist erheblich kleiner als das nominelle.

Überraschenderweise stellt sich das Bild schon ab einem Gesamt-N von 60 grundsätzlich anders dar. Der mittlere Approximationsfehler liegt bereits durchweg unter .04, der maximale Approximationsfehler größtenteils unter .05 (vgl. Tabelle 4.1). Das entspricht einer Genauigkeit, mit der man sich in der Praxis schon zufrieden geben kann. Wie nicht anders zu erwarten, steigt die Genauigkeit mit wachsendem N weiter an; ab $N = 480$ liegen die Abweichungen im Toleranzbereich des Standardschätzfehlers. Geringfügige Abweichungen vom Trend sind — wie die Abbildungen zeigen — in erster Linie auf Teststärkewerte im mittleren Bereich zurückzuführen, deren Schätzer einen größeren Standardschätzfehler aufweisen.

Im gegebenen Anwendungsfall ist die Log-Likelihood-Quotienten-Statistik in bezug auf die absolute Höhe der Teststärke und in bezug auf die Approximationsgenauigkeit besser als die Cressie-Read-Statistik, welche wiederum etwas besser als die Pearson-Statistik ist. Die Unterschiede sind jedoch nicht sehr groß und

aus Gründen, die im folgenden noch angesprochen werden, wahrscheinlich nicht auf andere Anwendungsfälle generalisierbar.

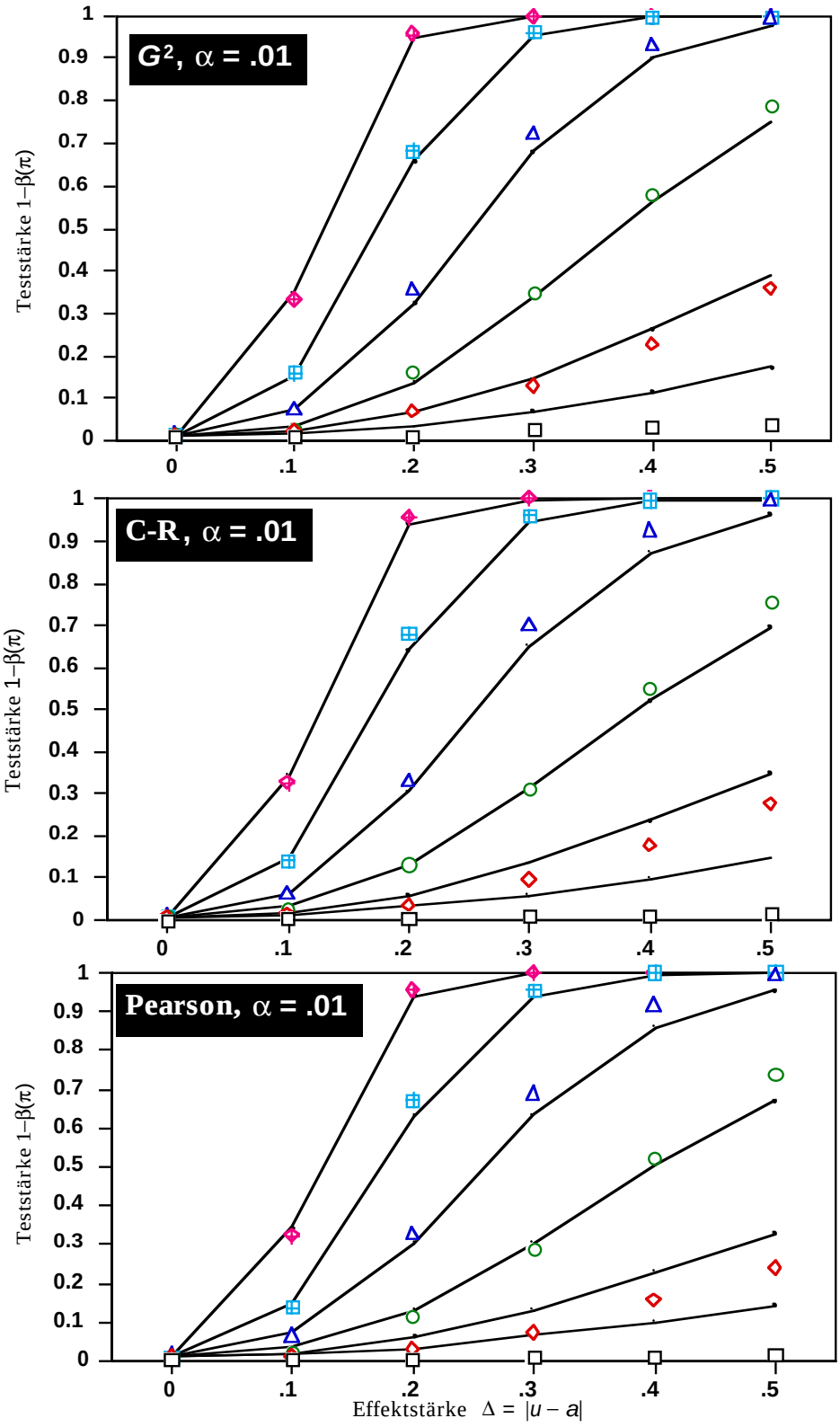

Abbildung 4.7:
Empirisch geschätzte (Symbole) und approximative Teststärkewerte (Linien) für N = 30, 60, 120, 240, 480, 960 (jeweils von unten nach oben) und α = .01; getrennt für die Statistiken G^2 (oben), Cressie-Read- (Mitte) und Pearson-χ^2 (unten).

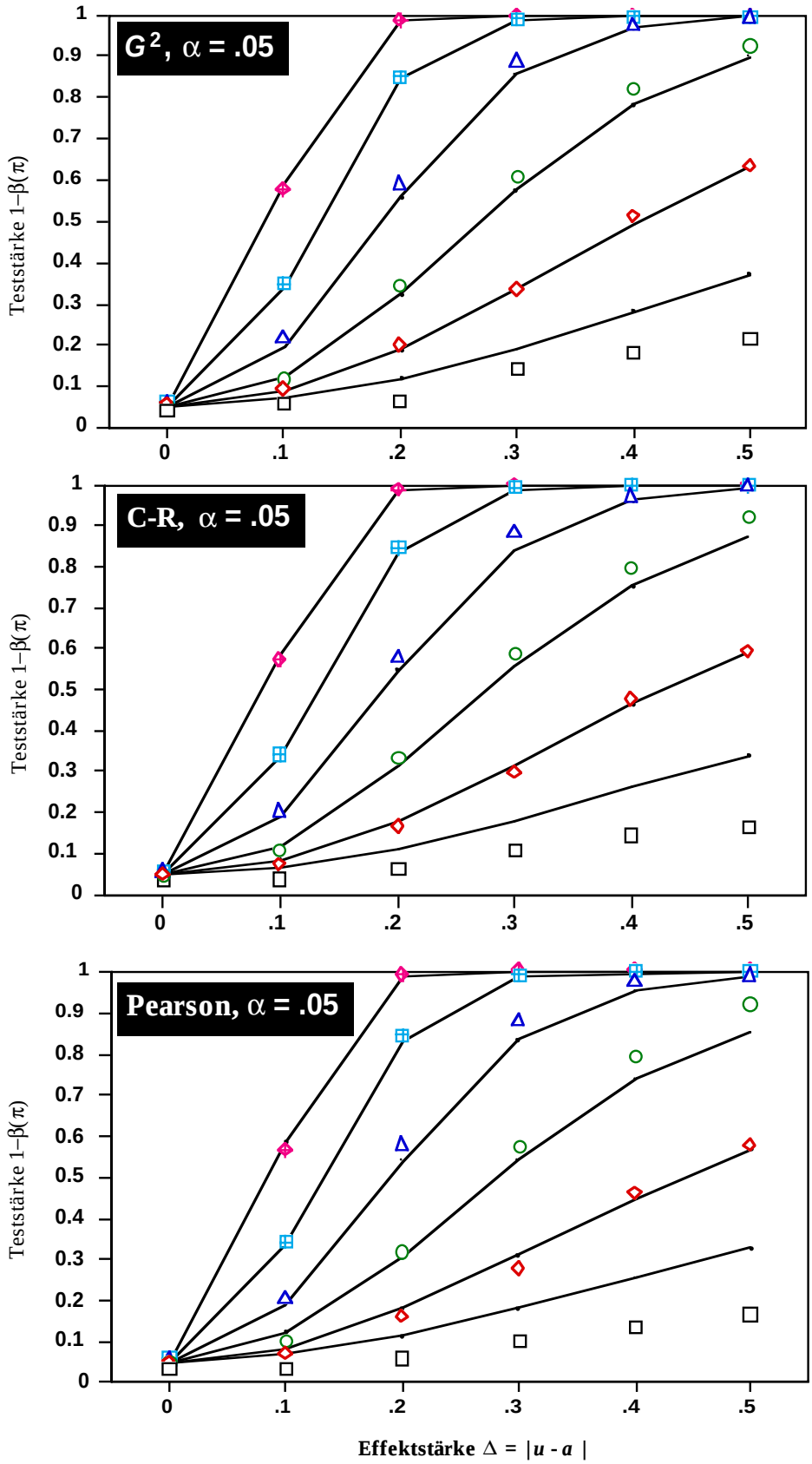

Abbildung 4.8:
Empirisch geschätzte (Symbole) und approximative Teststärkewerte (Linien) für N = 30, 60, 120, 240, 480, 960 (jeweils von unten nach oben) und α = .05; getrennt für die Statistiken G^2 (oben), Cressie-Read- (Mitte) und Pearson-χ^2 (unten).

Als Fazit der Monte-Carlo-Studie kann festgehalten werden, daß die nonzentrale χ^2-Verteilung mit dem Nonzentralitätsparameter $\gamma^{(\lambda)}$ gemäß Gleichung (4.3.3.4-1) für Teststärkeapproximationen bei Minimum-PD$^{\lambda}$-Modelltests sehr gut brauchbar zu sein scheint, sofern Gesamtstichprobenumfänge mit $N > 50$ vorliegen und $\lambda = 0, 2/3$ oder 1 gewählt wird. Möglicherweise kann dieses Ergebnis auch noch auf andere Minimum-PD$^{\lambda}$-Modelltests erweitert werden; Monte-Carlo-Studien zu einfachen multinomialen Modellen ohne freie Parameter legen aber die Vermutung nahe, daß die Approximation bei extremen λ–Werten versagt. Auch die klassische nonzentrale χ^2-Approximation auf der Basis des Nonzentralitätsparameters γ gemäß Gleichung (4.3.3-1) kann durchaus verwendet werden; sofern allerdings nicht Pearsons χ^2 als Prüfstatistik gewählt wird, ist diese Approximation im Vergleich zur genauso leicht berechenbaren $\gamma^{(\lambda)}$-Approximation suboptimal.

Lediglich bei sehr kleinem N unterhalb von 50 oder extremem λ-Parameter der Prüfstatistik muß von der Verwendung der hier evaluierten Approximationen abgeraten werden. Sowohl das faktische α-Niveau als auch die faktische Teststärke weichen dann erheblich von den nominellen Werten ab. Sollten Teststärkeanalysen für solche Fälle erforderlich sein, muß man wohl auf die mehr oder minder aufwendigen Approximationen von Drost et al. (1989b) zurückgreifen; die Normalverteilungsapproximation ist jedenfalls keine überzeugende Alternative und somit eigentlich generell verzichtbar.

In der Anwendungspraxis dürften auch die Drostschen Approximationen eher eine untergeordnete Rolle spielen, da (a) die meisten Gründe gegen die Verwendung extremer λ-Parameter sprechen (s.u.) und (b) Teststärkewerte im Regelfall erst ab einem Gesamt-N von 200 bis 500 in einem akzeptabel hohen Bereich liegen. Bei derartig hohen Stichprobenumfängen ist man aber — wie gezeigt — auch bei Verwendung der einfach berechenbaren nonzentralen χ^2-Approximation mit Nonzentralitätsparameter γ oder $\gamma^{(\lambda)}$ eindeutig auf der sicheren Seite.

4.3.3.6 Teststärkeoptimierung

Die aus versuchsplanerischer Sicht wichtigsten Einflußgrößen auf die Power einer statistischen Modellprüfung sind im allgemeinen der Umfang der gezogenen Stichprobe (N) und das α-Niveau des Tests. Da jedoch für dasselbe H_0-H_1-Hypothesenpaar mehrere potentielle Tests zur Verfügung stehen und der Grad der H_0-Verletzung (d.h. die Effektstärke w bzw. $w^{(\lambda)}$) oftmals nicht allein von einer speziell interessierenden H_1-Parameterhypothese abhängt, gibt es neben den o.g. Faktoren noch weitere versuchsplanerische und datenanalytische Möglichkeiten der Teststärkeoptimierung:. (a) die optimale Wahl des λ-Parameters der Prüfstatistik, (b) die optimale Aufteilung der Gesamtstichprobe auf die einzelnen Gruppen bzw. Ereignisbäume eines MVB-Modells, (c) Maßnahmen zur Optimierung bestimmter

Modellparameterwerte und (d) die Abstimmung des Tests (globaler Modelltest vs. χ^2-Differenzentest) auf ein gegebenes H_0-H_1-Hypothesenpaar. Diese Möglichkeiten sollen in diesem Abschnitt kurz diskutiert werden.

Maßnahme (a): λ-Optimierung. Eine naheliegende Frage ist die nach der unter Teststärkegesichtspunkten optimalen Wahl des λ-Parameters der Minimum-PD^{λ}-Prüfstatistik. Die asymptotischen Resultate zur sogenannten Pitman-Effizienz und Bahadur-Effizienz erlauben leider keine eindeutige Entscheidung. Ist N_2 der Stichprobenumfang, der bei der Prüfstatistik PD^{λ_2} die gleiche Teststärke erzeugt wie der Stichprobenumfang N_1 bei der Prüfstatistik PD^{λ_1}, so ist die Pitman-Effizienz (oder Asymptotische Relative Effizienz, A.R.E.) der asymptotische Quotient, gegen den das Verhältnis N_1 / N_2 für $N1 \rightarrow \infty$ unter lokalen Alternativhypothesen der Form (4.3.3-3) strebt. Die Pitman-Effizienz nimmt für beliebige Werte $\lambda_1 \in \mathbb{R}$ und $\lambda_2 \in \mathbb{R}$ den Wert A.R.E = 1 an, da alle Minimum-PD^{λ}-Statistiken unter lokalen Alternativhypothesen die gleiche asymptotische $\chi^2(\gamma)_{(df)}$-Verteilung aufweisen (Read & Cressie, 1988, S. 55). Unter dem Aspekt der Pitman-Effizienz sind also alle Minimum-PD^{λ}-Statistiken teststärkeäquivalent. Anders sieht es bei der Bahadur-Effizienz aus. Hier betrachtet man nicht lokale, sondern fixierte Alternativhypothesen und läßt stattdessen für $N1 \rightarrow \infty$ das α-Niveau des Tests gegen null gehen. Unter diesen Randbedingungen liefert G^2 die Obergrenze für die Effizienz, d.h. keine PD^{λ}-Statistik kann im Sinne der Bahadur-Effizienz teststärker als die Log-Likelihood-Quotienten-Statistik sein (Cressie & Read, 1984). Formal bewiesen wurde dies m.W. bislang allerdings nur für multinomiale Modelle ohne freie Parameter.

Bei endlichen Stichprobenumfängen kann prinzipiell jede PD^{λ}-Statistik maximal teststark sein. Entscheidend dafür, welches λ optimal ist, ist die Art der Abweichungen zwischen den erwarteten Häufigkeiten unter H_1 und H_0. Man kann hier in Anlehnung an die Terminologie bei approximativen MANOVA-F-Tests von „Nonzentralitätsstruktur" sprechen (vgl. etwa Bredenkamp & Erdfelder, 1985; Olson, 1976; Stevens, 1980). Gibt es im wesentlichen nur eine positive Abweichung zwischen erwarteten H_1- und H_0-Frequenzen, so liegt eine konzentrierte Nonzentralitätsstruktur vor. Unter diesen Umständen wächst die Teststärke monoton mit λ (Drost et al., 1989b; Read & Cressie, 1988). Für den speziellen Fall des Tests der Gleichverteilungsnullhypothese gegen eine Ausreißeralternativhypothese, die nur eine einzige positive Abweichung von der Gleichverteilung in einer unbekannten Beobachtungskategorie behauptet, ist verschiedentlich die maximale Zellfrequenz (unter Ignorierung aller anderen Zellfrequenzen) als optimale Teststatistik vorgeschlagen worden („Test für die Ausreißeralternative", vgl. Erdfelder & Funke, 1988; Levin, 1983). Man kann zeigen, daß der auf der maximalen Zellfrequenz basierende Test dem Minimum-$PD^{\lambda=\infty}$-Test äquivalent ist; es handelt sich um das

teststärkste Verfahren, das für diese Alternativhypothese zur Verfügung steht (Read & Cressie, 1988, S. 79).

Umgekehrt wächst die Teststärke mit monoton sinkendem, d.h. gegen $-\infty$ strebendem λ, wenn ein konzentrierte Nonzentralitätsstruktur mit einer einzigen *negativen* Ausreißerzelle vorliegt. Den Gegenpol zu konzentrierten Nonzentralitätsstrukturen mit positiven und negativen Ausreißerzellen bilden diffuse Nonzentralitätsstrukturen. Eine extrem diffuse Nonzentralitätsstruktur ist durch gleichmäßige positive und negative H_0-H_1-Diskrepanzen über alle Beobachtungskategorien hinweg gekennzeichnet. Die Beträge der logarithmierten Quotienten $y_{k,j} / (N_k \cdot p_{k,j}(\boldsymbol{\theta}))$ sind dann also in allen Kategorien j und Gruppen k in etwa gleich groß. Extrem diffuse Nonzentralitätsstrukturen lassen sich am effektivsten durch Wahl von $\lambda = 0$ in der Prüfstatistik aufdecken, d.h. mit der Log-Likelihood-Quotienten-Statistik (Drost et al., 1989b; Read & Cressie, 1988).

Bei typischen Modelltests parametrisierter multinomialer Modelle dürften eher diffuse als konzentrierte Nonzentralitätsstrukturen dominieren. Man kann dies im Prinzip vor der Durchführung eines Modelltests selbst ausprobieren, indem man die erwarteten Häufigkeiten zu einer H_1 generiert und für verschiedenen Werte von λ die Minimum-PD^{λ}-„Statistik" — d.h. den Nonzentralitätsparameter $\gamma^{(\lambda)}$ — zu einem spezifizierten H_0-Modell berechnet. Derjenige λ-Wert, der $\gamma^{(\lambda)}$ maximiert, ist der im Sinne der Teststärke optimale Wert. Bei der Monte-Carlo-Studie aus Abschnitt 4.3.3.5 wurde dies für $\lambda = 0, 2/3$ und 1 ausprobiert, mit dem Ergebnis, daß $\lambda = 0$ von diesen drei λ-Werten die größte Teststärke lieferte (vgl. Tabelle 4.1). Offenbar liegt in diesem Fall also eher eine diffuse als eine konzentrierte Nonzentralitätsstruktur mit positiver Ausreißerzelle vor. Inspiziert man die Residuen genauer, entdeckt man sogar eine leichte Tendenz zu einer negativen Ausreißerzelle (Kategorie $C_{1,3}$), die tendenziell leichter mit negativen λ-Parametern aufdeckbar sein sollte.

Obwohl der untersuchte Fall recht typisch sein dürfte, kann es vorkommen, daß der Vergleich unterschiedlicher Nonzentralitätsparameter ein extremes λ als optimal ausweist. Sollte man in einem solchen Fall tatsächlich den unter Teststärkegesichtspunkten optimalen Wert — z.B. $\lambda = 6$ — in die Minimum-PD^{λ}-Statistik einsetzen? Cressie und Read (1984) folgend würde ich hiervon abraten. Auf der negativen Seite ist nämlich die schlechtere Approximierbarkeit der exakten Minimum-$PD^{\lambda=6}$-Verteilung über die χ^2-Verteilung zu verbuchen, so daß man nur mit viel Aufwand genaue Angaben zu α- und β-Fehlerwahrscheinlichkeiten machen kann. Hinzu kommt die größere Ausreißeranfälligkeit der Parameterschätzungen und auch der Prüfstatistik selbst, die mit extremen λ-Werten zwangsläufig einhergeht. Man muß daher vermuten, daß statistische Analysen, die auf extremen λ-Werten beruhen, nicht so robust gegenüber Verletzungen des zugrundeliegenden Verteilungsmodells sind wie Analysen auf der Basis von $\lambda \in (-1, 1]$.

Im Regelfall sollte man also an den klassischen Minimum-PD$^{\lambda}$-Statistiken mit $\lambda = 0, 2/3$ oder 1 festhalten, auch wenn Teststärkebetrachtungen andere λ-Werte nahelegen (Cressie & Read, 1984). Innerhalb der Gruppe der klassischen Prüfstatistiken können Teststärkebetrachtungen die Präferenz für eine bestimmte Statistik unterstützen, beim Test des EA-Modells etwa die Präferenz für G^2. Allerdings muß man sich darüber im klaren sein, daß bedeutsame Teststärkegewinne durch optimale Wahl des λ-Parameters nicht möglich sind, wenn man sich von vornherein auf λ-Werte zwischen 0 und 1 festlegt. Sofern die Teststärke für Standard-λ-Parameter unbefriedigend ist und größere Stichprobenumfänge praktisch nicht realisierbar sind, mag dies in Einzelfällen ein Grund sein, sich für λ-Parameter außerhalb des Standardintervalls zu entscheiden. Allerdings muß diese Festlegung unbedingt *vor* der Datenerhebung bzw. vor der Durchführung des Tests erfolgen. Die Wahl von λ zu bereits vorliegenden Daten, derart, daß die Minimum-PD$^{\lambda}$-Statistik — je nach Wunsch — besonders groß oder besonders klein ausfällt, ist auf jeden Fall unzulässig. Im präasymptotischen Bereich werden die nominellen α- und β-Fehlerwahrscheinlichkeiten nämlich nicht mehr exakt eingehalten, wenn die Prüfstatistik auf gegebene Daten abgestimmt wird.

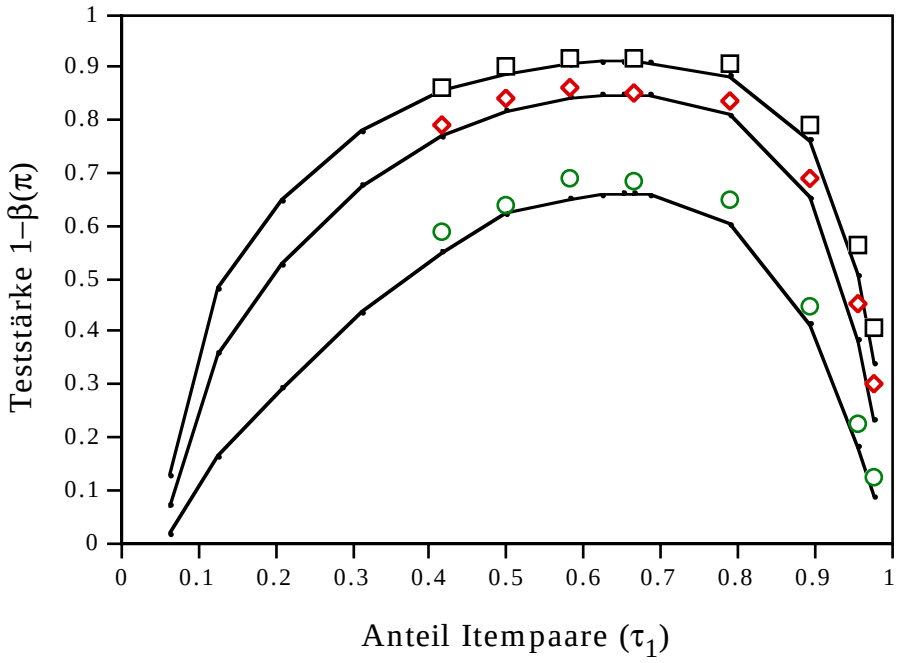

Abbildung 4.9:
Empirisch geschätzte (Symbole) und approximative Teststärkewerte (Linien) für den Log-Likelihood-Quotienten-Test des EA-Modells bei $N = 480$ und $\alpha = .01, .05, .10$ (von unten nach oben) in Abhängigkeit vom Anteil der Gesamtstichprobenumfangs, der auf Itempaare entfällt.

Maßnahme (b): τ_k-Optimierung. Eine wesentlich wichtigere Einflußmöglichkeit auf die Teststärke als die λ-Optimierung ist die τ_k-Optimierung, d.h. die geeignete Aufteilung des Gesamtstichprobenumfangs auf verschiedene Gruppen oder Ereignisbäume eines Gesamtmodells. Bei der Monte-Carlo-Studie zum EA-

Modell wurde beispielsweise in Anlehung an Batchelder und Riefer (1980; 1986) sowie Riefer und Batchelder (1991b) eine 2/3 versus 1/3 Aufteilung auf Itempaare und Einzelitems vorgenommen. Wir können nun fragen: Ist diese Aufteilung optimal im Sinne der Teststärkemaximierung? Wie Abbildung 4.9 zeigt, ist die Aufteilung *fast* optimal, jedenfalls bei den H_1-Parametern, die der Monte-Carlo-Studie zugrunde lagen. Abbildung 4.9 illustriert die Teststärke des Log-Likelihood-Quotienten-Tests zum Drei-Parameter-EA-Modell für die H_1-Wahrscheinlichkeiten in der dritten Zeile von Tabelle 4.1 (d.h. $\Delta = |u-a| = .20$), $N = 480$ und drei verschiedene Signifikanzniveaus ($\alpha = .10, .05, .01$) in Abhängigkeit vom Anteil $\tau_1 = N_1 / N$ des Gesamtstichprobenumfangs, der auf Itempaare entfällt. Die entsprechende Teststärkefunktion ist offensichtlich negativ beschleunigt. Hält man sich an die relativ genaue χ^2-Approximation mit Nonzentralitätsparameter $\gamma^{(\lambda=0)}$, so hat die Funktion ein Maximum bei $\tau_1 = .652$, was 313 Itempaaren und 167 Einzelitems entspricht (statt 320 vs. 160, die in der Monte-Carlo-Studie verwendet wurden). Bei dieser Aufteilung resultiert eine maximale Effektstärke von $w^{(\lambda=0)} = .1364$ und eine Nonzentralitätsparameter von $\gamma^{(\lambda=0)} = 8.9285$. Das Signifikanzniveau und das Gesamt-N beeinflussen die absolute Höhe der Teststärke, nicht aber die Lage des τ_1-Optimums. Die Lage des Optimums kann jedoch durch die zugrunde gelegten H_1-Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ und in geringem Umfang auch durch die Wahl des λ-Parameters beeinflußt werden. $\tau_1 = .652$ ist also kein optimaler Wert für den Modelltest des EA-Modells schlechthin, sondern lediglich ein Optimum *relativ* zu gegebenem $\boldsymbol{\pi}$ unter H_1.

Was bedeutet dies für die praktische Versuchsplanung im Vorfeld von Modelltests oder Parametertests bei verbundenen multinomialen Modellen? Zunächst einmal wird deutlich, daß Faustregeln der Stichprobenaufteilung — wie beispielsweise die oftmals fast routinemäßig eingesetzte Gleichaufteilungsregel — zurückgewiesen werden müssen. Eine Gleichaufteilung kann durchaus optimal sein, z.B. bei Zwei-Gruppen-Homogenitätstests und Proportionenvergleichen (vgl. Meng & Chapman, 1966), aber sie muß eben nicht immer optimal sein. Man sollte sich daher vor einer Entscheidung über die Stichprobenaufteilung zunächst die statistischen Kernhypothesen überlegen, die zu testen sind. Zu jeder H_0 definiert man eine plausible H_1 (s.o.) und sucht dann durch iterative Anpassung der τ_k-Anteile diejenige Aufteilung eines Gesamtstichprobenumfangs N, die für das betreffende H_0-H_1-Paar einen maximalen Nonzentralitätsparameter $\gamma^{(\lambda)}$ und damit (zumindest approximativ) maximale Teststärke garantiert. Rechnerisch verfährt man wieder so, daß man zu gegebenem N, τ_k und $\boldsymbol{\pi}$ die erwarteten Häufigkeiten unter H_1 berechnet und dafür die gewünschte Minimum-PD^{λ}-„Statistik" zum H_0-Modell bestimmt. Natürlich wird die optimale Wahl der τ_k im Regelfall nicht eindeutig sein, wenn mehrere Hypothesen zu testen sind. In diesem Fall muß man einen geeigneten Kompromiß zwischen den verschiedenen optimalen Aufteilun-

gen suchen, wobei „geeignet" bei Bedarf auch noch präzisiert werden kann (z.B. Maximierung des durchschnittlichen Nonzentralitätsparameters oder des Median-Nonzentralitätsparameters).

Maßnahme (c): Optimierung von Parameterwerten. Bei der Prüfung von Parameterhypothesen interessieren gewöhnlich Unterschiede zwischen wenigen (im Extremfall nur zwei) Parametern des Gesamtmodells, während zu den anderen Modellparametern keine Vorhersagen vorliegen. Dennoch beeinflussen diese „irrelevanten Parameter" den H_1-Parametervektor $\boldsymbol{\pi}$ und damit die Effektstärke und somit auch die Power des interessierenden Parametertests. Was liegt also näher, als versuchsplanerische Maßnahmen zu ergreifen, welche die „irrelevanten Parameter" in dem Sinne optimieren, daß die Teststärke für den interessierenden Parametertest möglichst groß wird? Praktisch sieht das so aus, daß man den „Ästen" eines MVB-Modells, in denen die interessierenden Parameter liegen, durch günstige Wahl der Werte „irrelevanter Parameter" möglichst hohe Astwahrscheinlichkeiten und damit möglichst hohe erwartete Häufigkeiten verschaffen muß. Wenn man also etwa die Hypothese testen möchte, daß der Antwortbias b_{2ht} im 2HT-Modell der Rekognition (vgl. Abschnitt 3.5.2) zwischen zwei experimentellen Bedingungen variiert, tut man im Sinne der Teststärkeoptimierung gut daran, erschwerte Rekognitionsbedingungen (z.B. mit Hilfe eines langen Retentionsintervalls) herzustellen, damit der in diesem Kontext irrelevante Rekognitionsparameter r_{2ht} möglichst gering ausfällt. Will man dagegen im Rahmen des Enkodierung-Abruf-Modells (Enkodierung-Abruf-Modell, vgl. Abschnitt 3.5.1) testen, ob sich der Cluster-Abrufparameter r zwischen zwei experimentellen Bedingungen unterscheidet, so sind Maßnahmen hilfreich, die den Cluster-Enkodierungsparameter c maximieren (z.B. Verwendung semantisch sehr stark assoziierter Itempaare).

Wie sich verschiedene Werte der irrelevanten Parameter auf die Stärke des Tests zu den interessierenden Parametern auswirken, kann über die beschriebene Methode der Generierung von erwarteten Häufigkeiten zum komplett spezifizierten H1-Modell herausgefunden werden. Im allgemeinen wird man in jedem Anwendungsfall verschiedene Parameterhypothesen testen, so daß die Möglichkeit besteht, daß es keine generell besten (teststärkeoptimierenden) Werte der irrelevanten Parameter gibt. Wie schon im Falle der optimalen τ_k-Wahl muß man dann einen geeigneten Kompromiß herbeiführen.

Maßnahme (d): Differenzentest statt Modelltest. Neben der optimalen Aufteilung des Gesamtstichprobenumfangs ist der optimale Einsatz von χ^2-Differenzentests eine sehr effektive und von unerwünschten Nebenwirkungen weitgehend freie Methode der Teststärkeoptimierung. Daß χ^2-Differenzentests eines restringierten Modells teststärker als globale Modelltests sind, wenn die Gültigkeit des Ausgangsmodells vorausgesetzt werden kann, wurde in Abschnitt 4.3.2.2 be-

reits erwähnt. Aber wie groß ist der Teststärkegewinn in einem konkreten Anwendungsfall genau und wovon hängt die Größe des Teststärkegewinns ab?

Beim Enkodierung-Abruf-Modell könnte man als restringiertes Modell beispielsweise ein Modell mit fixierter Cluster-Enkodierungswahrscheinlichkeit betrachten, sagen wir das Modell zur H_0: $c = .30 \wedge u = a$. Der globale Test dieses restringierten Modells hätte $df = M - S' = 4 - 2 = 2$, weil nur noch zwei freie Modellparameter (u und r) geschätzt werden. Der Differenzen-χ^2-Test hätte dagegen $df = S - S' = 3 - 2 = 1$. Wie groß sind die Teststärken beider Tests, wenn als H_1 ein gültiges Enkodierung-Abruf-Modell mit Parametern $c = r = u = a = .50$ betrachtet wird und $N_1 = 160$ Itempaare sowie $N_2 = 80$ Einzelitems vorliegen? Verwendet man G^2 als Modelltest und die auf dem Nonzentralitätsparameter $\gamma^{(\lambda=0)}$ basierende Teststärkeapproximation, so ergibt sich für beide Tests die gleiche Effektgröße $w^{(\lambda=0)} = .1895$ und natürlich auch der gleiche Nonzentralitätsparameter $\gamma^{(\lambda=0)} = 8.6219$. Dennoch resultieren wegen der unterschiedlichen Freiheitsgrade unterschiedliche Teststärken. Wie man mit GPOWER (Erdfelder et al., 1996) berechnen kann, hat der globale Modelltest bei $\alpha = .05$ die approximative Teststärke $Pr(\chi^2(\gamma^{(0)} = 8.62)_{(df=2)} \geq 5.99) = .75$, während der Differenzentest auf immerhin $Pr(\chi^2(\gamma^{(0)} = 8.62)_{(df=1)} \geq 3.84) = .84$ kommt. Auch wenn der Unterschied nicht überwältigend ist, ist er doch deutlich größer als Teststärkeunterschiede, die man durch geeignete Wahl von λ erreichen kann. Hinzu kommt, daß der Teststärkegewinn des Differenzentests relativ zum globalen Modelltest im gegebenen Fall ungewöhnlich gering ist, weil sich beide Tests nur um einen Freiheitsgrad unterscheiden. Generell gilt

$$Pr(\chi^2(\gamma)_{(df)} \geq c_{\alpha,df}) > Pr(\chi^2(\gamma)_{(df')} \geq c_{\alpha,df'}) \Leftrightarrow df < df' \qquad (4.3.3.6\text{-}1)$$

für alle möglichen Nonzentralitätsparameter $\gamma \in \mathbb{R}^+$ und α–Niveaus, wobei $c_{\alpha,df}$ den kritischen Wert der $\chi^2_{(df)}$-Verteilung zum Niveau α bezeichnet. Der Unterschied der beiden Überschreitungswahrscheinlichkeiten auf der linken Seite von (4.3.3.6-1) wächst für gegebenes γ und α monoton mit der Differenz $df' - df$. Hat der globale Modelltest also sehr viele Freiheitsgrade und der des Differenzentests sehr wenige, resultieren deutlich größe Teststärkegewinne des Differenzentests relativ zum globalen Test als im o.g. Beispielfall.

4.3.4 Die Prüfung von Omnibus-Nullhypothesen

Häufig bezieht sich eine Modellgeltungs- oder Parameterhypothese nicht auf einen einzige Stichprobe, sondern simultan auf mehrere unabhängige Datensätze, die beispielsweise aus verschiedenen Versuchsgruppen eines komplexen Experiments oder aus Replikationen eines Experiments anhand unabhängiger Vpn-Stichproben hervorgegangen sein können (vgl. etwa das Beispiel in Abschnitt 6.4.2). Wenn n

unabhängige Datensätze vorliegen, interessiert gewöhnlich der Test der Omnibus-Nullhypothese $H_0^{(o)}$, daß für jede zugrundeliegende Grundgesamtheit *j* die korrespondierende Nullhypothese $H_0^{(j)}$ gilt,

$$H_0^{(o)} \Leftrightarrow H_0^{(1)} \wedge H_0^{(2)} \wedge \ ... \ \wedge H_0^{(j)} \wedge \ ... \ \wedge H_0^{(n)} \ , \qquad (4.3.4\text{-}1)$$

gegen die Alternativhypothese $H_1^{(o)}$, daß die Modellgeltungs- bzw. Parameternullhypothese in mindestens einem Fall falsch ist:

$$H_1^{(o)} \Leftrightarrow H_1^{(1)} \vee H_1^{(2)} \vee \ ... \ \vee H_1^{(j)} \vee \ ... \ \vee H_1^{(n)} \ . \qquad (4.3.4\text{-}2)$$

Die Prüfung von Omnibus-Nullhypothesen zu multinomialen Modellen kann prinzipiell auf zweierlei Weise erfolgen: (a) durch multiple Signifikanztests der einzelnen Nullhypothesen oder (b) durch einen simultanen Test (sog. Omnibustest) aller Nullhypothesen. Bei multiplen Signifikanztests wird $H_0^{(o)}$ nur dann beibehalten, wenn keiner der *n* Einzeltests signifikant ausfällt; andernfalls wird sie verworfen (Erdfelder, 1984; Hager, 1987; Westermann & Hager, 1986). Wegen der Kumulierung des α-Fehlers bei multiplen Tests ist der α-Fehler zu adjustieren. Wird pro Einzeltest ein Signifikanzniveau α zugrunde gelegt, so ist die Wahrscheinlichkeit, bei *n* unabhängigen Tests mindestens einen α-Fehler zu begehen, gerade

$$\alpha^{(o)} = 1 - (1 - \alpha)^n \ . \qquad (4.3.4\text{-}3)$$

Man wählt daher zu einem vorgegebenen Gesamt-$\alpha^{(o)}$-Niveau für die Einzeltests das adjustierte Niveau

$$\alpha = 1 - \sqrt[n]{(1 - \alpha^{(o)})} \ . \qquad (4.3.4\text{-}4)$$

Konkret stellt also beispielsweise bei $n = 5$ Tests $\alpha = .002$ ein globales $\alpha^{(o)} = .01$, $\alpha = .01$ ein globales $\alpha^{(o)} = .05$ und $\alpha = .021$ ein globales $\alpha^{(o)} = .1$ sicher.

Eine Alternative hierzu sind Omnibustests. Bei Omnibustests konjunktiv verknüpfter Nullhypothesen zu multinomialen Modellen macht man sich die bekannte Eigenschaft von χ^2-Verteilungen zunutze, daß die Summe zweier unabhängiger χ^2-verteilter Zufallsvariablen $\chi^2_{(df1)}$ und $\chi^2_{(df2)}$ selbst wieder χ^2-verteilt ist, wobei sich die Freiheitsgrade ebenfalls addieren (z.B. Hays, 1973, S. 436; Johnson & Kotz, 1970b, S. 135):

$$\chi^2_{(df_1)} + \chi^2_{(df_2)} = \chi^2_{(df_1 + df_2)} \qquad (4.3.4\text{-}5)$$

Folglich kann man die Nullhypothese (4.3.4-1) ganz einfach dadurch testen, daß man zunächst PD^{λ}-Anpassungsstatistiken bzw. G^2-Differenzenstatistiken zu jedem Datensatz isoliert bestimmt. Unter $H_0^{(o)}$ sind alle Anpassungsstatistiken $PD^{\lambda}_{(j)}$ asymptotisch zentral $\chi^2_{(dfj)}$-verteilt und deren Summe ist — sofern stochastische

Unabhängigkeit vorausgesetzt werden kann — wegen der Additivitätseigenschaft (4.3.4-3) ebenfalls asymptotisch χ^2-verteilt. Die PD^λ-Summenstatistik kann also anhand der χ^2-Verteilung mit $df = df_1 + df_2 + ... + df_n$ in der üblichen Weise auf Signifikanz beurteilt werden.

Tatsächlich kann die in Abschnitt 4.2.2 eingeführte PD^λ-Anpassungsstatistik (4.2.2-1) für ein verbundenes multinomiales Modell mit K Gruppen auch als Summe von K Anpassungsstatistiken aufgefaßt werden, die getrennt für die Gruppen zu bestimmen sind:

$$PD^\lambda(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{k=1}^{K} PD^\lambda{}_{(k)}(\boldsymbol{\theta};\boldsymbol{y}), \text{ mit}$$
$$PD^\lambda{}_{(k)}(\boldsymbol{\theta};\boldsymbol{y}) = \frac{2}{\lambda(\lambda+1)} \cdot \sum_{j=1}^{J_k} y_{k,j}\left[\left(\frac{y_{k,j}}{N_k \cdot p_{k,j}(\boldsymbol{\theta})}\right)^\lambda - 1\right]. \qquad (4.3.4\text{-}4)$$

Es handelt sich also um einen Spezialfall des oben geschilderten Omnibustests. Der Omnibustest setzt jedoch im Gegensatz zum (4.3.4-4) nicht voraus, daß alle Datensätze als Teildatensätze eines Gesamtmodells aufzufassen sind. Er verlangt auch nicht, daß immer dasselbe Modell an verschiedene Datensätze angepaßt wird. Es kann sich prinzipiell um *beliebige* Nullhypothesen zu multinomialen Modellen handeln, die simultan evaluiert werden sollen. Die einzige kritische Bedingung für die Adäquatheit des Omnibustests ist die Unabhängigkeit der zugrundeliegenden Datensätze, genauer: die stochastische Unabhängigkeit der χ^2-Modellanpassungsstatistiken zu den Datensätzen. Dies kann z.B. durch Verwendung unterschiedlicher Untersuchungsmaterialien und Vpn sichergestellt werden. Beziehen sich dagegen mehrere Nullhypothesen auf den *gleichen* Datensatz — was z.B. bei mehrfachen Tests hierarchischer Submodelle zu einem gemeinsamen Obermodell der Fall ist — so sind die Prüfstatistiken nicht unabhängig, und der o.g. Omnibustest ist folglich nicht gerechtfertigt. Es bleibt dann nur der Weg über multiple Tests der einzelnen Hypothesen, wobei jedoch — anders als in den Gleichungen (4.3.4-3) und (4.3.4-4) — nicht von unabhängigen Tests ausgegangen werden kann. Es empfiehlt sich dann eine gewöhnliche Bonferroni-Adjustierung gemäß $\alpha = \alpha^{(o)} / n$ (vgl. Lienert, 1973).

Wenn von unabhängigen Tests ausgegangen werden kann, stellt sich die Frage, welcher der beiden skizzierten Wege — multiple Einzeltests versus Omnibustest — zu empfehlen ist. Da beide Methoden das $\alpha^{(o)}$-Niveau einhalten, sollte man die Entscheidung an der Teststärke der Methoden festmachen. Die Teststärke des Omnibustests für n Datensätze läßt sich auf die gleiche Weise approximieren, wie dies in Abschnitt 4.3.3.1 für verbundene multinomiale Modelle mit K Gruppen besprochen wurde. Bei der Bestimmung des Gesamt-Effektstärkemaßes w muß also die Aufteilung des Gesamtstichprobenumfangs auf die einzelnen Datensätze

gemäß Gleichung (4.3.3.1-3) berücksichtigt werden, sofern die Effektstärke w_j nicht für alle n Hypothesen gleich ist. Die Teststärke des multiplen Testverfahrens kann ganz einfach aus den Teststärken der unabhängigen Einzeltests hergeleitet werden, wenn man die Entscheidungsregel zwischen $H_0^{(o)}$ und $H_1^{(o)}$ berücksichtigt. Trifft bei jedem der n Einzeltests die $H_1^{(j)}$ zu und ist die Teststärke und folglich die β-Fehlerwahrscheinlichkeit bei jedem der Einzeltests gleich groß, so errechnet sich die Teststärke $(1-\beta^{(o)})$ für die gesamte Testprozedur gemäß

$$(1-\beta^{(o)}) = 1-\beta^{n} \,, \qquad (4.3.4\text{-}5)$$

denn nur wenn bei *jedem* der n Einzeltests ein β-Fehler gemacht wird, resultiert auch ein $\beta^{(o)}$-Fehler für die multiple Testprozedur. Gilt nur bei einem der n Einzeltests die Alternativhypothese, bei allen anderen $(n-1)$ dagegen die Nullhypothese, so ergibt sich die Teststärke des multiplen Testverfahrens wie folgt:

$$(1-\beta^{(o)}) = 1-(1-\alpha)^{(n-1)} \cdot \beta \,. \qquad (4.3.4\text{-}6)$$

Hierbei ist für α natürlich grundsätzlich das adjustierte α-Niveau der Einzeltests anzusetzen.

Tabelle 4.3:
Teststärke des multiplen Testverfahrens und des Omnibustests zur Prüfung der Konjunktion von n Nullhypothesen via χ^2-Tests, und zwar für unterschiedliches n, unterschiedliche Effektstärken in den n Grundgesamtheiten und $df = 1$ sowie $df = 5$ pro Einzeltest. Der Stichprobenumfang für jeden der n Datensätze ist jeweils auf $N = 100$ und das $\alpha^{(o)}$-Niveau für beide Testprozeduren auf .05 festgelegt.

		multiple Einzeltests			Omnibustest		
n	$w_1, w_2, ..., w_n$	α	$1-\beta^{(o)}$ ($df = 1$)	$1-\beta^{(o)}$ ($df = 5$)	w	$1-\beta^{(o)}$ ($df = 1$)	$1-\beta^{(o)}$ ($df = 5$)
2	.10, .10, ..., .10	.0253	.2056	.1098	.10	.2255	.1208
5	.10, .10, ..., .10	.0102	.2602	.1261	.10	.3627	.1744
10	.10, .10, ..., .10	.0051	.3076	.1403	.10	.5424	.2503
2	.30, .30, ..., .30	.0253	.9504	.7656	.30	.9745	.8478
5	.30, .30, ..., .30	.0102	.9959	.9137	.30	.9999	.9928
10	.30, .30, ..., .30	.0051	.9998	.9735	.30	1.000	1.000
2	.00, .00, ..., .10	.0253	.1313	.0803	.0707	.1327	.0820
5	.00, .00, ..., .10	.0102	.0963	.0657	.0447	.0985	.0819
10	.00, .00, ..., .10	.0051	.0795	.0594	.0316	.0819	.0623
2	.00, .00, ..., .30	.0253	.7829	.5281	.2121	.2255	.4895
5	.00, .00, ..., .30	.0102	.6802	.4119	.1342	.3627	.3204
10	.00, .00, ..., .30	.0051	.5985	.3358	.0949	.5424	.2234

Tabelle 4.3 zeigt, daß es leider keine allgemeingültige Antwort auf die Frage nach dem teststärkeren Verfahren gibt. In der oberen Hälfte der Tabelle sieht man, daß der Omnibustest immer teststärker als die multiple Testprozedur ist, wenn in allen n Grundgesamtheiten die $H_1^{(j)}$ mit gleichen Effektstärken $w_1 = w_2 = ... = w_n = w$ gilt. Man kann dies als „diffuse Nonzentralitätsstruktur" bezeichnen, wenn man im Blick behält, daß der Begriff hier eine etwas andere Bedeutung hat als in Abschnitt 4.3.3.6. Analog kann man von einer „konzentrierten Nonzentralitätsstruktur" sprechen, wenn die Verletzung der $H_0^{(o)}$ auf eine einzige Grundgesamtheit beschränkt ist. Die untere Tabellenhälfte macht deutlich, daß sich in diesem Fall der Teststärkeunterschied abschwächt und bei großer Effektstärke w_j im „Ausreißerdatensatz" sogar umkehrt, so daß nunmehr die multiple Testprozedur teststärker ist (s. unteres Viertel der Tabelle). Der auf nur einen Datensatz beschränkte große Effekt w_j wird durch die simultane Analyse aller Datensätze gewissermaßen „verwischt", so daß die Gesamt-Effektstärke w nicht ausreicht, um dem Omnibustest eine hinreichende Teststärke zu geben.

Dieses Ergebnismuster erinnert an die Teststärkeunterschiede der verschiedenen Prüfstatistiken zur multivariaten Varianzanalyse (MANOVA): Roys „largest root criterion", das nur für den effektstärksten Mittelwertskontrast sensitiv ist, ist bei konzentrieren Nonzentralitätsstrukturen optimal, während das Pillai-Bartlett-V-Kriterium, welches auf Effekte in allen Mittelwertskontrasten reagiert, bei diffusen Nonzentralitätsstrukturen teststärker ist (Bredenkamp & Erdfelder, 1985; Olson, 1976; Stevens, 1980). Bei der Prüfung von MANOVA-Nullhypothesen kann man davon ausgehen, daß diffuse Nonzentralitätsstrukturen der Regelfall sind, was zur Empfehlung führt, routinemäßig das Pillai-Bartlett-Kriterium zu wählen (vgl. Bredenkamp, 1980; Olson, 1976). Ähnliches gilt für die Prüfung von Omnibus-Nullhypothesen zu multinomialen Modellen. Denn wenn man von vornherein die Vermutung hat, daß die Ungültigkeit der Nullhypothese auf eine einzige Grundgesamtheit beschränkt ist, wird man zunächst diese Gruppe untersuchen und sich nicht der Mühe unterziehen, mehrere Datensätze zu erheben. Wenn mehrere Datensätze erhoben werden, dann geschieht dies also typischerweise mit der Erwartung, daß das Modell entweder in allen Grundgesamtheit (annähernd) zutrifft oder aber in (fast) allen untersuchten Grundgesamtheiten in ungefähr gleichem Maße verletzt ist. Letztere Erwartung entspricht einer diffusen Nonzentralitätsstruktur, so daß die Durchführung des Omnibustests klar indiziert ist. Selbst wenn sich die Nullhypothesenverletzung unerwarteterweise auf einen Datensatz konzentrieren sollte, ist bei kleinen Effektstärken immer noch ein Vorteil des Omnibustests gegeben. Bei großen Effektstärken ist der Teststärkeverlust relativ zur multiplen Testprozedur gering (vgl. Tabelle 4.3). Man riskiert also nicht viel, wenn man generell auf den Omnibustest vertraut.

Nicht selten liegen Verbundhypothesen oder „Metahypothesen" (Erdfelder, 1984) zu $n = n_1+n_2$ unabhängigen Datensätze vor, die einer Konjunktion von n_1 Nullhypothesen und n_2 Alternativhypothesen entsprechen. In solchen Fällen kommt nur eine multiple Testprozedur in Frage, wobei die n_1 Nullhypothesen aus den oben genannten Gründen im Regelfall durch einen Omnibustest simultan geprüft werden sollten. Zusätzlich müssen die n_2 Alternativhypothesen durch Einzeltests abgeprüft werden. Die Entscheidungsregel lautet, daß die Verbundhypothese genau dann beizubehalten ist, wenn die Omnibus-Teststatistik zu den n_1 Nullhypothesen nicht in den Rejektionsbereich fällt, die n_2 Teststatistiken zu den Alternativhypothesen dagegen schon. Die Fehlerwahrscheinlichkeiten für eine solche komplexe Entscheidungsregel ergeben sich aus den Fehlerwahrscheinlichkeiten der Einzeltests. Wenn erwünscht, können die α- und β-Fehler der Einzeltests so adjustiert werden, daß a priori festgelegte Typ-1- und Typ-2-Fehlerrisiken für die gesamte multiple Testprozedur eingehalten werden. Da darauf an anderer Stelle ausführlich eingegangen wurde, kann dies hier ausgeklammert werden (z.B. Erdfelder, 1984; Hager, 1987; Westermann & Hager, 1986).

Beachtet werden sollte, daß es bei multiplen Tests nicht immer angezeigt ist, generell dasselbe α-Niveau für alle Tests zugrundezulegen. Dies gilt insbesondere dann, wenn sich die Tests auf denselben Datensatz beziehen. Sind beispielsweise Parametergleichsetzungshypothesen zu prüfen, die auf verschiedenen Ebenen eines Verarbeitungsbaummodells angesiedelt sind, so muß damit gerechnet werden, daß die Teststärke für Parameterunterschiede auf höheren Ebenen des Ereignisbaumes geringer ausfällt als auf den unteren Ebenen. Die Auswirkung eines bestimmten Parameterunterschieds auf das Effektstärkemaß w ist nämlich um so eher vernachlässigbar, je mehr irrelevante Parameter dem kritischen Parameter im Baummodell vorgeordnet sind. Dies ist auch daran erkennbar, daß die Standardschätzfehler und Konfidenzintervalle typischerweise mit der Ebene des Baumes, auf der die zugehörigen Parameter angesiedelt sind, wachsen.

Ein partieller Ausgleich für diese im allgemeinen unerwünschten Teststärkeunterschiede kann bei unabhängigen Datensätzen dadurch erreicht werden, daß man unterschiedliche Stichprobenumfänge für die verschiedenen Datensätze wählt. Beziehen sich die Tests dagegen auf denselben Datensatz oder sind die Stichprobenumfänge verschiedener Datensätze nicht beeinflußbar, so bleibt nur die Möglichkeit, das α-Niveau so zu anzupassen, daß das β-Fehlerrisiko für alle Tests in einem vergleichbaren Bereich liegt (vgl. Erdfelder et al., 1996). Betrachten wir als Beispiel die Anwendung des Enkodierung-Abruf-Modells (Batchelder & Riefer, 1980; 1986) auf die Frage nach Altersunterschieden bei der Speicherung und beim Abruf von Informationen aus dem Gedächtnis (vgl. Abschnitt 3.5.1). Hier kann es durchaus sinnvoll sein, die Nullhypothese nicht vorhandener Altersunterschiede bei der Cluster-Speicherung (H_0: $c_{alt} = c_{jung}$) auf dem Niveau $\alpha = .05$, die

Nullhypothese gleicher Abrufparameter (H_0: $r_{alt} = r_{jung}$) dagegen auf dem Niveau $\alpha = .10$ zu testen, und zwar vor allem dann, wenn die *c*-Parameter im unteren Bereich liegen. Welche Typ-1-Fehlerwahrscheinlichkeiten im Sinne einer Balancierung des α- und β-Fehlerrisikos für gegebene Parameterunterschiedshypothesen (z.B. H_1: $c_{jung} - c_{alt} = .10$ und H_1: $r_{jung} - r_{alt} = .10$) optimal sind, kann durch vergleichende Teststärkeanalysen (wie im letzten Abschnitt beschrieben) entschieden werden.

4.3.5 Welche Prüfstatistik sollte man wählen?

Die Antwort auf die Frage nach der besten Prüfstatistik fällt eher leichter als die Frage nach der optimalen Schätzmethode (vgl. Abschnitt 4.2.4). Viele Argumente sprechen für die Wahl der Log-Likelihood-Quotienten-Statistik G^2 als Standard-Prüfstatistik. Insbesondere ist die Verfügbarkeit teststarker und einfach berechenbarer ΔG^2-Differenzentests zur Prüfung spezieller Parameterhypothesen hervorzuheben. Derartige Tests werden in der Praxis sehr häufig benötigt. Hinzu kommt die optimale Bahadur-Effizienz sowie die vergleichsweise hohe Teststärke bei diffusen Nonzentralitätsstrukturen, welche bei Modelltests parametrisierter multinomialer Modelle vermutlich eher vorkommen als konzentrierte Nonzentralitätsstrukturen. Auf der negativen Seite ist lediglich die in der Regel schlechtere χ^2-Approximierbarkeit der exakten zentralen und nonzentralen Verteilungen von G^2 bei kleinen Stichprobenumfängen zu vermerken. Doch dieses Problem fällt nicht stark ins Gewicht, da (a) Korrekturformeln verfügbar sind und (b) kleine Stichprobenumfänge ohnehin vermieden werden sollten, um ausreichende Teststärke zu gewährleisten.

Auch hier gilt natürlich, daß es keine Faustregel ohne Ausnahme gibt. In Einzelfällen — z.B. bei mutmaßlich konzentrierten Nonzentralitätsstrukturen — kann man an andere Minimum-PD^λ-Statistiken denken, wobei man sich allerdings auch in diesem Fall auf den Bereich $\lambda \in (-1, 2]$ beschränken sollte. Außerhalb dieses Bereichs ist die χ^2-Approximation unbefriedigend und es liegt zudem eine große Sensitivität gegenüber Ausreißerzellen vor (Cressie & Read, 1984; Read & Cressie, 1988, S. 63). Je nach H_0-H_1-Paar werden unterschiedliche λ–Parameter im mittleren Bereich optimal sein. Manche Autoren halten $\lambda = 2/3$ für den bestmöglichen Kompromiß bei Routineanwendungen (z.B. Read & Cressie, 1988, S. 63 & 96f; Garcia-Pérez, 1994). Die Cressie-Read-Statistik hat einerseits den Vorteil, auch schon im Falle kleiner Stichprobenumfänge gut durch die zentrale χ^2-Verteilung approximierbar zu sein, so daß man auf Korrekturformeln generell verzichten kann; andererseits ist sie auch unter Teststärkegesichtspunkten ein guter Kompromiß, wenn man a priori nichts über die Nonzentralitätsstruktur weiß (Cressie & Read, 1984). Allerdings ist mehr als fraglich, ob das für G^2 sprechende Argument

einfach berechenbarer G^2-Differenzentests durch diese Argumente aufgewogen wird, zumal auch G^2 bei Routineanwendungen unter Teststärkegesichtspunkten nicht schlecht abschneidet.

Wichtiger noch als die Wahl des λ-Parameters der Prüfstatistik ist eine teststärkeoptimierende Stichprobenaufteilung auf verschiedene Versuchsbedingungen und der Einsatz von χ^2-Differenzentests, wann immer dies durch eine spezielle Hypothese nahegelegt wird (vgl. Abschnitt 4.3.3.6). Das α-Niveau sollte niemals blind auf ein Standardniveau (z.B. $\alpha = .05$) festgelegt werden. Bei den in der Praxis nicht seltenen Stichprobenumfängen im Bereich $N > 1000$ kann mit Standardalphaniveaus schon für triviale Verletzungen der H_0 eine sehr große Teststärke verbunden sein. Sorgfältige Teststärkebetrachtungen sind deshalb nötig, um zu vernünftigen α-Festlegungen zu kommen. Im Regelfall sollte man das α-Niveau so wählen, daß für das realisierte N und eine „gerade bedeutsame Verletzung" der H_0 die β-Fehlerwahrscheinlichkeit etwa so groß wie die α-Fehlerwahrscheinlichkeit ist.

Prinzipiell gibt es keinen zwingenden Grund, die Wahl der Prüfstatistik an die Wahl der Parameterschätzmethode zu koppeln. Die verschiedenen Minimum-PD$^\lambda$-Prüfstatistiken verlangen lediglich BAN-geschätzte Parameter; der konkrete λ-Wert in der Minimierungsfunktion ist unerheblich (Read & Cressie, 1988, S. 32). Wer sich also beispielsweise für G^2 als Prüfstatistik entscheidet, muß nicht unbedingt die ML-Methode der Parameterschätzung wählen. Garcia-Pérez (1994) fand in seiner Simulationsstudie sogar Hinweise darauf, daß die Verwendung von Freeman-Tukey-Schätzern (identisch mit Minimum-PD$^{\lambda=-1/2}$-Schätzern) anstelle von ML-Schätzern zu geringfügig besseren Übereinstimmungen zwischen nominellem und faktischen α-Niveaus bei Log-Likelihood-Quotienten-Tests führte. Ähnliches berichtete er auch für die Cressie-Read- und die Pearson-Statistik. Allerdings gibt es kein mir bekanntes statistisches Argument, welches die Vermutung rechtfertigen könnte, daß dieser geringfügige Vorteil einer λ-Diskrepanz zwischen Prüfstatistik und Schätzmethode auf andere Anwendungsfälle generalisierbar ist. In Anbetracht der Tatsache, daß die Verwendung unterschiedlicher λ-Werte bei der Parameterschätzung und der Prüfstatistikberechnung einen beträchtlich erhöhten Zeitaufwand bei der Datenanalyse zur Folge hat, sollte man davon eher Abstand nehmen. Viele Argumente sprechen für ML-Schätzer (vgl. Abschnitt 4.2.4) und noch mehr Argumente für den Log-Likelihood-Quotienten-Test. Warum sollte man also nicht beides in zeitsparender Weise miteinander kombinieren?

Wird die statistische Entscheidung approximativ anhand der χ^2-Verteilung durchgeführt, müssen die Regularitätsbedingungen von Birch (1964) vorausgesetzt werden, egal welche Minimum-PD$^\lambda$-Statistik verwendet wird. Besteht an der Gültigkeit dieser Regularitätsbedingungen begründeter Zweifel, weil mindestens eine Parameterschätzung auf dem Rand des Parameterraumes liegt, oder ist der Ge-

samtstichprobenumfang zu klein, um auf asymptotische Approximationen setzen zu können, so kann man entweder auf informationstheoretische Maße der Anpassungsgüte zurückgreifen, auf die im folgenden Abschnitt 4.3.6 eingegangen wird, oder die statistische Entscheidung mit der Methode des parametrischen Bootstrap absichern (vgl. die Abschnitte 4.3.1 und 4.3.2.2).

Noch nicht sehr gut geklärt ist die Frage nach der optimalen Wahl der Anpassungsstatistik beim parametrischen Bootstrap. Klar ist bislang nur, daß die Wahl der Prüfstatistik beim Bootstrap durchaus Einfluß auf die Ergebnisse haben kann. Von Davier (1997) hat in einer Simulationsstudie zum Latent-Class-Modell und zum Mixed-Rasch-Modell (Rost, 1990) deutliche Unterschiede zwischen parametrischen Bootstraps für vier verschiedene Minimum-PD$^{\lambda}$-Anpassungsstatistiken gefunden, wenn die analysierten Kontingenztafeln extrem niedrige erwartete Häufigkeiten aufwiesen („sparse data"). Die Pearson- und die Cressie-Read-Statistik schnitten deutlich besser ab als die Log-Likelihood-Quotienten-Statistik und die Freeman-Tukey-Statistik (d.h. die Minimum-PD$^{\lambda=-1/2}$-Statistik). Faktische α- und β-Fehlerraten wichen in den beiden letztgenannten Fällen deutlich von den nominellen Werten ab, nicht jedoch bei den Prüfstatistiken mit $\lambda > 0$. Es spricht einiges dafür, daß dieses Problem auf Anwendungsfälle beschränkt ist, in denen sehr viele Beobachtungskategorien mit nur einer Beobachtung (d.h. $y_{k,j} = 1$) besetzt sind (von Davier, 1997, S. 45-47). Dieser Fall dürfte bei Anwendungen multinomialer Modelle in der kognitiven Psychologie sehr selten vorkommen. Ob das Problem auch bei größeren Zellenbesetzungen auftreten kann, muß durch zukünftige Monte-Carlo-Studien noch genauer untersucht werden. Es ist jedoch zu erwarten, daß Unterschiede zwischen Bootstrap-Verfahren für verschiedene Teststatistiken mit zunehmendem Stichprobenumfang tendenziell kleiner werden.

Bootstrap-Verfahren müssen im übrigen nicht mit einem Verzicht auf Teststärkebetrachtungen verbunden sein. Prinzipiell gibt es auch im Bereich der Teststärkekontrolle die Möglichkeit, von parametrischen Bootstrap-Verfahren Gebrauch zu machen (Beran, 1986). Hierzu erzeugt man eine Bootstrap-Stichprobenverteilung der verwendeten Teststatistik nicht unter dem Nullhypothesenmodell, sondern statt dessen unter dem Alternativhypothesenmodell, das die Grenze zur bedeutsamen Verletzung der H_0 markiert. Man zählt aus, wie häufig die Stichproben aus dieser nonzentralen Bootstrap-Verteilung den aus der zentralen (H_0-konformen) Bootstrap-Verteilung ermittelten kritischen Wert der Prüfstatistik überschreiten. Die relative Häufigkeit dieser Ereignisse ist ein konsistenter Schätzer für die Teststärke unter der betreffenden H_1 (vgl. Beran, 1986).

Gelegentlich kommt es vor, daß ein multinomiales Modell für sehr viele unterschiedliche Datensätze erstaunlich positive Anpassungsergebnisse erzielt. Dann liegt der Verdacht nahe, daß das betreffende Modell sehr unspezifische Verteilungsvorhersagen macht, so daß es nahezu beliebige Datensätze gut „vorhersa-

gen" kann, auch wenn es sich nicht um ein saturiertes Modell handelt. Unter diesen Umständen ist es hilfreich, die Nullhypothese testen zu können, daß das Modell einen gegebenen empirischen Datensatz nicht besser als einen Zufallsdatensatz „erklärt". Kann man diese Nullhypothese beibehalten, so ist das Modell trotz guter Anpassungsergebnisse relativ uninteressant, da es zu empirischen Daten nachweislich nicht besser als zu Zufallsdatenmustern paßt. Kann die Nullhypothese dagegen verworfen werden, so ist damit gezeigt, daß das Modell durchaus spezifische, nichttriviale Vorhersagen bezüglich zulässiger Multinomialverteilungen macht, denen Zufallsdaten nicht ohne weiteres genügen.

Die Umsetzung eines solchen Tests verlangt zunächst einmal eine Präzisierung des Begriffs „Zufallsdaten". Erdfelder und Buchner (1998a) haben einen konditionalen Zufallsdatenbegriff vorgeschlagen, der an den beobachteten Häufigkeiten in einer Stichprobe festhält, jedoch die Bindung an die Beobachtungskategorien aufhebt. „Zufallsdaten" erhält man demnach, wenn man die beobachteten Häufigkeiten in einer Stichprobe den Beobachtungskategorien zufällig zuordnet. Relativ zu diesem Zufallsdatenbegriff ist die Prüfung der o.g. Nullhypothese leicht möglich. Man generiert ganz einfach viele (z.B. 1000) Zufallsdatensätze aus einer gegebenen Stichprobe und berechnet für jeden dieser Zufallsdatensätze die interessierende Modellanpassungsstatistik für das zu evaluierende Modell. So erhält man eine Monte-Carlo-Stichprobenverteilung der Anpassungsstatistik unter der Nullhypothese, daß das Modell an Zufallsdaten im geschilderten Sinne angepaßt wird. Liegt die Anpassungsstatistik für den empirisch wirklich beobachteten Datensatz im Bereich der oberen $(1-\alpha)\cdot 100\%$ Werte der Monte-Carlo-Stichprobenverteilung, so behält man die Nullhypothese bei, daß das Modell die empirischen Daten nicht besser als Zufallsdaten beschreibt. Liegt die Anpassungsstatistik dagegen im Bereich der unteren $\alpha\cdot 100\%$ der Stichprobenverteilung, so verwirft man die H_0 und entscheidet sich für die Alternativhypothese, daß die empirisch beobachteten Daten signifikant besser vorhergesagt werden als Zufallsdaten (vgl. Erdfelder & Buchner, 1998a, für eine konkrete Anwendung des Verfahrens).

Zu beachten ist, daß der geschilderte Test gegen die Zufallsnullhypothese zwar unter Umständen eine sinnvolle Ergänzung eines Modellanpassungstests ist, jedoch diesen nicht ersetzen kann. Für die Annahme eines Modells reicht der Nachweis nicht aus, daß das Modell empirische Daten besser als Zufallsdaten beschreibt. Gezeigt ist damit lediglich, daß es für die Anpassungsgüte eines MVB-Modells nicht belanglos ist, in welche Beobachtungskategorien man die Äste des Modells einmünden läßt. Wenn ein Modell diesen Test besteht, kann es natürlich immer noch ein schlechtes Modell in dem Sinne sein, daß die vom Modell vorhergesagte and die faktische Multinomialverteilung der Daten weit voneinander entfernt liegen.

4.3.6 Informationstheoretische Maße der Anpassungsgüte

Standardanwendungen parametrisierter multinomialer Modelle sind dadurch gekennzeichnet, daß verschiedene Modelle hinsichtlich der Anpassungsgüte an gegebene Daten zu vergleichen sind. Sofern es sich hierbei um Submodelle einer gemeinsamen hierarchischen Modellfamilie handelt, liegt es nahe, mit sukzessiven G^2-Differenzentests nach dem sparsamsten (parameterärmsten) Modell der Familie zu suchen, das die Daten noch adäquat abbildet. Allerdings ist die praktische Anwendung von G^2-Differenzentests nicht immer unproblematisch. Zunächst einmal ergeben sich Schwierigkeiten, wenn ein Submodell am Rand des Parameterraumes des übergeordneten Modells liegt. Bei MVB-Modellen ist das z.B. immer dann der Fall, wenn Parameter des übergeordneten Modells im untergeordneten Modell auf die Grenzwerte 0 oder 1 fixiert werden, was nicht selten vorkommt[47]. In solchen Fällen ist nicht mehr garantiert, daß die betreffende ΔG^2-Differenzenstatistik bei Gültigkeit des Submodells asymptotisch zentral χ^2-verteilt ist (vgl. Abschnitt 4.3.1). Die statistische Rechtfertigung für den asymptotischen Modellvergleichstest entfällt also.

Doch selbst wenn die Voraussetzungen für eine asymptotische χ^2-Verteilung der ΔG^2-Statistik erfüllt sind, ist die Anwendung von Differenzentests problematisch, sofern mehr als zwei Modelle zu vergleichen sind. Das Ergebnis des sukzessiven paarweisen Modellvergleichs kann nämlich durchaus davon abhängen, mit welchem Modellvergleich man beginnt und in welcher Reihenfolge man die restlichen Vergleiche durchführt. Da es normalerweise keinen Grund gibt, eine bestimmte Reihenfolge als verbindlich auszuzeichnen, können verschiedene Experimentatoren bei ausschließlicher Verwendung von G^2-Differenzentests zu unterschiedlichen Schlußfolgerungen über das beste Modell der Modellfamilie gelangen, auch wenn sie auf die gleichen empirischen Daten Bezug nehmen. Wir haben es hier gewissermaßen mit einem „Bedeutsamkeitsproblem vierter Art" zu tun, nämlich einem Bedeutsamkeitsproblem bei der statistischen Entscheidungsstrategie. Dies ist unbefriedigend.

Vollends unübersichtlich wird der Modellvergleich bei nichthierarchischen Modellen, die man nicht durch Parameterrestriktionen auseinander herleiten kann. G^2-Differenzentests lassen sich in diesen Fällen nicht durchführen. Ein deskriptiver Modellvergleich auf der Basis der globalen Modellanpassungsstatistiken liegt nahe, ist aber nur dann aussagekräftig, wenn die zu vergleichenden Modelle gleiche Parameterzahl aufweisen oder das bestpassende Modell zugleich das parameterärmste Modell ist (Wickens, 1998a, S. 542f.). Ist dagegen das bestpassende Mo-

[47] Beispielsweise ist das Ein-Hochschwellenmodell für Quellendiskriminationstests (Batchelder & Riefer, 1990) ein Submodell des Zwei-Hochschwellenmodells (Bayen et al., 1996), basierend auf der Restriktion $D_3 = 0$ (vgl. dazu Abschnitt 7.1.2). Modellvergleiche können daher nicht über asymptotische ΔG^2-Differenzentests erfolgen.

dell nicht das parameterärmste Modell — was der Regelfall sein dürfte —, so kann es immer sein, daß die hohe Anpassungsgüte nur eine scheinbare ist, d.h. in Wirklichkeit auf der Vielzahl freier Modellparameter beruht, mit denen sich nahezu beliebige Datenstrukturen gut anpassen lassen. Saturierte Modelle sind sicherlich der krasseste Fall, da sie immer „optimale" Anpassungsstatistiken liefern, auch dann, wenn eine inhaltliche Theorie zugrunde liegt, die gänzlich inadäquat ist.

Im Bereich der Statistik zum Allgemeinen Linearen Modell liegt ein strukturell ähnliches Problem vor. Das multiple R^2 als Maß der aufgeklärten Varianz bei Regressionsanalysen kann mit zunehmender Zahl freier Regressionsgewichte in einer Regressionsgleichung nur steigen, niemals fallen. Es ist deshalb als Maßstab für den Vergleich von Regressionsmodellen mit unterschiedlicher Parameterzahl ungeeignet. Als Ausweg aus dem Dilemma wurden sogenannte „shrinkage formulas" vorgeschlagen, welche darauf abzielen, die Höhe des empirischen R^2 um die Anzahl freier Parameter zu korrigieren (Carter, 1979). Auch wenn derartigen korrigierten R^2-Werten kein einheitliches Rationale zugrunde liegt und außerdem nicht davon ausgegangen werden kann, daß die Korrekturen erwartungstreue Schätzungen für den aufgeklärten Varianzanteil in der zugrundeliegenden Population liefern (Bredenkamp, 1980, S. 52), haben sie sich dennoch als nützliche Maße für deskriptive Modellvergleiche erwiesen. Diejenigen Regressionsmodelle, die optimale korrigierte R^2-Werte aufweisen, schneiden in Kreuzvalidierungsstudien oftmals besser ab als Modelle, die aufgrund zusätzlicher Prädiktoren zwar ein höheres R^2, jedoch ein niedrigeres korrigiertes R^2 aufweisen.

Kann dieser Ansatz auch im Bereich der multinomialen Modellierung fruchtbar gemacht werden? Tatsächlich wurden und werden Modellevaluationen auch in diesem Bereich gelegentlich anhand erklärter Varianzanteile vorgenommen. Man berechnet dann im Regelfall die einfache lineare Regression der beobachteten Kategoriehäufigkeiten auf die erwarteten Kategoriehäufigkeiten unter Annahme der Modellgültigkeit und verwendet die quadrierte Korrelation als Maß des „goodness of fit" (z.B. Rips, 1983; Yonelinas, 1999). Obwohl bislang meines Wissens noch nicht praktiziert, ließe sich dieser Ansatz prinzipiell so modifizieren, daß die Anzahl freier Modellparameter als Korrekturfaktor einbezogen wird. Allerdings ist dieses Vorgehen — unabhängig von der Einbeziehung eines Korrekturfaktors — sehr problematisch. Auch für multiple Regressionsanalysen sind R^2-Werte bzw. korrigierte R^2-Werte nur dann brauchbare Datenanpassungsmaße, wenn die Gültigkeit des linearen Regressionsmodells vorausgesetzt werden kann und lediglich die Anzahl der Prädiktoren in der Regressionsgleichung offen gelassen wird. Ist die lineare Modellgleichung ungültig, so können erklärte Varianzanteile erheblich in die Irre führen. Gilt beispielsweise statt eines linearen Modells $E(Y \mid X) = a + b \cdot X$ ein quadratisches Modell $E(Y \mid X) = a' + b' \cdot X^2$, so kann bei Anpassung des linearen Modells dennoch ein extrem hohes R^2 deutlich über .90 resultie-

ren, sofern alle Realisationen des Prädiktors X positiv sind. Umgekehrt kann sich auch bei Gültigkeit eines linearen Modells ein relativ niedriges R^2 ergeben, sofern die Fehlervarianz um die Regressionsgerade entsprechend groß ist. Mit anderen Worten: Ohne weitere Zusatzannahmen erlaubt die Höhe des R^2 keinen Rückschluß auf die Gültigkeit oder Ungültigkeit von Modellen. Ganz analoge Probleme ergeben sich, wenn man R^2-Werte zur Evaluation multinomialer Modelle heranzieht. Sie geben lediglich Auskunft über die *Approximierbarkeit* empirischer Kategoriehäufigkeiten durch die Modellgleichungen; diese kann auch dann gegeben sein, wenn das Modell selbst ungültig ist.[48] Die Modellgültigkeit kann nur evaluiert werden, wenn man statt erklärter Varianzanteile die nicht erklärte Datenvarianz — erfaßt durch G^2 oder ein anderes Modell-Daten-Divergenzmaß — in den Mittelpunkt des Interesses rückt.

Sind Modelle mit unterschiedlicher Anzahl freier Parameter zu evaluieren, liegt es nahe, einen zur R^2-Adjustierung analogen Ansatz auf ein Potenz-Divergenzmaß wie z.B. G^2 anzuwenden. Diese Idee ist nicht neu. Akaike (1973) hat vorgeschlagen, die mit –2 multiplizierte mittlere erwartete Log-Likelihood gegebener Daten unter einem Modell als Maß für die Anpassungsgüte zu verwenden. Je kleiner dieses Maß ausfällt, um so näher liegt die vom Modell vorhergesagte Verteilung der tatsächlichen Verteilung. „Näher" ist hierbei im Sinne der gerichteten Kullback-Divergenz zu verstehen (vgl. dazu Read & Cressie, 1988, S. 124f.). Liegen S freie Parameter im zu evaluierenden Modell vor, so kann man das sogenannte Akaike Information Criterion (AIC)

$$\mathrm{AIC} = -2 \cdot \ln(L(\hat{\boldsymbol{\theta}}; \boldsymbol{y})) + 2 \cdot S \qquad (4.3.6\text{-}1)$$

als stichprobenbasiertes Maß der Modellgüte verwenden, denn AIC / N ist ein erwartungstreuer Schätzer für den Akaike-Modellgüteparameter (Akaike, 1973).

Das AIC-Maß kann wie das korrigierte R^2 als parameteradjustierte Anpassungsstatistik aufgefaßt werden. Der linke Summand in Gleichung (4.3.6-1) ist unschwer als lineare Funktion von G^2 zu erkennen (vgl. Gleichung (4.3.2.2-2)). Der rechte Summand kann als „Penalty-Funktion" für freie Modellparameter aufgefaßt werden. Mit jedem zusätzlichen Parameter im Modell muß der linke Summand (und damit G^2) um mindestens 2 fallen, damit das AIC-Maß nicht schlechter (d.h. nicht größer) wird. Umgekehrt kann die Elimination eines freien Parameters mit einem G^2-Anstieg kleiner oder gleich 2 einhergehen, ohne daß dadurch das AIC-Anpassungsmaß insgesamt schlechter wird.

[48] Ein Beispiel ist die Arbeit von Yonelinas (1999). Er stellt ein Modell für Rekognitionsexperimente vor, das bis zu 99% der Varianz in den Trefferraten und falschen Alarmraten erklären kann (vgl. dazu Abschnitt 7.1.3). Trotzdem nennen Glanzer und Mitarbeiter (1999) überzeugende Argumente gegen die Gültigkeit dieses Modells.

Der rechte Summand in Gleichung (4.3.6-1) ist der Korrekturterm, der das globale Anpassungsmaß um die Anzahl freier Parameter im Modell adjustiert. Von verschiedenen multinomialen Modellen zum gleichen Datensatz nennen wir dasjenige AIC-optimal, das den kleinsten AIC-Wert liefert. Natürlich können wir auf diese Weise auch ein bestimmtes H_0-Modell (M_0) mit einem saturierten H_1-Modell (M_1) vergleichen. Das entspricht exakt der Fragestellung eines globalen Modellanpassungstests zu M_0 (vgl. Abschnitt 4.3.3.2). Ein Vergleich von M_0 und M_1 auf der Basis des AIC-Maßes würde genau dann zur Annahme von M_0 führen, wenn die AIC-Differenz

$$\begin{aligned} &\mathrm{AIC}(M_0) - \mathrm{AIC}(M_1) = (-2 \cdot \ln(L(M_0)) + 2 \cdot S) - (-2 \cdot \ln(L(M_1)) + 2 \cdot M) = \\ &= -2 \cdot \ln(L(M_0) / L(M_1)) - 2 \cdot (M - S) = G^2(M_0) - 2 \cdot df \end{aligned} \qquad (4.3.6\text{-}2)$$

kleiner oder gleich null ist. Ansonsten wird M_0 verworfen. Praktisch läuft das darauf hinaus, die G^2-Modellanpassungsstatistik nicht — wie beim gewöhnlichen Log-Likelihood-Quotienten-Test — mit dem entsprechenden kritischen Wert c_α der $\chi^2_{(M-S)}$-Verteilung zu vergleichen, sondern statt dessen immer mit dem Zweifachen des Erwartungswerts dieser Verteilung, nämlich $2 \cdot df$. Man kann zeigen, daß statistische Entscheidungen auf der Basis der AIC-Differenz gewöhnlichen Log-Likelihood-Quotienten-Tests zum Niveau $\alpha = .05$ nahezu äquivalent sind, wenn $df = M - S = 7$ (Read & Cressie, 1988, S. 127). Für $df < 7$ ist der „AIC-Differenzentest" liberaler ($\alpha_{AIC} > .05$), für $df > 7$ konservativer ($\alpha_{AIC} < .05$) als ein G^2-Modelltest zum Niveau $\alpha = .05$. Entscheidungstheoretisch betrachtet liegt also eine gewisse Voreingenommenheit zugunsten einfacher, parameterarmer Modelle vor, was jedoch eine durchaus erwünschte Eigenschaft des AIC-Maßes ist.

Allerdings hat das AIC-Maß zwei miteinander zusammenhängende Nachteile. Es ist zum einen nicht konsistent, führt also für $N \to \infty$ nicht zu einer korrekten statistischen Entscheidung zugunsten des zutreffenden Modells (Azencott & Dacunha-Castelle, 1986). Zum zweiten berücksichtigt es den Stichprobenumfang nicht, der in eine bestimmte G^2-Anpassungsstatistik eingeht. Somit werden bei Verwendung des AIC-Kriteriums bereits triviale Modellverletzung immer zu einer Verwerfung von M_0 führen, wenn man den Stichprobenumfang nur groß genug wählt. Umgekehrt werden substantielle Modellverletzungen von M_0 nur mit geringer Wahrscheinlichkeit entdeckt, wenn der Stichprobenumfang klein ist. Unter Teststärkegesichtspunkten ist das AIC-Maß also äußerst problematisch.

Beide Nachteile werden bei Verwendung des von Schwarz (1978) vorgeschlagen Bayesian Information Criterion (BIC) vermieden. Das BIC-Maß zu einem Modell M_0 lautet

$$\mathrm{BIC}(M_0) = G^2(M_0) - df \cdot \ln(N) \qquad (4.3.6\text{-}3)$$

Wie im Falle der AIC-Differenz wird M_0 genau dann verworfen, wenn das BIC-Maß hierzu positiv ist. Offensichtlich muß $G^2(M_0)$ mit wachsendem Gesamtstichprobenumfang N immer größer werden, damit eine Verwerfung des Modells auf der Basis von BIC in Frage kommt. Raftery (1986) hat gezeigt, daß das BIC-Kriterium im Falle multinomialer Modelltests konsistent ist, d.h. asymptotisch zu korrekten Entscheidungen über M_0 führt.

Diese positiven Eigenschaften legen die Empfehlung nahe, das BIC-Maß beim Vergleich verschiedener multinomialer Modelle zu einem Datensatz zugrundezulegen. Man wählt das Modell als optimales Modell aus, dessen BIC-Statistik am kleinsten ist. Da der Stichprobenumfang im BIC-Maß berücksichtigt wird, ist im Prinzip auch ein Vergleich von BIC-Werten aus verschiedenen Untersuchungen (mit unterschiedlichem N) denkbar, sofern in den Untersuchungen jeweils die gleiche kategoriale Datenstruktur erhoben wurde.

BIC kann besonders hilfreich sein, wenn Submodelle am Rand des Parameterraumes übergeordneter Modelle liegen, beispielsweise beim Vergleich verschiedener Latent-Class-Modelle oder finiter Mischverteilungsmodelle mit unterschiedlicher Klassenanzahl. Da für diesen Fall — wie oben schon ausgeführt — der herkömmliche asymptotische G^2-Differenzentests statistisch unbegründet ist, hat jüngst Nagin (1999) die BIC-Statistik als Alternative propagiert, mit der die bestpassende Modellvariante leicht und ohne viel Rechenaufwand identifiziert werden kann.

Wird statt G^2 eine andere Minimum-PD$^\lambda$-Anpassungsstatistik präferiert (vgl. Abschnitt 4.3.5), so liegt es nahe, ein BIC$^\lambda$-Maß zu definieren. BIC$^\lambda$ resultiert aus Gleichung (4.3.6-3), indem man die G^2-Statistik durch die gewünschte Minimum-PD$^\lambda$-Anpassungsstatistik ersetzt (Read & Cressie, 1988, S. 127). Verbindliche Empfehlungen hierzu können derzeit noch nicht gegeben werden. Ein statistisches Rationale, das sich zum AIC-Kriterium (Akaike, 1973) und zum BIC-Kriterium (Schwarz, 1978) angeben läßt, existiert bislang zum BIC$^\lambda$-Kriterium noch nicht. Die Verwendung von BIC$^\lambda$ läßt sich also lediglich pragmatisch rechtfertigen. Wünschenswert wäre deshalb eine Evaluation der verschiedenen BIC$^\lambda$-Modellanpassungsmaße anhand von Kreuzvalidierungsstudien. Ist dasjenige Modell, das einen bestimmten BIC$^\lambda$-Wert minimiert, zugleich auch dasjenige, das neue, unabhängig gewonnene Daten am besten vorhersagt? Einschlägige Publikationen zu dieser Frage stehen noch aus.

4.4 Datenäquivalenz von Modellen und verwandte Probleme

Beim Vergleich verschiedener Modelle kann ein besonderes Problem entstehen, wenn die Werte der Modellanpassungsstatistiken zu gegebenen Daten numerisch exakt identisch sind. Offenbar ist eine Entscheidung zwischen den Modellen dann

zunächst einmal nicht möglich. Auch die AIC- und die BIC-Anpassungsmaße helfen nicht weiter, wenn die betreffenden Modelle gleich viele freie Parameter aufweisen.

Die exakte Übereinstimmung des goodness-of-fits verschiedener Modelle ist häufig ein Hinweis auf Äquivalenz oder zumindest Datenäquivalenz der betreffenden Modelle. Diese Begriffe lassen sich wie folgt definieren:

> Seien $M = \langle M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$ und $M' = \langle M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}', \boldsymbol{\Gamma}, f' \rangle$ zwei parametrisierte multinomiale Modelle zur gleichen verbundenen Multinomialverteilung $M(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi})$ mit $\boldsymbol{\Gamma}$ als Raum der Kategoriewahrscheinlichkeiten (vgl. Abschnitt 3.3). M und M´ heißen genau dann datenäquivalent, wenn
>
> $$f(\Omega) = f'(\boldsymbol{\Omega}') \tag{4.4-1}$$
>
> gilt. Haben M und M´ darüber hinaus gleiche Parameterzahl $S = S'$ und gilt
>
> $$\exists(h: \boldsymbol{\Omega}' \rightarrow \boldsymbol{\Omega}): \forall \boldsymbol{\theta} \in \boldsymbol{\Omega}, \forall \boldsymbol{\theta}' \in \boldsymbol{\Omega}': f(\boldsymbol{\theta}) = f'(\boldsymbol{\theta}') \Rightarrow \boldsymbol{\theta} = h(\boldsymbol{\theta}') , \tag{4.4-2}$$
>
> wobei $h: \boldsymbol{\Omega}' \rightarrow \boldsymbol{\Omega}$ eine Permutation der Vektorelemente von $\boldsymbol{\theta}'$ (oder die Identitätsfunktion) ist, so heißen die Modelle äquivalent.

Datenäquivalenz zweier Modelle bedeutet also, daß jeder denkbare Vektor von Kategoriewahrscheinlichkeiten $\boldsymbol{\pi} \in \boldsymbol{\Gamma}$ entweder durch beide Modelle erklärt werden kann ($\boldsymbol{\pi} \in f(\boldsymbol{\Omega})$ und $\boldsymbol{\pi} \in f'(\boldsymbol{\Omega}')$) oder durch beide Modelle nicht erklärt werden kann ($\boldsymbol{\pi} \notin f(\boldsymbol{\Omega})$ und $\boldsymbol{\pi} \notin f'(\boldsymbol{\Omega}')$). Ausgeschlossen sind die Kombinationen $\boldsymbol{\pi} \in f(\boldsymbol{\Omega})$ und $\boldsymbol{\pi} \notin f'(\boldsymbol{\Omega}')$ sowie $\boldsymbol{\pi} \in f'(\boldsymbol{\Omega}')$ und $\boldsymbol{\pi} \notin f(\boldsymbol{\Omega})$. Mit anderen Worten: M und M´ sind datenäquivalent, wenn die Modellgeltungshypothesen zu M und M´ äquivalent sind. Trifft die Modellgeltungshypothese zu M und M´ nicht zu, so sind die Abweichungen zwischen $\boldsymbol{\pi}$ und dem nächstliegenden Punkt aus $f(\boldsymbol{\Omega})$ einerseits sowie die zwischen $\boldsymbol{\pi}$ und dem nächstliegenden Punkt aus $f'(\boldsymbol{\Omega}')$ andererseits wegen der Gleichheit $f(\boldsymbol{\Omega}) = f'(\boldsymbol{\Omega}')$ immer gleich groß, egal welches Distanzmaß zugrunde gelegt wird. Aus diesem Grund müssen Modellanpassungsstatistiken für datenäquivalente Modelle numerisch zwangsläufig exakt identisch ausfallen, wobei es unerheblich ist, welche PD^{λ}-Anpassungsstatistik gewählt wird.

Sind zwei Modelle M und M´ datenäquivalent, so sagt man auch, M´ sei eine Reparametrisierung von M und vice versa (vgl. Bamber & van Santen, 1985, Definition 18, S. 463). Die Äquivalenz von M und M´ ist ein Spezialfall der Datenäquivalenz.[49] Sie verlangt einerseits, daß man jedes der beiden Modelle durch das andere reparametrisieren kann, und andererseits, daß zusätzlich die Werte der Para-

[49] Viele Autoren setzen die Begriffe „Datenäquivalenz" und „Äquivalenz" von Modellen gleich. Ich weiche davon ab, weil der Versuch einer empirischen Entscheidung zwischen substanzwissenschaftlich motivierten multinomialen Modellen unterschiedliche Maßnahmen erfordert, wenn die betreffenden Modelle (a) datenäquivalent oder (b) äquivalent (im hier definierten Sinne) sind. Dies wird im nächsten Abschnitt näher ausgeführt.

meter beider Modelle nach geeigneter Zuordnung übereinstimmen. Äquivalente Modelle sind also eigentlich identische Modelle, nur daß die Parameter in beiden Modellen unterschiedlich benannt oder indiziert sein können. In jedem Fall existiert bei äquivalenten Modellen eine eineindeutige Zuordnung der Parameter von M´ zu denen von M, so daß die einander zugeordneten Parameter numerisch identisch sind, sofern sie auf das gleiche $\boldsymbol{\pi} \in \boldsymbol{\Gamma}$ abgebildet werden.

Die Äquivalenz zweier Modelle erzwingt per definitionem die Gleichheit der Parameteranzahl in beiden Modellen, d.h. es gilt immer $S = S'$. Bei datenäquivalenten Modellen muß $S = S'$ dagegen nicht unbedingt erfüllt sein (Bamber & van Santen, 1985; 2000). Zum Beispiel ist das Bradley-Terry-Luce (BTL) Modell für Wahlwahrscheinlichkeiten (vgl. Luce, 1959) in seiner unrestringierten Form ein Modell mit m positiv-reellwertigen „Attraktivitätsparametern" $\theta_1, \theta_2, ..., \theta_m$, wenn ein kompletter Paarvergleich für m Objekte durchgeführt wird. Das BTL-Modell behauptet, daß die Wahrscheinlichkeit, ein beliebiges Objekt i einem beliebigen anderen Objekt j vorzuziehen, der Modellgleichung $p_{i,j} = \theta_i / (\theta_i + \theta_j)$ genügt. Offensichtlich ist dieses Modell in der unrestringierten Version nicht identifizierbar, denn durch Multiplikation aller Parameter mit einer beliebigen positiven Konstanten c erhält man einen zweiten Parametervektor $c \cdot \boldsymbol{\theta}$, der exakt die gleichen Wahlwahrscheinlichkeiten vorhersagt wie $\boldsymbol{\theta}$. Die Konstante c läßt sich nämlich aus der genannten Modellgleichung herauskürzen. Lösen kann man dieses Identifizierbarkeitsproblem dadurch, daß man beispielsweise die „Attraktivität" des ersten Objekts als Einheit auszeichnet, was durch die Parameterfixierung $\theta_1 = 1$ erreicht werden kann. Man erhält dann ein restringiertes BTL-Modell mit $m-1$ freien Parametern $\theta_2, ..., \theta_m$, das zu dem unrestringierten Ausgangsmodell datenäquivalent ist. Darüber hinaus ist dieses restringierte Modell auch identifizierbar, denn die Parameter können wegen der Restriktion $\theta_1 = 1$ nicht mehr mit einer beliebigen Konstanten c multipliziert werden. Man sieht, daß zwei datenäquivalente Modelle durchaus unterschiedlich viele freie Parameter haben können. Praktisch bedeutet dies aber immer, daß (mindestens) eines der beiden Modelle — nämlich das mit der größeren Parameterzahl — nicht identifizierbar ist (Bamber & van Santen, 1985; 2000).

Weitere Beispiele für datenäquivalente Modelle sind Schwellen- und Signalentdeckungsmodelle (mit $d' \geq 0$) für Ja-Nein-Rekognitionstests, genauer: für *eine* Trefferrate und *eine* Rate falschen Alarms. Für alle diese Modelle ist das Bild des Parameterraumes die Menge möglicher Trefferraten und Raten falschen Alarms, bei denen die Trefferrate nicht kleiner als die Rate falschen Alarms ist. Anschaulich entspricht dies dem Dreieck oberhalb der Winkelhalbierenden im Raum möglicher Receiver Operating Characteristics (ROCs). So sind beispielsweise die Standardversion der Signalentdeckungstheorie (mit Parametern $d' \geq 0$ und β) und das Zwei-Hochschwellen-Modell (mit Parametern r_{2ht} und b_{2ht}) datenäquiva-

lent, weil sie exakt identische Mengen von Treffern und falschen Alarmen erklären. Obwohl in diesem Fall die Anzahl der Modellparameter gleich ist ($S = S' = 2$), handelt es sich nicht um äquivalente Modelle, denn es existiert keine Zuordnung der Parameter beider Modelle derart, daß die Parameter paarweise numerisch identisch ausfallen.

Beispiele für äquivalente MVB-Modelle sind die beiden Ereignisbaummodelle aus Abbildung 3.4. Man kann sich zwei inhaltlich inkompatible kognitive Prozeßtheorien vorstellen, die zu den unterschiedlichen Ereignisbäumen führen. Trotz ihrer strukturellen Unterschiede implizieren die Ereignisbäume — wie in Abschnitt 3.6 gezeigt wurde — identische MVB-Modellgleichungen (3.6-1) und sind deshalb zwangsläufig äquivalent. Analoge Fälle von Modelläquivalenzen lassen sich durch Vertauschen der Parameterreihenfolge oder Permutation der Kategorien, in die gleichwahrscheinliche Äste eines Baummodells einmünden, erzeugen, sofern sich dadurch die Modellgleichungen nicht ändern (vgl. Abschnitt 3.6).

Eine vielleicht noch informativeres Beispiel für die Äquivalenz zweier MVB-Modelle, denen substanzwissenschaftlich ganz unterschiedliche Ideen zugrunde liegen, ist die formale Identität des Zwei-Hochschwellen-Modells für Ja-Nein-Rekognitions- oder Signalentdeckungstest einerseits und des von Jacoby (1991) vorgeschlagenen Modells zur Prozeßdissoziationsprozedur (PDP) andererseits (Bredenkamp, 1998a, Vaterrodt-Plünnecke, 1994; Windmann & Krüger, 1998). Die PDP existiert in verschiedenen Varianten. Bei der Rekognitionsvariante der PDP lernen Vpn zunächst zwei disjunkte Itemlisten L1 und L2, die zwecks besserer Diskriminierbarkeit z.B. visuell (L1) versus akustisch (L2) dargeboten werden können. Anschließend folgt ein Rekognitionstest, der L1-Items, L2-Items und neue Distraktoritems umfaßt. Der Rekognitionstest wird unter zwei Bedingungen durchgeführt, die prinzipiell innerhalb einer Vp oder auch zwischen Vpn-Gruppen variiert werden können: der Inklusions- und der Exklusionsbedingung. In der Inklusionsbedingung sollen die Vpn rekognizierte L1- und L2-Items mit „ja“ beantworten, Distraktoren dagegen mit „nein“. In der Exklusionsbedingung sind von der Vp L1-Items auszuschließen, d.h. nur L2-Items sollen mit „ja”, L1-Items und Distraktoren dagegen mit „nein” beantwortet werden.

Für die Auswertung nach dem von Jacoby (1991) vorgeschlagenen Modell interessieren nur die Wahrscheinlichkeiten von „Ja”-Antworten zu L1-Items in der Inklusionsbedingung (p_i) und in der Exklusionsbedingung (p_e). Antworten zu L2-Items und Distraktoren bleiben in diesem Modell — anders als in Weiterentwicklungen des Ansatzes (Buchner et al., 1997; Buchner et al., 1995; Erdfelder & Buchner, 1998b; Mulligan & Hirshman, 1997; Vaterrodt-Plünnecke, Krüger, Gerdes & Bredenkamp, 1996; Yonelinas & Jacoby, 1996a; Yonelinas et al., 1995) — unberücksichtigt. Jacoby (1991) nahm an, daß die Wahrscheinlichkeit einer L1-Item-Rekognition in der Inklusionsbedingung einerseits durch bewußte Erinnerung mit

Wahrscheinlichkeit c und andererseits durch ein starkes Vertrautheitsgefühl ohne bewußte Erinnerung mit Wahrscheinlichkeit $(1-c)\cdot u$ erfolgen kann. In der Exklusionsbedingung kann nach Jacoby (1991) nur im letztgenannten Fall (Vertrautheit ohne bewußte Erinnerung) eine „Ja"-Antwort resultieren. Zusätzlich unterstellte Jacoby stochastische Unabhängigkeit von Vertrautheit und bewußter Erinnerung, eine Annahme, die jedoch prinzipiell verzichtbar ist (Buchner & Erdfelder, 1996; Buchner et al., 1995; Erdfelder & Buchner, 1998b). Aus allen Annahmen zusammen resultieren die beiden Modellgleichungen

$$p_i = c + (1-c)\cdot u \tag{4.4-3}$$

$$p_e = (1-c)\cdot u\ , \tag{4.4-4}$$

wobei c und u unbedingte Wahrscheinlichkeiten einer bewußten Erinnerung bzw. eines starken Vertrautheitsgefühls sind.

Wie man durch Vergleich mit den Modellgleichungen des Zwei-Hochschwellen-Modells leicht feststellen kann, sind die Gleichungen (4.4-3) und (4.4-4) mit den Gleichungen (3.5.2-3) und (3.5.2-4) äquivalent (Bredenkamp, 1998a; Buchner et al., 1995; Vaterrodt-Plünnecke, 1994; Vaterrodt-Plünnecke et al., 1996). Man muß nur die Trefferrate mit p_{i}, die Rate falschen Alarms mit p_{e}, den Rekognitionsparameter $r_{2\mathrm{ht}}$ mit c und den Antworttendenzparameter $b_{2\mathrm{ht}}$ mit u identifizieren, um das Zwei-Hochschwellen-Modell in das PDP-Modell von Jacoby (1991) zu überführen. Man beachte, daß die unterschiedlichen Instruktionen der Vpn und die unterschiedlichen psychologischen Interpretationen der Parameter c und $r_{2\mathrm{ht}}$ sowie $b_{2\mathrm{ht}}$ und u an der Tatsache der Modelläquivalenz ebensowenig ändern wie der Sachverhalt, daß der Antworttendenzparameter $b_{2\mathrm{ht}}$ im Zwei-Hochschwellen-Modell eine *bedingte* Wahrscheinlichkeit ist (Erraten von „ja", gegeben keine Rekognition), während der korrespondierende Parameter u bei Jacoby (1991) durch die hinzugenommene stochastische Unabhängigkeitsannahme zu einer *unbedingten* Wahrscheinlichkeit wird. Modelläquivalenz bedeutet also nicht automatisch auch inhaltliche (d.h. substanzwissenschaftliche) Äquivalenz. Die Unterscheidung von substanzwissenschaftlicher und formaler Äquivalenz ist für die Frage, ob zwischen verschiedenen äquivalenten oder datenäquivalenten Modellen *prinzipiell* eine empirische Entscheidung herbeigeführt werden kann, von zentraler Bedeutung. Solange zwei Modelle zwar formal äquivalent sind, die numerisch einander korrespondierenden Parameter jedoch unterschiedlich psychologisch interpretiert werden, ist eine empirische Entscheidung prinzipiell möglich. Wir kommen im nächsten Abschnitt hierauf zurück.

Ein eng mit dem Problem der Datenäquivalenz zusammenhängendes Problem ist das der Gültigkeits-Implikationsbeziehung zwischen Modellen, die gleiche Parameterzahl aufweisen. Erdfelder, Murnane und Bayen (1995) haben dieses Pro-

blem am Beispiel dreier Quellendiskriminationsmodelle untersucht, die sich auf einfache Quellendiskriminationsaufgaben (source monitoring tasks) beziehen. Vpn lernen hierbei zunächst Items aus zwei Quellen — beispielsweise zwei Itemlisten L1 und L2 wie bei der Prozeßdissoziationsprozedur — und müssen dann im anschließenden Quellendiskriminationstest angeben, ob das betreffende Testitem (a) alt ist und aus Quelle L1 stammt, (b) alt ist und aus Quelle L2 stammt oder (c) neu ist. Im Prinzip handelt es sich also um einen erweiterten Ja-Nein-Rekognitionstest, bei dem im Falle einer Ja-Antwort zusätzlich die Quelle des Items anzugeben ist. Die von Erdfelder at al. (1995) untersuchten Quellendiskriminationsmodelle sind Vier-Parameter-BVB-Modelle, die jeweils einen Parameter für das Itemgedächtnis (Alt-Neu-Diskriminierung), einen für das Quellengedächtnis (L1-L2-Diskriminierung eines als alt erkannten Items), einen Rateparameter für die Alt-Neu-Entscheidung im Falle der Rekognitionsunsicherheit und einen Rateparameter für die Quellenentscheidung bei Quellenunsicherheit umfassen. Die korrespondierenden Parameter sollen in den drei Modellen psychologisch identische Konstrukte erfassen. Die Modelle unterscheiden sich jedoch insofern, als entweder ein Ein-Hochschwellen-Modell (1HTSM-Modell), ein Zwei-Hochschwellen-Modell (2HTSM-Modell) oder ein Niedrigschwellen-Modell (LTSM-Modell) für das Itemgedächtnis unterstellt wird (Details in Bayen, Murnane & Erdfelder, 1996; Erdfelder, Murnane & Bayen, 1995). Diese Unterschiede verhindern, daß die drei Modelle äquivalent sind. Im allgemeinen sind also die numerischen Werte korrespondierender Parameter zu gegeben Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ nicht gleich. Die Modelle sind darüber hinaus auch nicht datenäquivalent. Man kann zwar alle Kategoriewahrscheinlichkeit, die mit dem 2HTSM-Modell konsistent sind, auch durch das 1HTSM-Modell erklären. Die Umkehrung gilt jedoch nicht, solange die Parameter des 2HTSM-Modells auf das Intervall [0, 1] beschränkt werden. Analog impliziert die Gültigkeit des 1HTSM-Modells die des LTSM-Modells, während die Umkehrung wegen der Beschränkung der Parameter auf das Intervall [0, 1] nicht gilt (Erdfelder et al., 1995, S. 544). Wir haben es hier also mit identischen Parameterräumen für alle drei Modelle zu tun, nämlich $\boldsymbol{\Omega} = [0, 1]^4$, jedoch mit unterschiedlichen Bildmengen von $\boldsymbol{\Omega}$. Bezeichnen $f_{1\mathrm{ht}}$, $f_{2\mathrm{ht}}$ und f_{lt} die modellspezifischen Abbildungen von $\boldsymbol{\Omega}$ in $\boldsymbol{\Gamma}$ für das 1HT-, das 2HT- und das LTSM-Modell, so gilt für die Bildmengen von $\boldsymbol{\Omega}$ die Teilmengenrelation

$$f_{2\mathrm{ht}}(\boldsymbol{\Omega}) \subset f_{1\mathrm{ht}}(\boldsymbol{\Omega}) \subset f_{lt}(\boldsymbol{\Omega}) \subset \boldsymbol{\Gamma} \qquad (4.4\text{-}5)$$

Würden sich die drei Bildmengen bedeutsam unterscheiden, so daß die paarweisen Schnittmengen zwischen ihnen hinreichend klein sind, wäre die Entscheidung für oder gegen ein Modell auf der Basis von Modellanpassungsstatistiken oder AIC- und BIC-Maßen oftmals möglich. Im gegebenen Falle der Vier-Parameter-Quellendiskrimationsmodelle sind jedoch die Schnittmengen fast so groß

wie $f_{lt}(\boldsymbol{\Omega})$, so daß der typische Anwendungsfall auf identische Anpassungsstatistiken für alle drei Modelle hinausläuft (vgl. Bayen et al., 1996). Dies bedeutet, daß — genau wie bei datenäquivalenten Modellen — entweder alle Modelle zu den Daten passen oder alle Modelle nicht passen. Mit anderen Worten: Obwohl die Modelle im strengen Sinne nicht datenäquivalent sind, sind sie doch „fast datenäquivalent", so daß sich das Entscheidbarkeitsproblem im Regelfall genauso wie bei datenäquivalenten Modellen stellen.

Bei MVB- und BVB-Modellen ist dieser Fall der „Beinahe-Datenäquivalenz" fast noch häufiger als vollständige Datenäquivalenz. Dies liegt an der Beschränkung der Modellparameter auf das Einheitsintervall [0, 1]. Modelle, die für unbeschränkte reellwertige Parameter datenäquivalent wären — so z.B. die drei oben geschilderten Quellendiskriminationsmodelle — verlieren die Datenäquivalenzeigenschaft als Folge der Ungleichheitsrestriktionen, die mit der Einschränkung der Parameter auf das Intervall [0, 1] verbunden sind. Die Konsequenz sind Teilmengenbeziehungen zwischen den Gültigkeitsbereichen der betreffenden Modelle, wobei allerdings der kleinere Gültigkeitsbereich den größeren oftmals fast vollständig ausschöpft.

Die Probleme der Äquivalenz, Datenäquivalenz und „Beinahe-Datenäquivalenz" zwischen Modellen zeigen, daß eine befriedigende Modellevaluation mit Modellanpassungstests und informationstheoretischen Anpassungsmaßen wie AIC und BIC allein nicht erreicht werden kann. Zu oft kommt es vor, daß die betreffenden Modellgüteindikatoren nicht oder nicht hinreichend zwischen konkurrierenden Modellen trennen. Liegen zunächst keine konkurrierenden Modelle vor, so ist es oftmals relativ leicht, ein inhaltlich abweichendes, jedoch hinsichtlich der Anpassungsgüte gleichwertiges Modell zu einem zu evaluierenden Modell zu konstruieren. Mit anderen Worten: Das Problem, zwischen gleich gut oder nahezu gleich gut zu gegebenen Daten passenden Modellen eine Entscheidung fällen zu müssen, ist eher der Regelfall als die Ausnahme.

Aber die Situation ist nicht so hoffnungslos wie es zunächst scheinen mag. Ein bestimmtes Bündel empirischer Resultate kann potentiell immer durch mehrere Theorien erklärt werden. Dies ist ein generelles Problem empirischer Forschung und nicht auf stochastische Modellierungsansätze beschränkt (Wickens, 1998b). Der Ausweg aus diesem Problem liegt auf der Hand: „When two theories make the same prediction, all is not lost. We can try to find a place where the implications of the two theories differ and run a new experiment to decide between them" (Wickens, 1998b, S. 607). Diesem Ansatz werde ich mich im folgenden Abschnitt zuwenden.

4.5 Modellvalidierung

Die bloße Tatsache, daß die Modellgeltungshypothese eines substanzwissenschaftlich motivierten multinomialen Modells in einem konkreten Anwendungsfall relativ zu einem bestimmten Typ-1-Risiko α nicht zurückgewiesen werden kann, ist schon allein wegen der Möglichkeit eines β-Fehlers bei der statistischen Entscheidung kein hinreichender Grund, das Modell für psychologisch bestätigt zu halten. Das Mindestkriterium, dem ein Modell genügen sollte, ist die *wiederholte Bewährung* bei mehrfacher Anwendung auf unabhängige Datensätze, wobei deren Stichprobenumfänge groß genug sein sollten, um bedeutsame Modellverletzungen mit hinreichender Teststärke aufdecken zu können (vgl. Abschnitt 4.3.3). „Wiederholte Bewährung" bedeutet dabei aber nicht, daß der Modellgeltungstest nie signifikant ausgehen darf. Es bedeutet lediglich, daß der kritische $\chi^2_{(df)}$-Wert c_α im Durchschnitt nicht deutlich häufiger als $\alpha \cdot n$ Mal überschritten werden sollte, wenn n unabhängige Modelltests zum gleichen Niveau α vorliegen. Werden also beispielsweise zwanzig Modelltests zu konventionellen Signifikanzniveaus durchgeführt, so sind ein, zwei oder auch drei signifikante Teststatistiken bei Modellgültigkeit statistisch durchaus zu erwarten und deshalb noch kein Grund, das Modell zu verwerfen. Das bestmögliche Ergebnis bei mehrfacher Modellanwendung ist immer eine empirische Verteilung von Anpassungsstatistiken, die der betreffenden zentralen $\chi^2_{(df)}$-Referenzverteilung bei Modellgültigkeit maximal ähnlich ist. Es sollten also annähernd $\alpha \cdot 100\%$ der Statistiken signifikant sein, der Mittelwert der Anpassungsstatistiken sollte ungefähr beim Erwartungswert df liegen usw. Ein Kolmogoroff-Smirnov (KS) Anpassungstest kann Klarheit darüber schaffen, ob die theoretische und die empirische Verteilung der Teststatistiken über die unabhängigen Datensätze hinweg hinreichend nah beieinander liegen (vgl. z.B. Bredenkamp, 1984, für eine ähnliche Anwendung von KS-Tests zur summarischen Evaluation multipler Signifikanztests einer zugrundeliegenden psychologischen Hypothese).[50]

Wenn ein in diesem Sinne positives Ergebnis vorliegt, ist das betreffende Modell *potentiell* psychologisch gültig. Der vollständige Gültigkeitsnachweis steht aber wegen des Problems äquivalenter, datenäquivalenter oder „nahezu datenäquivalenter" Modelle, die zu anderen psychologischen Schlußfolgerungen aus den gleichen Daten führen können, noch aus. Offenbar helfen Modellanpassungstests und auch AIC- oder BIC-Anpassungsmaße hier nicht weiter, auch wenn sie in großer Zahl vorliegen. Allein ein Forschungsprogramm mit dem Ziel einer syste-

[50]Die denkbare Alternative, die Daten aller Datensätze zunächst zu aggregieren und dann auf der Basis der aggregierten Daten eine einzige statistische Entscheidung zu fällen, ist ausgesprochen problematisch, wenn von bedeutsamen Parameterunterschieden zwischen den einzeln Untersuchungen ausgegangen werden muß. Denkbar ist, daß sich ein in den Einzeluntersuchungen jeweils gültiges Modell als unzutreffendes Modell für die aggregierten Daten herausstellt (vgl. Kapitel 5).

matischen Modellvalidierung in dem von Wickens (1998) angedeuteten Sinn kann Klarheit über die psychologische Gültigkeit verschaffen.

Was heißt es, ein Modell psychologisch zu validieren? Im Grunde genommen bedeutet es, den Nachweis der Konstruktvalidität im Sinne von Cronbach und Meehl (1955) zu führen, und zwar für jeden substanzwissenschaftlich interessierenden Modellparameter separat. Allerdings geschieht dies nicht durch Korrelationsstudien wie im Falle psychologischer Tests, sondern durch Experimente, die darauf abzielen, die primär interessierenden Modellparameter durch Manipulationen geeignet gewählter unabhängiger Variablen (UV) in vorhersagbarer Weise zu beeinflussen (vgl. bereits Chechile & Meyer, 1976, sowie Brainerd, 1985, S. 156).

Man kann zwischen einem schwachen und einem strengen Kriterium für die Validierung eines Modells unterscheiden (Batchelder, 1994; Erdfelder, 1992a; Erdfelder et al., 1995). Nach dem schwachen Kriterium sollte für jeden substanzwissenschaftlich interessierenden Modellparameter *mindestens* eine experimentelle UV benennbar sein, mit der der betreffende Parameter in vorhersagbarer Weise beeinflußt werden kann. Die Vorhersage muß durch gut etablierte Forschungsergebnisse oder allgemein akzeptierte Theorien so stark abgesichert sein, daß ein Scheitern der Vorhersage zwingend auf die mangelnde Validität des betreffenden Modellparameters zu attribuieren ist und nicht etwa auf die psychologische Fragwürdigkeit der Vorhersage (Erdfelder & Buchner, 1998a; Klauer & Wegener, 1998). Es geht also in Validierungsexperimenten nicht darum, neue empirische Effekte oder verblüffende Ergebnisse zu präsentieren, sondern es sollen offensichtliche Vorhersagen und schon mehrfach replizierte Befunde anhand neuer abhängiger Variablen — den betreffenden Modellparametern — erneut repliziert werden. Derartige Experimente sind zugegebenermaßen „often boring and trivial" (Batchelder, 1994) in bezug auf die zu erwartenden Effekte. Große Überraschungen und neue theoretische Einsichten kann man sich von den Ergebnissen der Validierungsexperimente nicht erhoffen. Dennoch zeigen Validierungsexperimente im positiven Fall, daß die Modellparameter sich so verhalten wie man es psychologisch von ihnen erwartet — ein keinesfalls triviales Resultat —, und im negativen Fall, daß auch ein empirisch adäquates, die Daten korrekt beschreibendes Modell psychologisch unbrauchbar sein kann, weil offensichtliche Effekte unabhängiger Variablen in den Modellparametern nicht widergespiegelt werden.

Bewährt sich ein Modell auf der Grundlage des schwachen Validitätskriteriums, so kann man sich nicht ganz sicher sein, ob das Dekomponierungsproblem, das der Modellkonstruktion zugrunde lag, wirklich erfolgreich gelöst wurde (vgl. Abschnitt 2.6). Der Zielparameter ist zwar nachweislich sensitiv für den psychologischen Prozeß, den er erfassen soll, möglicherweise sind es aber auch andere Parameter des Modells, die eigentlich andere psychologische Prozesse erfassen sollen. Genau dieses Problem hat man in der designbasierten Forschung häufig, und ein

Ziel der modellbasierten Forschung besteht darin, dieses Problem zu überwinden. Eine Verschärfung des schwachen Validitätskriteriums zum strengen Validitätskriterium liegt deshalb nahe. Während beim schwachen Kriterium nur Vorhersagen zum Effekt jeweils einer UV auf jeweils einen Modellparameter gemacht werden und die Effekte auf andere Modellparameter nicht interessieren, verlangt man beim strengen Kriterium einen *selektiven* Effekt der UV auf genau einen Modellparameter, d.h. die anderen Modellparameter dürfen durch die Manipulation der UV nicht beeinflußt werden (Bayen et al., 1996; Bredenkamp, 1998a; Buchner et al., 1995; Buchner et al., 1997; Erdfelder & Buchner, 1998a; Klauer & Wegener, 1998). Dies entspricht der Forderung nach konvergenter und diskriminanter Validität im Sinne von Campbell und Fiske (1959): Der manipulierte psychologische Prozeß soll in dem Zielparameter reflektiert werden (konvergente Validität), während er sich in den Parametern, die andere Prozesse erfassen sollen, nicht niederschlagen darf (diskriminante Validität) (Klauer & Wegener, 1998).

Übersteht ein multinomiales Modell eine Serie von Validierungsexperimenten nach dem strengen Validitätskriterium, so kann mit Recht davon gesprochen werden, daß dieses Modell das Dekomponierungsproblem in dem betreffenden Forschungsbereich erfolgreich löst (vgl. Abschnitt 2.6). Allerdings — das sollte klar sein — ist jede positive Antwort zur Validitätsfrage eine vorläufige. Die Menge der experimentellen UVn, die sich vorhersagbarer Weise auf die einzelnen Modellparameter auswirken müssen, ist prinzipiell nicht abgeschlossen, sondern offen. Die Validitätshypothese ist folglich eine Allaussage, die sich auf eine Grundmenge mit überabzählbar vielen Elementen bezieht: Für alle experimentellen UVn mit eindeutigen, selektiven Effekten auf Prozeß X muß gelten, daß sie auch selektiv den Modellparameter affizieren, der Prozeß X erfassen soll. Aufgrund ihrer logischen Struktur kann man die Validitätshypothese nur falsifizieren, nicht jedoch verifizieren (Popper, 1934/1982). Falsifizieren kann man sie, indem man zeigt, daß ein trivialerweise zu erwartender Effekt — etwa die Länge des Retentionsintervalls auf das Nachlassen der Erinnerungsleistung — in den Modellparametern wiederholt nicht den Niederschlag findet, der theoretisch zu erwarten ist. Umgekehrt kann man die Validitätshypothese nicht verifizieren, weil es auch nach einer großer Zahl von Validierungsexperimenten mit positivem Ausgang denkbar ist, daß der nächste Validierungsversuch scheitert. Dies ist erneut kein Spezifikum stochastischer Modellierungsansätze in der Psychologie, sondern ein generelles Problem empirischer Forschung, weil Allsätze — deterministische und stochastische — einen zentralen Stellenwert in den Erfahrungswissenschaften — gerade in der Psychologie — einnehmen (Bredenkamp, 1972; 1980; 1986; Gadenne, 1976; 1984).

Inwiefern sind Validierungsexperimente nach dem schwachen oder dem starken Validitätskriterium geeignet, eine Entscheidung zwischen Modellen her-

beizuführen, die bei Anwendung auf einzelne Datensätze identische Modellanpassungsstatistiken aufweisen? Betrachten wir zunächst datenäquivalente oder „beinahe datenäquivalente" Modelle, die jedoch nicht äquivalent sind, so daß sich die Werte der korrespondierenden (d.h. mit gleichen psychologischen Interpretationen versehenen) Modellparameter voneinander unterscheiden. In diesem Fall sucht man für jeden substanzwissenschaftlich interessierenden Parameter mindestens eine experimentelle UV, die den psychologischen Prozeß, den der betreffende Parameter erfassen soll, möglichst sicher beeinflußt. „Möglichst sicher" ist hierbei natürlich an dem jeweils verfügbaren fachwissenschaftlichen Wissen zu relativieren. Man führt dann ein Experiment oder mehrere Experimente durch, in dem man die ausgewählte UV und gegebenenfalls weitere UVn orthogonal variiert. Jedes der konkurrierenden datenäquivalenten Modelle wird so auf ein verbundenes Modell für den kompletten Versuchsplan verallgemeinert, daß jeder Modellparameter in jeder Versuchsbedingung beliebige Werte annehmen kann. Praktisch bedeutet das einfach eine „Multiplikation" des Ausgangsmodells um den Faktor v, wobei v die Anzahl der Versuchsbedingungen ist. Hat das Ausgangsmodell S Modellparameter und M funktional unabhängige Kategoriewahrscheinlichkeiten, so hat das unrestringierte Modell für den kompletten Versuchsplan folglich $S \cdot v$ Parameter (jeder der S Ausgangsparameter kann in jeder der v Versuchsbedingungen beliebige Werte annehmen) und entsprechend $M \cdot v$ funktional unabhängige Kategoriewahrscheinlichkeiten. Für dieses unrestringierte verbundene Modell berechnet man eine Modellanpassungsstatistik, am besten die Log-Likelihood-Quotienten-Statistik. Anschließend formuliert man für jeden zu validierenden Modellparameter und jede manipulierte UV ein restringiertes Modell, wobei der betreffende Parameter über die Stufen der betreffenden UV gleichgesetzt wird (vgl. Abschnitt 4.3.2.1). Liegen r_1 UVs für r_2 zu validierende Parameter vor, so ergeben sich folglich insgesamt $r_1 \cdot r_2$ restringierte Modelle. Auch für jedes dieser restringierten Modelle berechnet man eine Log-Likelihood-Quotienten-Anpassungsstatistik und vergleicht sie dann mit der des unrestringierten Ausgangsmodells via ΔG^2-Differenzentests.

Ein Modell hat sich im Sinne des schwachen Validitätskriteriums bewährt, wenn (a) das unrestringierte Ausgangsmodell die Daten des kompletten Versuchsplans adäquat beschreibt (d.h. keine Zurückweisung der globalen Modellgeltungshypothese), (b) diejenigenr_1 χ^2-Differenzentests signifikant ausfallen, die Effekte der r_1 UVs auf ihre korrespondierenden Zielparameter reflektieren, und (c) die Richtung des signifikanten Effekts (d.h. das Vorzeichen der Differenzen zwischen Parameterschätzwerten) den Vorhersagen entspricht. Gilt darüber hinaus, daß alle anderen $r_1 \cdot (r_2 - 1)$ Differenzentests, die Effekte der UVn auf Nicht-Zielparameter reflektieren, insignifikant ausfallen, so hat sich ein Modell auch im Sinne des strengen Validitätskriteriums bewährt. In allen Fällen ist natürlich vorauszuset-

zen, daß das α–Niveau der Tests und der Stichprobenumfang vernünftig aufeinander abgestimmt sind, so daß die Teststärke für bedeutsame Modellverletzungen — und möglichst nur für diese — hinreichend groß ist (vgl. Abschnitt 4.3.3).

Der „Trick", mit dem eine Entscheidung zwischen datenäquivalenten Modellen gelingen kann, ist also die Erhöhung des empirischen Gehalts der Modelle durch (a) Verallgemeinerung auf komplexe Versuchspläne in Verbindung mit (b) substanzwissenschaftlich begründeten Gleichheitsrestriktionen zu speziellen Modellparametern. Schritt (a) allein ist hierbei wohlgemerkt noch nicht zielführend; wenn verschiedene Ausgangsmodelle datenäquivalent sind, so sind es auch ihre unrestringierten Verallgemeinerungen auf mehrere Versuchsbedingungen. Erst die Kombination der verallgemeinerten Modelle mit gezielten Gleichheitsrestriktionen führt zur Aufhebung der Datenäquivalenz, weil die Gleichheitsrestriktionen wegen der numerischen Unterschiede korrespondierender Modellparameter modellspezifisch unterschiedliche Implikationen für die Gesamt-Datenstruktur des kompletten Versuchsplans haben.

Natürlich gibt es keine Garantie, daß ein Validierungsexperiment genau ein Modell aus einer Gruppe datenäquivalenter oder nahezu datenäquivalenter Modelle bestätigt, während alle anderen Modelle mangelhaft abschneiden. Dieser Fall wird in Publikationen zwar erstaunlich oft berichtet (Bayen et al., 1996; Buchner et al., 1995), was seine Ursache jedoch einfach darin haben kann, daß Untersuchungen mit eindeutigen Ergebnissen leichter publizierbar sind als Untersuchungen, die die Frage nach dem besten Modell unbeantwortet lassen. Denkbar ist immer, daß mehrere oder auch keines der evaluierten Modelle die Hürden des strengen oder schwachen Validitätskriteriums nehmen. Im erstgenannten Fall bleibt keine andere Wahl, als mit dem Validierungsprogramm fortzufahren, bis irgendwann einmal ein Modell über die anderen dominiert. Im letztgenannten Fall wird man nach gänzlich neuen Modellen oder Modellvarianten Ausschau halten müssen, wenn nicht einmal das schwache Validitätskriterium erfüllt ist. Scheitern alle Modelle am strengen, jedoch eines oder einige nicht am schwachen Validitätskriterium, so sollte man zunächst überlegen, ob das nicht auf die mangelnde Selektivität des Treatments, das mit der UV-Manipulation verbunden ist, zurückgeführt werden kann. Erhärtet sich dieser Verdacht, so sollte man nach besseren, d.h. selektiveren Manipulationen der interessierenden kognitiven Prozesse suchen und mit dem Validierungsprogramm fortfahren.

Ein Beispiel mag das illustrieren. Es erscheint prinzipiell denkbar, daß die angestrebte Manipulation der Antworttendenz (kognitiver Prozeß) durch Erhöhung des Anteils alter Items in Rekognitionstests (experimentelles Treatment) nicht nur die Ratewahrscheinlichkeit für „Alt"-Antworten erhöht, sondern zusätzlich auch die Erinnerungsmotivation und dadurch mittelbar auch die Erinnerungsleistung von Vpn beeinflußt. Yonelinas und Jacoby (1996a) haben genau die-

ses Argument benutzt, um unbefriedigende Validierungsergebnisse für das von Ihnen vorgeschlagene Dual-Process Signal-Detection Model (DPSDM) zur Prozeßdissoziationsprozedur zu erklären. Auch wenn das Argument bei Anwendung auf die Experimente von Buchner et al. (1995) nicht stichhaltig ist (vgl. Erdfelder & Buchner, 1998b), ist an der Logik des Arguments prinzipiell nichts auszusetzen. Käme man zu der Überzeugung, daß das Argument berechtigt ist, müßte man nach einer neuen experimentellen UV suchen, welche Antworttendenzen gezielter beeinflussen kann als es die Manipulation des Anteils alter Items im Rekognitionstest vermochte.

Wie ist zu verfahren, wenn ein Modell nur partiell, d.h. nur für einen oder einige Modellparameter den Validitätstest besteht? Es versteht sich von selbst, daß man ein Modell *generell* nicht verwenden sollte, wenn auch nur ein einziger relevanter Modellparameter den Validitätstest nicht bestanden hat. War die betreffende experimentelle Manipulation durchschlagskräftig genug, um den relevanten kognitiven Prozeß zu affizieren, und verhält sich der Zielparameter im Modell nicht hypothesengemäß, so muß man vermuten, daß das Modell inhaltlich fehlspezifiziert ist. Dies begründet den starken Verdacht, daß auch andere Modellparameter nicht reine Maße der psychologischen Größen sein können, die sie erfassen sollen. Dieser Verdacht wird nicht dadurch gegenstandslos, daß die betreffenden Modellparameter ihren „eigenen" Validitätstest in einem konkreten Anwendungsfall bestanden haben. Es kann sich hierbei um ein Zufallsergebnis handeln, das auf die jeweilige Auswahl der UVn zurückzuführen ist, die in das Validierungsexperiment eingehen. Mit anderen Worten: „Psychologische Gültigkeit" ist eine Eigenschaft von Modellen, nicht von einzelnen Parametern eines Modells. Es wäre daher mehr als fragwürdig, Parameter unterschiedlicher Modelle zu einer „Testbatterie" zusammenzustellen, die zwar ihren spezifischen Validitätstest bestanden haben, jedoch aus in ihrer Gesamtheit nicht validierten Modellen stammen. Parameter, die in unterschiedlichen Modellen definiert sind, sind nicht einfach gegeneinander austauschbar. Dies gilt auch dann, wenn sie inhaltlich ähnlich interpretiert werden, wie beispielsweise die Rekognitionsparameter des Ein- und des Zwei-Hochschwellen-Modells der Rekognition oder die entsprechenden Rateparameter beide Modelle. Am besten erkennt man das daran, daß man die aus einem Modell (z.B. dem 1HT-Modell) gewonnenen Parameterschätzwerte in die Modellgleichungen eines anderen Modells (z.B. des 2HT-Modells) für die inhaltlich korrespondierenden Parameter einsetzt. Die vorhergesagten Kategoriewahrscheinlichkeiten werden sich dann im Regelfall deutlich von den relativen Häufigkeiten der Stichprobe unterscheiden. Folglich kann nicht die Rede davon sein, daß die so gewonnenen Parameterschätzwerte die vorliegenden Daten erklären können.

Leider funktioniert das oben geschilderte Prinzip der Entscheidung zwischen datenäquivalenten Modellen bei äquivalenten Modellen oftmals nicht. Sind die einander korrespondierenden Parameter äquivalenter Modelle mit gleichen psychologischen Interpretationen ausgestattet, so hilft ein Validierungsprogramm nicht weiter, da die Vorhersagen für alle Modelle dieselben sind. Wegen der numerischen Gleichheit der Parameter müssen sich die Vorhersagen entweder bei allen Modellen bestätigen oder aber nicht bestätigen, so daß es auf diese Weise nicht gelingen kann, zwischen den Modellen empirisch zu diskriminieren. Als Ausweg bleibt nur eine Erweiterung der Modelle über rein multinomiale Modellierungsansätze hinaus. Beispielsweise erscheint es prinzipiell denkbar, daß die beiden äquivalenten MVB-Modelle aus Abbildung 3.4 empirisch diskriminierbar werden, wenn man sie auf Reaktionszeitmodelle über in der Zeit seriell ablaufende Prozesse erweitert (Luce, 1986; Townsend & Ashby, 1983). Die Modelle beziehen sich dann nicht mehr nur auf Häufigkeiten für die Beobachtungskategorien C_1 bis C_5, sondern zusätzlich auch auf die Antwortzeitverteilungen für die fünf Kategorien.

Eine günstigere Situation ist gegeben, wenn zwischen äquivalenten Modellen zu entscheiden ist, bei denen die numerisch einander korrespondierenden Parameter psychologisch unterschiedlich interpretiert werden. Dies ist im oben geschilderten Beispielfall des Zwei-Hochschwellen-Modells und des Modells von Jacoby (1991) der Fall. Nach dem Zwei-Hochschwellen-Modell sollte sich beispielsweise eine Manipulation von Antworttendenzen selektiv auf den b_{2ht}-Parameter auswirken. Ein entsprechender Effekt auf den korrespondierenden *u*-Parameter des Modells von Jacoby (1991) sollte demgegenüber nicht eintreten, wenn es sich bei *u* wirklich um ein „prozeßreines" Vertrautheitsmaß handelt. Man kann entsprechende Validierungsuntersuchungen zu Rekognitionstests (vgl. Snodgrass & Corwin, 1988) und zur Prozeßdissoziationsprozedur (vgl. Buchner et al., 1995) durchführen und auf diese Weise eine Entscheidung zwischen den Interpretationen herbeiführen. Die Experimente von Buchner et al. (1995) zeigen beispielsweise, daß Jacobys *u*-Parameter durchaus durch Antworttendenzen beeinflußt wird, so daß das Modell in der von Jacoby (1991) vorgeschlagenen Form kein gültiges Maß für Vertrautheit bzw. unbewußte Gedächtnisprozesse ist.[51]

Zur Verdeutlichung sollen abschließend einige Arbeiten exemplarisch erwähnt werden, die Validierungsexperimente berichten. Eine inzwischen schon klassische Validierungsstudie zu verschiedenen Rekognitionsmodellen ist die bereits erwähnte Arbeit von Snodgrass und Corwin (1988). In ihrem ersten Experiment präsentierten die Autoren den Vpn zu lernende Wörter mit hohem versus

[51]Umgekehrt gilt nicht unbedingt, daß Manipulationen unbewußter Verarbeitungsprozesse keinen Effekt auf Antworttendenzparameter ausüben dürfen. Windmann und Krüger (1998) haben jüngst argumentiert, daß ein hoher positiver Antwortbias auf subliminal dargebotene angstauslösende Wörter als Indikator unbewußter Entdeckungen von Bedrohungen angesehen werden kann.

niedrigem Bildhaftigkeitsgrad im Sinne von Paivio, Yuille und Madigan (1968). Im anschließenden Rekognitionstest mußten die Vpn alte und neue Wörter beider Bildhaftigkeitsstufen (UV1) unter drei Auszahlungsbedingungen beurteilen. Die Auszahlungsbedingungen sollten liberale, neutrale und konservative Entscheidungskriterien für „Alt"-Antworten induzieren (UV2). Beide UVn wurde innerhalb der Vpn (über mehrere Tage hinweg) orthogonal manipuliert. Für das Zwei-Hochschwellen-Modell und die Signalentdeckungstheorie in Verbindung mit d' als Sensitivitätsmaß und dem Antwortkriterium C (wobei $C = 0$ eine neutrale Tendenz indiziert) fielen die Validitätstest wie erwartet aus: Die Bildhaftigkeitsmanipulation beeinflußte selektiv die Rekognitionsparameter r_{2ht} und d' aus beiden Modellen, die Antworttendenzmanipulation dagegen ausschließlich die entsprechenden Antworttendenzmaße b_{2ht} und C. Alle anderen untersuchten Rekognitionsmodelle scheiterten an dem strengen Validitätskriterium, und zwar vor allem deshalb, weil deren Antworttendenzparameter nicht nur durch die Auszahlungsmanipulation, sondern zusätzlich auch noch durch die Bildhaftigkeitsmanipulation beeinflußt wurden.

Buchner et al. (1995) berichteten eine partielle Validierungsstudie zu zwei Prozeßdissoziationsmodellen: dem bereits erwähnten Modell von Jacoby (1991) und einem erweiterten Meßmodell (extended measurement model, EMM), das neben Maßen für bewußte Erinnerung (c) und Vertrautheitseffekte (u) separate Parameter für Antworttendenzen in der Inklusions- und der Exklusionsbedingung (g_i und g_e) bereitstellt. Möglich wurde das dadurch, daß nicht nur Antworten zu Items aus der ersten Itemliste im Modell berücksichtigt wurden, sondern zusätzlich auch Distraktoritems. Buchner et al. (1995) ging es vor allem um die Validität der Modellparameter in bezug auf Antworttendenzen. Auf dreierlei Weise wurden Antworttendenzen manipuliert (über den Anteil alter Items im Rekognitionstest, über unterschiedliche Auszahlungsmatrizen und durch Hinweise auf die Basisraten erforderlicher „Alt"-Antworten in der Inklusions- versus Exklusionsbedingung). In allen drei Experimenten erwies sich das EMM als besser als das Modell von Jacoby (1991), d.h. die Antworttendenzmanipulation wirkte sich entgegen der Validitätshypothese auf die Gedächtnisparameter c und u des Modells von Jacoby (1991) aus, nicht jedoch (oder zumindest nicht so stark) auf die entsprechenden Parameter des EMM. Zusätzlich wurde in jedem der drei Experimente die Tiefe der Verarbeitung der Items aus Liste 1 dadurch manipuliert, daß die Hälfte der Items nur gelesen werden mußten während die andere Hälfte als Anagramme dargeboten wurden, die zunächst aufgelöst werden mußten. Wie erwartet hatten die Anagramm-Items im EMM höhere bewußte Erinnerungsraten (c) als die nur gelesenen Items, während sich keine Effekte auf andere Modellparameter zeigten.

Insgesamt sind dies überzeugende Argumente für das EMM. Da die Antworttendenzmanipulation jedoch auch beim EMM zumindest in einer experimen-

tellen Bedingung (Manipulation der Auszahlungsmatrix in Experiment 2, Anagramm-Items) unerwarteterweise einen signifikanten Effekt auf die Gedächtnisparameter hatte, liegt es nahe, nach noch besseren Modellen zu suchen. Signalentdeckungstheoretische Alternativen zum EMM (Yonelinas & Jacoby, 1996a; Yonelinas et al., 1995) zeigen in empirischer Hinsicht keine wesentlichen Vorteile gegenüber dem EMM (Erdfelder & Buchner, 1998b). Eine Erweiterung des EMM, welches auf einer Ein-Hochschwellen-Konzeption für bewußte Rekognition basiert, erscheint jedoch vielversprechend: Zwei-Hochschwellen-Varianten des EMM führen zu deutlich besseren Validierungsergebnissen, und zwar sowohl bei entsprechender Reanalyse der Experimente von Buchner et al. (1995; vgl. Erdfelder & Buchner, 1995) als auch bei Analyse unabhängig gewonnener Datensätze (Vaterrodt-Plünnecke et al., 1996).

Die Ergebnisse zur Überlegenheit der Zwei-Hochschwellen-Konzeption des EMM im Vergleich zur Ein-Hochschwellen-Konzeption harmonieren sehr gut mit den Ergebnissen von Snodgrass und Corwin (1988) zu Ja-Nein-Rekognitionstests einerseits und mit denen von Bayen et al. (1996) zu Quellendiskriminationstests andererseits. Bayen et al. (1996) haben eine Kritik von Kinchla (1994) am Quellendiskriminationsmodell von Batchelder und Riefer (1990) zum Anlaß eines Validierungsexperiments genommen. Das Batchelder-Riefer-Modell zielt darauf ab, Effekte des Itemgedächtnisses, des Quellengedächtnisses und verschiedener Antworttendenzen auf das Antwortverhalten in Zwei-Quellen-Diskriminationstests (source monitoring tests) unter Zugrundelegung einer Ein-Hochschwellen-Konzeption des Item-Gedächtnisses zu erfassen (1HTSM-Modell, siehe oben). Kinchla (1994) kritisierte, daß sich das Ein-Hochschwellen-Modell für Ja-Nein-Rekognitionstests als inadäquat erwiesen habe (Kintsch, 1970; Murdock, 1974) und deshalb auch nicht zur Modellierung der Alt-Neu-Entscheidung innerhalb von Quellendiskriminationsmodellen Verwendung finden sollte. Obwohl dieses Argument für sich genommen nicht überzeugend ist, weil sich die relevanten psychologischen Prozesse in Rekognitionstests und Quellendiskriminationstests durchaus unterscheiden können (Batchelder & Riefer, 1999; Batchelder et al., 1994), zeigt es doch die Notwendigkeit einer Validierungsstudie auf. Bayen et al. (1996) verglichen deshalb die drei bereits erwähnten Quellendiskriminationsmodelle — nämlich das 1HTSM von Batchelder und Riefer (1990), eine Niedrigschwellen-Variante des Modells (LTSM) von Batchelder, Riefer und Hu (1994) und eine Zwei-Hochschwellen-Verallgemeinerung des 1HTSM-Modells — anhand eines zweifaktoriellen Experiments, in dem das Itemgedächtnis durch die Ähnlichkeit von alten Items und Distraktoren und das Quellengedächtnis durch die visuelle Ähnlichkeit der beiden Quellen beeinflußt werden sollte. Beide UVn des Designs wurden zwischen den Vpn und orthogonal zueinander variiert. Bei den beiden Quellen handelte es sich um zwei optisch ähnliche oder unähnliche Personen, die Aussagen zu

einem vermeintlichen Kriminalfall machten. Im anschließenden Quellendiskriminationstest bekamen die Vpn diese Aussagen (Zielitems) zusammen mit mehr oder minder ähnlichen Distraktoraussagen dargeboten und mußten jeweils angeben, ob die betreffende Aussage neu oder alt ist und von welcher Person sie gegebenenfalls stammt.

Es stellte sich heraus, daß nur das 2HTSM-Modell die Manipulation der Zielitem-Distraktor-Ähnlichkeit korrekt im Itemgedächtnisparameter und die der Quellenähnlichkeit korrekt im Quellengedächtnisparameter reflektierte. Die beiden anderen Modelle (1HTSM und LTSM) zeigten keinen signifikanten Effekt der Zielitem-Distraktor-Ähnlichkeit auf den Itemgedächtnisparameter und erfüllten somit nicht einmal das schwache Validitätskriterium. Zu unterstellen ist hierbei natürlich, daß man die Veränderung der Zielitem-Distraktor-Ähnlichkeit als geeignete Manipulation des Itemgedächtnisses akzeptiert, was aus der Sicht einer Ein-Hochschwellen-Konzeption vielleicht debattierbar ist. Möglicherweise verbirgt sich hinter dieser UV eher eine Manipulation der Antworttendenz in Richtung „Alt"-Antwort. Diese Vermutung wird durch die Ergebnisse von Bayen et al. (1996) insofern gestützt, als die Rate falscher Alarme signifikant mit der Zielitem-Distraktor-Ähnlichkeit anstieg. Eine exklusive Beeinflussung des Itemgedächtnisses lag also sicher nicht vor, sondern allenfalls eine kombinierter Effekt auf das Itemgedächtnis und die Antworttendenz.

Ungeachtet dieses Interpretationsproblems sprechen die Ergebnisse von Bayen et al. (1996) insgesamt eher für die Zwei-Hochschwellen-Konzeption des Itemgedächtnisses, weil allein diese Konzeption alle Befunde des Experiments ohne Post-hoc-Zusatzannahmen erklären kann. Die Ergebnisse harmonieren mit einer Reihe von anderen Studien, welche ebenfalls die Annahme bestätigen, daß Vpn offenbar durchaus Distraktoren als „sicher neu" entdecken können (z.B. Strack & Bless, 1994). Genau dies wird von Zwei-Hochschwellen-Konzeptionen vorausgesagt.

Ein weiteres Beispiel für die erfolgreiche Anwendung von Zwei-Hochschwellen-Konzeptionen ist die jüngst von Klauer und Wegener (1998) vorgeschlagene Erweiterung des 2HTSM-Modells auf das sogenannte „Who-said-what?"-Paradigma zur Untersuchung sozialer Kategorisierungsprozesse. Im „Who-said-what?"-Paradigma müssen Vpn Aussagen von vielen verschiedenen Personen, die zwei sozialen Gruppen entstammen (z.B. Männer vs. Frauen), nach einem Retentionsintervall ihren mutmaßlichen Urhebern zuordnen. Im multinomialen Modell von Klauer und Wegener (1998) wird das Quellengedächtnis in das Gedächtnis für die Person einerseits und das Gedächtnis für die soziale Gruppe, falls die Person nicht mehr erinnert werden kann, andererseits dekomponiert („Ich weiß nicht mehr, wer das gesagt hat, aber es war sicher eine Frau und kein

Mann"). Ansonsten ist das Modell weitgehend isomorph zum 2HTSM-Modell für Quellendiskriminationstests aufgebaut.

In einer extensiven Serie von Validierungsexperimenten zeigten Klauer und Wegener (1998), daß — im Einklang mit den Validierungshypothesen — (1) der Anteil von Distraktoritems im Gedächtnistest die „Neu"-Ratewahrscheinlichkeit *b*, (2) die Ähnlichkeit von alten Items und Distraktoren den Itemgedächtnisparameter *D*, (3) die Ähnlichkeit der sozialen Gruppen den Gruppendiskriminationsparameter *d*, (4) die Korrelation von Aussageinhalt und sozialer Gruppe die Ratewahrscheinlichkeit *a* für die Gruppenzugehörigkeit und schließlich (5) die Ähnlichkeit der Personen innerhalb einer sozialen Gruppe den Personendiskriminationsparameter *c* signifikant in die vorhergesagte Richtung beeinflußte[52], wobei alle diese Effekte exklusiv nur die Modellparameter betrafen, deren Manipulation angestrebt wurde. Somit war das strenge Validitätskriterium in fünf Tests ausnahmslos erfüllt. Darüber hinaus erwies sich das Modell als adäquat im Sinne der Datenanpassung, so daß es insgesamt sehr viel Argumente gibt, die dafür sprechen, das Modell zukünftig routinemäßig bei Untersuchungen mit dem „Who-said-What?"-Paradigma einzusetzen. Klauer und Wegener (1998) haben an dem Einfluß der kognitiven Beanspruchung auf Kategorisierungsprozesse (manipuliert durch die Darbietungsgeschwindigkeit der Items) zugleich exemplarisch aufgezeigt, daß konventionelle, varianzanalytische Auswertungen der Daten und modellbasierte Analysen zu ganz unterschiedlichen psychologischen Schlußfolgerungen führen können: Während konventionelle Analysen einen Effekt der kognitiven Beanspruchung auf die Gruppenenkodierung nahelegen, zeigt die modellbasierte Analyse keinen solchen Effekt. Letzteres legt den Schluß nahe, daß soziale Kategorisierung im wesentlichen ein automatischer Prozeß ist, der keine Aufmerksamkeitsressourcen erfordert. Genau das Gegenteil wird durch die konventionelle Analyse nahegelegt. Nach den von Klauer und Wegener (1998) vorgelegten Befunden muß man letzteres mit Skepsis betrachten, da die modellbasierte Analyse auf einem empirisch bewährten und psychologisch weitgehend validierten Modell aufbaut.

Ähnlich extensive Serien von Validierungsexperimenten wurden von Erdfelder und Buchner (1998a) zu einem multinomialen Hindsight-Bias-Modell (vgl. Abschnitt 7.1.9) und von Riefer und Batchelder (1995) zu einem Modell zum Recognition-Failure-Paradigma vorgelegt. Beide Modelle beziehen sich — anders als die zuvor geschilderten Modelle — nicht nur auf Rekognitions- sondern zusätzlich

[52]Der Personenähnlichkeitseffekt war allerdings unerwarteterweise auf Frauen beschränkt. Bei Männern war die Ähnlichkeitsmanipulation offenbar nicht durchschlagskräftig genug, d.h. auch die ähnlichen Männer konnten noch sehr gut diskriminiert werden. Klauer und Wegener (1998, S. 1168) argumentierten, daß dieser Geschlechterunterschied auf die durchschnittlich bessere visuelle Diskriminierbarkeit von Männern im Vergleich zu Frauen zurückzuführen ist, für die es in der Literatur Belege gibt.

auch auf Reproduktionsurteile von Vpn. Die Validierungsstudie von Riefer und Batchelder (1995) ist insofern besonders interessant, als sie eine einfache Möglichkeit aufzeigt, Validierungsprogramme schnell und ohne großen empirischen Aufwand zu realisieren. Voraussetzung ist lediglich, daß auf die Ergebnisse zuvor publizierter Arbeiten zum gleichen Paradigma zurückgegriffen werden kann. Hierzu müssen die Häufigkeiten der relevanten Beobachtungskategorien für möglichst viele der zuvor durchgeführten Studien publiziert vorliegen. Bei Riefer und Batchelder (1995) war das der Fall. Sie konnten auf ein Sammelreferat von Nilsson und Gardiner (1993) zurückgreifen, das einschlägige Häufigkeitsdaten für insgesamt 302 experimentelle Bedingungen zum Recognition-Failure-Paradigma zusammenfaßt. Diese Datenbank ermöglichte multiple Validierungsanalysen zu jedem der drei primär interessierenden Modellparameter des von Riefer und Batchelder (1995) vorgeschlagenen Modells (eine Speicherwahrscheinlichkeit, zwei Abrufparameter für Reproduktions- und Rekognitionstests). Auch dieses Modell wurde durch die Validierungsuntersuchungen in überzeugender Form gestützt. Ich werde auf das Modell und das zugrundeliegende Paradigma im folgenden Abschnitt 4.6 (Fall 3) ausführlicher eingehen.

4.6 Behandlung von Identifizierbarkeitsproblemen

In den bisherigen Abschnitten dieses Kapitels sind wir davon ausgegangen, daß ein zu analysierendes Modell wenigstens empirisch identifizierbar, d.h. an der Stelle des Parameterschätzwertes lokal identifizierbar ist. Gerade in der Frühphase einer Modellkonstruktion kommt es aber sehr häufig vor, daß ein kognitionspsychologisch gut begründbares Modell weder global noch lokal identifizierbar ist. Wie ist in solchen Fällen zu verfahren? Die Antwort auf diese Frage hängt von den konkreten Randbedingungen ab. Vier Situationen sind zu unterscheiden:

1) Interesse besteht allein an der Modellprüfung, die Parameterwerte sind von untergeordneter Bedeutung oder sogar völlig irrelevant;
2) Interesse besteht an der Modellprüfung und an einzelnen Parameterwerten oder speziellen Funktionen von Parameterwerten;
3) einzelne Parameter sind in psychologisch begründbarer Weise restringierbar, die Gültigkeit des restringierten Modells und die Werte der restlichen Parameter interessieren;
4) alle Parameterwerte interessieren, das empirische Paradigma ist erweiterbar.

Adäquate Methoden der Behandlung des Identifizierbarkeitsproblems für die vier genannten Situationen sollen im folgenden kurz besprochen werden. Lediglich wenn keiner der Lösungsvorschläge greift, muß man die Modellkonstruktion als vorläufig gescheitert ansehen und sich gegebenenfalls um einen komplett neuen Ansatz bemühen.

Fall 1): Die Parameterwerte sind von untergeordeter Bedeutung. Ist man nur an der Frage der Modellgültigkeit interessiert, nicht jedoch an der Schätzung der Parameterwerte, so kann man mit einem nicht identifizierbaren Modell problemlos weiterarbeiten, sofern es testbar ist. Natürlich kann ein nicht identifizierbares Modell durchaus testbar sein (Bamber & van Santen, 1985; 2000), d.h. das Bild des Parameterraumes $f(\boldsymbol{\Omega})$ kann eine echte Teilmenge von $\boldsymbol{\Gamma}$ sein, auch wenn f nicht injektiv ist. Das Ergebnis einer solchen Modellprüfung kann im Kontext einer psychologischen Hypothesen- oder Theorieprüfung sehr informativ und wichtig sein, auch wenn die Parameterwerte selbst nicht schätzbar sind (Erdfelder, 1990a, S. 485). So ist es beispielsweise für bestimmte Aussagen zur Geometrie des Sehraumes wichtig, daß die Beziehung zwischen subjektiver Größe, Netzhautbildgröße und Distanz eines Objekts durch eine additiv-verbundene Meßstruktur beschreibbar ist (vgl. Lukas, 1996); die konkreten Parameterwerte, welche den Einfluß von Netzhautbildgröße und Distanz auf die subjektive Größe beschreiben, sind dagegen nebensächlich.

Auf einen weiteren Punkt haben Bamber und van Santen (1985, S. 466) hingewiesen: Durch die Suche nach einem gemeinsamen Obermodell für mehrere identifizierbare Modelle kann sich unter Umständen ein sehr allgemeines Modell ergeben, das zwar nicht identifizierbar, aber dennoch testbar ist. Unter einem „Obermodell" ist das Pendant zu einem „Submodell" zu verstehen: Ein Modell M_1 ist ein Obermodell zu einem Modell M_0 genau dann, wenn M_0 Submodell zu M_1 ist, sich also durch Parameterrestriktionen aus M_1 herleiten läßt. Bamber und van Santen (1985, S. 451) präsentierten ein verallgemeinertes Drei-Gruppen-Schwellenmodell für dreistuftige Ratingskalen als Beispiel, das ein gemeinsames Obermodell von Hoch- und Niedrigschwellenmodellen für Reizentdeckungs- und Rekognitionstests ist. Ihr Obermodell umfaßt $S = 7$ funktional unabhängige Parameter und bezieht sich auf $M = 6$ funktional unabhängige Kategoriewahrscheinlichkeiten. Dieses Modell ist wegen $M < S$ natürlich weder lokal noch global identifizierbar (vgl. Abschnitt 4.1). Dennoch ist das Modell testbar. Das liegt daran, daß der maximale Rang der Jacobischen Matrix in diesem Fall $\rho_{max} = 5$ beträgt und damit kleiner als M ist (Bamber & van Santen, 1985, S. 459). Bamber und van Santen (1985, Theorem 9, S. 455) haben gezeigt, daß ein Modell genau dann quantitativ testbar ist, wenn ρ_{max} kleiner als die Anzahl funktional unabhängiger Kategoriewahrscheinlichkeiten M ist.[53] „Quantitativ testbar" bedeutet dabei, daß $f(\boldsymbol{\Omega})$ nicht nur eine echte Teilmenge von $\boldsymbol{\Gamma}$ ist, sondern zusätzlich auch noch gegenüber $\boldsymbol{\Gamma}$ dimensionsreduziert ist, so daß $f(\boldsymbol{\Omega})$ einem „infinitesimal dünnen Film" innerhalb von $\boldsymbol{\Gamma}$ entspricht. Wenn man nun plausibel machen kann, daß die faktisch

[53]Die von Bamber und van Santen (1985) vorgenommene Unterscheidung von schwacher und starker quantitativer Testbarkeit wird hier ausgeklammert, da sie für MVB-Modelle und viele andere parametrisierte multinomiale Modelle irrelevant ist.

vorliegenden Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ nicht in $f(\boldsymbol{\Omega})$ liegen, so hat man damit das Modell widerlegt. Dies wäre trotz fehlender Identifizierbarkeit der Modellparameter ein sehr informatives Ergebnis, denn die Inadäquatheit des Obermodells impliziert die Inadäquatheit aller identifizierbaren Submodelle dieses Modells. Man braucht sie also nicht mehr separat zu untersuchen, wenn das Obermodell widerlegt ist.

Allerdings ist die Nichtidentifizierbarkeit eines Modells bei der Wahl der Methode der Modellprüfung im Blick zu behalten. Die in Abschnitt 4.3 erläuterten inferenzstatistischen Verfahren setzen — wie bereits betont — die lokale Identifizierbarkeit des zu prüfenden Modells voraus. Bamber und van Santen (1985, S. 467) haben als Alternative einen allgemeinen Modellanpassungstest vorgeschlagen, der auch bei nicht identifizierbaren multinomialen Modellen einsetzbar ist. Er basiert auf der Neyman-modifizierten χ^2-Statistik $NM^2(\boldsymbol{\theta}, \boldsymbol{y})$, bei der im Nenner nicht durch die erwartete, sondern durch die beobachtete Häufigkeit in einer Kategorie dividiert wird (vgl. Gleichung (4.2.2-4)). $NM^2(\boldsymbol{\theta}, \boldsymbol{y})$ ist — wie schon erwähnt — mit $\mathrm{PD}^{\lambda=-2}(\boldsymbol{\theta}, \boldsymbol{y})$ äquivalent. Man wählt einen Parameterschätzwert $\hat{\boldsymbol{\theta}}$ so, daß diese modifizierte Statistik für gegebene Kategoriehäufigkeiten $\boldsymbol{y}$ minimiert wird. Daß der Schätzwert $\hat{\boldsymbol{\theta}}$ möglicherweise nicht eindeutig ist, ist in diesem Zusammenhang ohne Belang; es kommt allein auf den minimalen Wert der Teststatistik an. Bamber und van Santen (1985, S. 467, Theorem 23) haben gezeigt, daß die Statistik $NM^2(\hat{\boldsymbol{\theta}}, \boldsymbol{y})$ bei Modellgültigkeit unter sehr milden Regularitätsbedingungen asymptotisch zentral χ^2-verteilt ist, wobei als Freiheitsgrade nun allerdings nicht mehr $df = M-S$ zugrundezulegen sind (vgl. Abschnitt 4.3.1), sondern stattdessen $df = M-\rho_{max}$. Hierbei bezeichnet $\rho_{max} \leq S$ den maximalen Rang der Jacobischen Matrix der Modellgleichungen (vgl. hierzu Abschnitt 4.1.2 und zur Bestimmung von ρ_{max} Abschnitt 4.1.3). Praktisch bedeutet das, daß man die statistische Modellgeltungsprüfung bei nicht identifizierbaren Modellen durch Korrektur der Freiheitsgrade „nach oben" in den Griff bekommen kann, so als hätte man weniger Parameter geschätzt als tatsächlich im Modell vorkommen. Ist das Modell dagegen in einer Umgebung von $\hat{\boldsymbol{\theta}}$ lokal identifizierbar (d.h. $\rho[\mathbf{J}(\hat{\boldsymbol{\theta}}, f)] = S$, vgl. Abschnitt 4.1.2), so geht der von Bamber und van Santen (1985) vorgeschlagene Test in den gewöhnlichen asymptotischen Neyman-modifizierten χ^2-Test über (vgl. Abschnitt 4.3.1).

Ein äquivalentes Vorgehen, das immer zu denselben statistischen Entscheidungen führt, besteht darin, ein Ausgangsmodell $\langle \mathrm{M}(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}, \boldsymbol{\Gamma}, f \rangle$ mit S Parametern, welches testbar, aber wegen $\rho_{max} = S' < S$ nicht identifizierbar ist, durch willkürliche Fixierungen nicht identifizierbarer Parameter in ein identifizierbares Modell $\langle \mathrm{M}(K, \boldsymbol{J}, \boldsymbol{N}, \boldsymbol{\pi}), \boldsymbol{\Omega}', \boldsymbol{\Gamma}, f' \rangle$ mit S' Parametern zu überführen, derart, daß beide Modelle datenäquivalent sind, so daß $f(\boldsymbol{\Omega}) = f'(\boldsymbol{\Omega}') \subset \boldsymbol{\Gamma}$ gilt. Das erstgenannte Modell ist also genau dann gültig, wenn das restringierte Modell gültig ist. Beide

Modelle weisen notwendigerweise immer den gleichen Wert der Anpassungsstatistik auf (vgl. Abschnitt 4.4). Folglich kann man den Neyman-modifizierten χ^2-Test des identifizierbaren Modells mit S' Parametern auch als Test des nicht identifizierbaren Ausgangsmodells mit S Parametern interpretieren und vice versa. Wie man leicht sieht, ist die Wahl der Prüfstatistik bei diesem Vorgehen prinzipiell unerheblich. Um einen asymptotischen Modelltest zum Niveau α zu erhalten, kann man die NM^2-Teststatistik oder auch eine beliebige andere Minimum-PD$^\lambda$-Statistik wählen. Der entscheidende Punkt ist die Spezifikation der Freiheitsgrade: Man darf beim Test des Ausgangsmodells nicht — wie im Falle eines identifizierbaren Modells — $df = M-S$ zugrunde legen, sondern muß statt dessen $df = M-S'$ wählen, wobei $S' = \rho_{max}$.

Fall 2): Einige identifizierbare Parameter(funktionen) sind von Interesse. Wenn man zusätzlich zur Frage der Modellgültigkeit auch noch an den Werten einiger, aber nicht aller Modellparameter interessiert ist, muß fehlende Identifizierbarkeit des Modells kein zwingender Grund sein, vom gegebenen Modell abzurücken. Denkbar ist, daß sich die Nichtidentifizierbarkeit auf die nicht interessierenden Modellparameter beschränkt, während die psychologisch interessierenden Parameter identifizierbar sind. Ein parametrisiertes multinomiales Modell ist hinsichtlich eines bestimmten Parameters θ_S global identifizierbar, wenn

$$\begin{aligned}&\forall(\theta_1,...,\theta_s,...\theta_S),(\theta_1',...,\theta_s',...\theta_S') \in \mathbf{\Omega}: \\ &\qquad f(\theta_1,...,\theta_s,...\theta_S) = f(\theta_1',...,\theta_s',...\theta_S') \Rightarrow \theta_s = \theta_s'\end{aligned} \tag{4.6-1}$$

gilt. Analog ist lokale Identifizierbarkeit eines bestimmten Parameters θ_S an der Stelle $\boldsymbol{\theta^0} \in \mathbf{\Omega}$ gegeben, wenn es einen offene Umgebung $\mathbf{U} \subseteq \mathbf{\Omega}$ zu $\boldsymbol{\theta^0} \in \mathbf{U}$ gibt, so daß

$$\begin{aligned}&\forall(\theta_1,...,\theta_s,...\theta_S),(\theta_1',...,\theta_s',...\theta_S') \in \mathbf{U}: \\ &\qquad f(\theta_1,...,\theta_s,...\theta_S) = f(\theta_1',...,\theta_s',...\theta_S') \Rightarrow \theta_s = \theta_s'\end{aligned} \tag{4.6-2}$$

gilt. Offenbar sind globale und lokale Identifizierbarkeit im Sinne der Bedingungen (4.1.1-1) und (4.1.1-3) äquivalent mit globaler bzw. lokaler Identifizierbarkeit *aller* Modellparameter θ_s, $s = 1, ..., S$, im Sinne von (4.6-1) bzw. (4.6-2). Der Nachweis für die Identifizierkeit einzelner Parameter folgt dem Prinzip des Identifizierbarkeitsnachweises für das gesamte Modell. Der globale Identifizierbarkeitsnachweis für einen bestimmten Parameter ist natürlich insofern einfacher, als er bereits erbracht ist, wenn man diesen Parameter als Funktion der Kategoriewahrscheinlichkeiten darstellen kann. So ist etwa der *a*-Parameter aus dem Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986) trivialerweise global identifizierbar, weil $a = Pr(C_{2,1}) = p_{2,1}$ gelten muß (vgl. Abbildung 3.3).

Sind die interessierenden Parameter eines Modells global oder lokal identifizierbar, so können die betreffenden Parameterschätzungen problemlos interpretiert werden, auch wenn andere Parameter des Modells nicht identifizierbar sein sollten. Bei der Modellprüfung ist allerdings ein Modell, das auch nur hinsichtlich eines einzigen Parameters nicht identifizierbar ist, wie ein Modell zu behandeln, daß hinsichtlich aller Modellparameter nicht identifizierbar ist. Anstelle von Standardmethoden der Modellprüfung muß also beispielsweise auf den oben geschilderten Modelltest für nicht identifizierbare Modelle zurückgegriffen werden, d.h. die Freiheitsgrade berechnen sich nicht nach $df = M - S$, sondern statt dessen nach $df = M - \rho[\mathbf{J}(\hat{\boldsymbol{\theta}}, f)]$.

Richtet sich das Interesse nicht auf einzelne Parameter, sondern auf Funktionen von Modellparametern (z.B. Summen, Differenzen oder Produkte zweier Parameter), so gelten die obigen Ausführungen analog. Funktionen von Parameterschätzungen können problemlos interpretiert werden, sofern die entsprechenden Funktionen (mindestens lokal) identifizierbar sind. Globale und lokale Identifizierbarkeit einer reellwertigen Funktion $\varphi(\boldsymbol{\theta})$ von $\boldsymbol{\theta}$ ist dabei ganz analog zu (4.6-1) und (4.6-2) zu definieren. Denkbar ist somit, daß kein einziger Modellparameter identifizierbar ist, statt dessen jedoch eine oder mehrere Funktionen der Modellparameter. Beim unrestringierten BTL-Modell sind die Attraktivitätsparameter θ_s beispielsweise weder global noch lokal identifizierbar (vgl. Abschnitt 4.4). Die Funktionen $\varphi_s(\theta_s / \theta_1)$, $s = 2, ..., m$, und auch die Quotienten beliebiger anderer Paare von Attraktivitätsparametern sind dagegen global identifizierbar. Aussagen wie „Objekt *s* ist fünf mal so attraktiv wie Objekt 1" sind also bei Bezugnahme auf BTL-Modellparameter substanzwissenschaftlich sinnvoll und im Falle $\varphi_s(\theta_s / \theta_1) = 5$ auch empirisch wohlbegründet, obwohl keiner der Attraktivitätsparameter für sich genommen identifizierbar ist. Man sieht an diesem Beispiel, wie eng das Bedeutsamkeitsproblem der Meßtheorie (Niederée & Mausfeld, 1996; Steyer & Eid, 1993; Suppes & Zinnes, 1963) mit dem Identifizierbarkeitsproblem bei Meßmodellen zusammenhängt. Grob kann man sagen, daß Aussagen über numerische Größen dann sinnvoll oder „bedeutsam" sind, wenn die in der Aussage verwendeten Funktionen numerischer Größen identifizierbar sind.

Fall 3): Einige Parameter sind restringierbar. Die Fälle 1) und 2) sind bei Anwendungen multinomialer Modelle in der Kognitionspsychologie im allgemeinen selten gegeben. Häufiger ist der Fall, daß die interessierenden Modellparameter erst dann identifizierbar werden, wenn man einem oder mehreren der nicht interessierenden Parametern bestimmte Restriktionen auferlegt. Typisch sind hier etwa Gleichheitsrestriktionen oder Parameterfixierungen (vgl. Abschnitt 4.3.2). Beides führt dazu, daß die Anzahl *S* freier Parameter reduziert wird, mit dem Ergebnis, daß beispielsweise ein überparametrisiertes Modell in ein saturiertes oder gar unterparametrisiertes Modell übergeht. So wird etwa das auf alte Items beschränkte

1HT-Modell (vgl. Abschnitt 4.1) zu einem saturierten und zugleich auch global identifizieren Modell mit $S = (J–1) = 1$, wenn man die Ratewahrscheinlichkeit b_{1ht} auf einen bestimmten Wert fixiert, beispielsweise auf $b_{1ht} = .5$. Man schränkt mit dieser Restriktion die vielen Punkte, die im unrestringierten Modell auf eine bestimmte Trefferwahrscheinlichkeit abgebildet werden (z.B. $p_{11} = .8$, vgl. Abbildung 4.1) auf denjenigen ein, bei dem $b_{1ht} = .5$ gilt.

Dieses Vorgehen scheint auf den ersten Blick die einfachste Möglichkeit zu sein, Identifizierbarkeit herzustellen. Es hat allerdings einen gravierenden Nachteil: Der Wert, der sich nach Herstellung der Identifizierbarkeit für die interessierenden Modellparameter ergibt, hängt wesentlich von der Art der Restriktion auf die nichtinteressierenden Modellparameter ab. Im Beispiel des 1HT-Modells für alte Items ergibt sich z.B. immer ein anderer Wert für den primär interessierenden Vertrautheitsparameter r_{1ht}, wenn man b_{1ht} auf unterschiedliche Werte fixiert. Damit dies nicht zu einem neuen Bedeutsamkeitsproblem in der psychologischen Forschung führt (vgl. Abschnitt 2.4), muß mindestens eine der beiden folgenden Bedingungen erfüllt sein:

1) Die Restriktion der nicht interessierenden Parameter muß psychologisch gut begründet sein, am besten so, daß damit eine Behauptung verbunden ist, die empirisch testbar ist.[54]
2) Man muß den Nachweis erbringen, daß die Art der Restriktion zwar den Wert bzw. die Werte der interessierenden Parameter beeinflußt, aber nicht so gravierend, daß die psychologischen Schlußfolgerungen, die aus empirischen Modellanwendungen resultieren, damit in Frage gestellt werden.

Ein sehr schönes Anwendungsbeispiel für den zweiten Weg ist die bereits erwähnte Arbeit von Riefer und Batchelder (1995), die ein BVB-Modell zum sogenannten Recognition-Failure-Paradigma (RF-Paradigma) vorgestellt haben. Im RF-Paradigma werden A-B-Itempaare gelernt, z.B. „Wissenschaftler — Darwin". Später folgt zunächst ein Rekognitionstest für die B-Items und dann ein geförderter Reproduktionstest (cued recall) für die B-Items, bei dem die A-Items als Abrufhilfen dargeboten werden. Wenn man nun für verschiedene Itempaare und Vpn die bedingte Wahrscheinlichkeit $p(\mathrm{Rn}\,|\,\mathrm{Rc})$ einer erfolgreichen Rekognition, gegeben eine erfolgreiche geförderte Reproduktion eines B-Items aus Stichprobendaten schätzt und gegen den Schätzer für die Trefferrate $p(\mathrm{Rn})$ im Rekognitionstest abträgt, ergibt sich eine leicht kurvilineare Beziehung, die als Tulving-Wiseman-Funktion bekannt geworden ist (Flexser & Tulving, 1978; Tulving & Wiseman, 1975). Sie kommt einer linearen Beziehung mit Steigung 1 recht nahe, was offenbar die sogenannte Generate-Recognize-Theorie der (geförderten) Reproduktion

[54] Wählt man diesen Weg, so läuft das letztlich auf eine Erweiterung des empirischen Paradigmas hinaus, d.h. man hat es automatisch mit Situation 4) zu tun, die weiter unten behandelt wird.

widerlegt, wonach erfolgreiche Reproduktion immer auch erfolgreiche Rekognition impliziert (d.h. $p(\mathrm{Rn}\,|\,\mathrm{Rc}) = 1$).

Riefer und Batchelder (1995) haben nun ein Modell entwickelt, das die Theorie unabhängiger Abrufprozesse (retrieval-independence theory) von Flexser und Tulving (1978) formalisiert. Kernidee ist die stochastische Unabhängigkeit von Abrufprozessen bei der geförderten Reproduktion und bei der Rekognition: Leichte Zusammenhänge zwischen $p(\mathrm{Rn}\,|\,\mathrm{Rc})$ und $p(\mathrm{Rn})$ ergeben sich nur deshalb, weil nicht alle dargebotenen Itempaare auch gespeichert werden. Das Modell von Riefer und Batchelder (1995) umfaßt folglich eine Speicherwahrscheinlichkeit s für die A-B-Paare, zwei unabhängige[55] Abrufwahrscheinlichkeiten r_1 und r_2 für den Rekognitionstest und den geförderten Reproduktionstest sowie zusätzlich eine Ratewahrscheinlichkeit g. Der letztgenannte Parameter entspricht der Wahrscheinlichkeit einer „Alt"-Antwort im Rekognitionstest, wenn ein Item nicht abrufbar ist. Das ergibt insgesamt $S = 4$ Parameter. Dem stehen nur vier Beobachtungskategorien und damit $M = 4 - 1 = 3$ funktional unabhängige Kategoriewahrscheinlichkeiten gegenüber, die aus der Kreuzklassifikation der dichotomen Ereignisse „erfolgreiche Rekognition" (ja / nein) und „erfolgreiche geförderte Reproduktion" (ja / nein) resultieren. Kein Parameter dieses überparametrisierten Modells ist identifizierbar. Wenn jedoch die psychologisch nicht interessierende Ratewahrscheinlichkeit g auf einen bestimmten Wert fixiert wird, wird das restringierte Modell global identifizierbar. Riefer und Batchelder (1995) wählten $g = .0$, zeigten aber zugleich auf, daß ihre Interpretationen der Ergebnisse von Nilsson und Gardiner (1993) nicht wesentlich tangiert werden, wenn tatsächlich andere Werte für g zutreffen sollten. Zwar ändern sich die Parameterschätzungen für s, r_1 und r_2, wenn man g auf andere Werte fixiert. Die vorhergesagten Korrelationen zwischen (a) der Trefferrate und der Quote erfolgreicher freier Reproduktionen sowie (b) den r_1- und r_2-Schätzern sind kaum betroffen, ebenso wie die Form der Tulving-Wiseman-Funktion, die aus dem Modell von Riefer und Batchelder (1995) ableitbar ist. Vorauszusetzen ist allerdings, daß Faktoren, die auf die Höhe von g Einfluß nehmen, nicht auch andere Modellparameter beeinflussen und vice versa (vgl. Riefer & Batchelder, 1995, S. 624). Diese Annahme ist im gegebenen Anwendungsfall jedoch plausibel. Damit erscheint die Parameterfixierung $g = .0$ pragmatisch vertretbar, auch wenn sie faktisch falsch sein sollte. Gravierende Fehlinterpretationen der Ergebnisse sind aufgrund der Robustheitsanalysen von Riefer und Batchelder (1995) nicht zu erwarten.

Fall 4): *Das empirische Paradigma ist erweiterbar*. Die im letzten Abschnitt genannten Absicherungsmaßnahmen bei Parameterfixierungen sind zwar

[55] „Unabhängig" heißt hier, daß die Abrufwahrscheinlichkeit r_2 immer gleich groß ist, egal ob der Abruf im Rekognitionstest erfolgreich war oder nicht. Zur Repräsentation stochastisch unabhängiger Ereignisse in MVB-Modellen vgl. insbesondere Abschnitt 4.3.2.

hilfreich, können aber nicht sicherstellen, daß Parameterschätzungen in konkreten Einzelfällen sinnvoll interpretierbar sind. Die Robustheitsanalysen von Riefer und Batchelder (1995) hätten z.B. ebensogut zu dem Ergebnis führen können, daß die Wahl von g wesentlichen Einfluß auf die psychologischen Schlußfolgerungen aus den Ergebnissen hat. Dann wäre eine willkürliche Parameterfixierung natürlich äußerst problematisch. Aus diesem Grund empfiehlt es sich, wann immer möglich nach Erweiterungen des empirischen Paradigmas zu suchen, welche die Anzahl M funktional unabhängiger Kategoriewahrscheinlichkeiten erhöhen, ohne zugleich eine Erweiterung des Parameterraumes $\mathbf{\Omega}$ zu erzwingen. Es ist also nicht damit getan, die Anzahl der Beobachtungskategorien „künstlich" zu erhöhen, indem man etwa im Falle des RF-Paradigmas zusätzlich noch Sicherheitsurteile zu den Reproduktions- und Rekognitionsdaten erhebt. Diese erhöhen zwar M, erzwingen aber zusätzliche Modellparameter zur Modellierung des Sicherheitsurteils, was im allgemeinen nicht in ein identifizierbares Modell einmünden wird.

Die Kunst „adäquater" Erweiterungen des empirischen Paradigmas besteht gerade darin, zusätzliche unabhängige Beobachtungsbedingungen zu finden, in die jedoch keine zusätzlichen psychologischen Prozesse einfließen, sondern nur diejenigen, die auch schon im Ausgangsparadigma wirksam sind. Beim Modell von Riefer und Batchelder (1995) könnte man etwa daran denken, im Rekognitionstest neben alten Items auch neue Items zu präsentieren und das Modell dann dahingehend zu ergänzen, daß die Rate falscher Alarme im Rekognitionstest entweder mit g oder mit $(1-r_1)\cdot g$ gleichgesetzt wird, was einem 1HT- bzw. einem 2HT-Konzept des Rekognitionsgedächtnisses entspräche. Genau diesen Weg haben auch Klauer und Wegener (1998) gewählt, um ihr multinomiales Modell zum „Who-said-what"-Paradigma in ein identifizierbares Modell zu überführen.

Sowohl die 1HT- als auch die 2HT-Erweiterung auf falsche Alarme würde das Vier-Parameter-Modell von Riefer und Batchelder (1995) in ein global identifizierbares Modell überführen. Faktisch entspricht die 1HT-Erweiterung dem oben genannten Fall 3), da die Gleichsetzung $g = Pr$(„alt"; neues Item) einer Fixierung des Parameters g auf die in der Stichprobe beobachtete relative Häufigkeit falscher Alarme gleichkommt. In diesem Fall läßt sich die Parameterfixierung gut begründen, wenn man einmal die Grundannahme von 1HT-Konzeptionen akzeptiert, daß die Rate falschen Alarms ein geeigneter Schätzer der Ratewahrscheinlichkeit ist.

Natürlich kann man auf die gleiche Weise auch das auf alte Items beschränkte Rekognitionsmodell aus Abbildung 4.1 in ein identifizierbares Modell überführen. Hierzu wird es einfach auf neue Items erweitert und z.B. die Annahme ergänzt, daß die Rate falschen Alarms mit der Ratewahrscheinlichkeit $b_{1\mathrm{ht}}$ identisch ist. Man erhält dann das volle 1HT-Modell, das den gleichen Parameterraum $\mathbf{\Omega}$ wie das in Abbildung 4.1 dargestellte Modell umfaßt, jedoch einen um

eine Dimension — die Rate falschen Alarms — erweiterten Raum $\mathbf{\Gamma}$. Das Modell ist mit $S = M = 2$ Parametern vollparametrisiert und außerdem global identifizierbar. Dies folgt unmittelbar aus der Existenz der Umkehrfunktion f^{-1} zu den Modellgleichungen (3.5.2-1) und (3.5.2-2):

$$r_{1\mathrm{ht}} = \frac{p_{1,1} - p_{2,1}}{1 - p_{2,1}} = \frac{Pr(\text{Treffer}) - Pr(\text{falscher Alarm})}{1 - Pr(\text{falscher Alarm})} \tag{4.6-3}$$

$$b_{1ht} = p_{2,1} = Pr(\text{falscher Alarm}) \,. \tag{4.6-4}$$

Abbildung 4.10 veranschaulicht die Tatsache, daß das volle 1HT-Modell jede Strecke im Parameterraum, die durch eine konstante Vertrautheitswahrscheinlichkeit $r_{1\mathrm{ht}}$ bei variablen Ratewahrscheinlichkeiten $b_{1\mathrm{ht}} \in [0, 1]$ gekennzeichnet ist, auf eine andere ROC-Kurve in $\mathbf{\Gamma}$ abbildet. Bei den ROC-„Kurven" handelt es sich in diesem Fall um unterschiedliche Geraden, die alle durch den Punkt $(1, 1) \in \mathbf{\Gamma}$ gehen. Ihr Schnittpunkt mit der Ordinate entspricht gerade dem Wert des Parameters $r_{1\mathrm{ht}}$. Von jedem Punkt im „ROC-Raum" $f(\mathbf{\Omega})$, der trotz $S = M$ eine echte Teilmenge von $\mathbf{\Gamma}$ mit $p_{1,1} \geq p_{2,1}$ ist, kann man eindeutig zurückverfolgen, welche Parameter $r_{1\mathrm{ht}}$ und $b_{1\mathrm{ht}}$ ihn „erzeugt" haben.

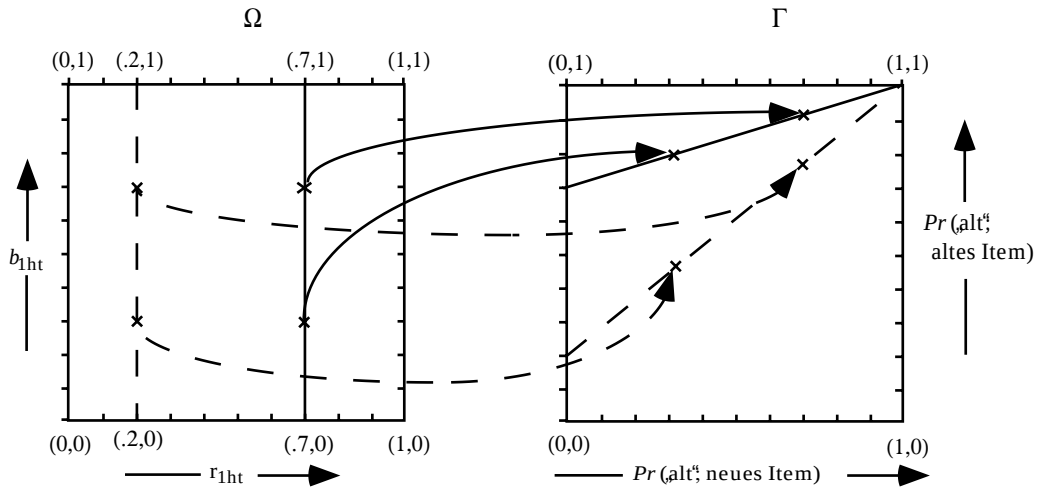

Abbildung 4.10:
Veranschaulichung der Abbildung des Parameterraums $\mathbf{\Omega}$ in den Raum möglicher Kategoriewahrscheinlichkeiten $\mathbf{\Gamma}$ beim Ein-Hochschwellen-Modell für Ja-Nein-Rekognitionstests mit alten und neuen Items.

Der oben benutzte Begriff einer „adäquaten" Erweiterung des Paradigmas mag fälschlich suggerieren, als gebe es im Regelfall nur einen Weg der Erweiterung des Ausgangsparadigmas, der zum Ziel, d.h. zu einem identifizierbaren Modell führt. Dies ist jedoch nicht der Fall. Es gibt fast immer mehrere Wege, nicht identifizierbare Modelle durch Paradigmenerweiterungen in identifizierbare Modelle zu überführen. Wenn man sich auf bestimmte zusätzliche Beobachtungsbe-

dingungen festgelegt hat — beispielsweise auf die Einbeziehung von Distraktoritems in Rekognitionstests —, dann sind zumeist unterschiedliche Modellerweiterungen auf diese zusätzlichen Beobachtungsbedingungen denkbar, die zu nichtäquivalenten identifizierbaren Modellen führen. Beispielsweise kann das Modell aus Abbildung 4.1 nicht nur zu einem 1HT-Modell, sondern alternativ auch zu einem 2HT-Modell erweitert werden. Beide Modelle wären identifizierbar und datenäquivalent, aber nicht äquivalent. Will man entscheiden, welcher der beiden Ansätze die bessere Erweiterung zu einem identifizierbaren Modell liefert, muß man Validierungsexperimente wie in Abschnitt 4.5 beschrieben durchführen. Konkret könnte das bei der Entscheidung zwischen unterschiedlichen Rekognitionsmodellen so aussehen, daß man das Rekognitionsgedächtnis und die Antworttendenz durch geeignete experimentelle Manipulationen orthogonal zueinander variiert. Man schaut dann nach, ob die empirisch ermittelten ROC-Kurven die Form der Geraden aus Abbildung 4.10 haben oder eher Geraden mit Steigung 1 (d.h. Parallelen zur Winkelhalbierenden in $\mathbf{\Gamma}$) entsprechen, wie es vom 2HT-Modell vorhergesagt werden würde. Vorliegende Befunde sprechen dafür, daß letzteres der Fall ist (vgl. Abschnitt 4.5), so daß die 2HT-Erweiterung des Rekognitionsmodells aus Abbildung 4.1 vorzuziehen ist.

Nicht immer fällt die Erweiterung des Ausgangsparadigmas so leicht wie im Falle von Ja-Nein-Rekognitionstests. Bei dem in Abschnitt 2.6 vorgestellten Modell zum Effekt nachträglicher Falschinformation, das sich auf einen Forced-Choice-Rekognitionstests für eine Versuchs- und eine Kontrollgruppe bezieht (Loftus-Paradigma), ist z.B. nicht ohne weiteres klar, welche Beobachtungsbedingungen hinzugenommen werden müssen, damit das Drei-Parameter-Modell identifizierbar wird. Unter Umständen ist man dann gezwungen, vom Idealweg einer „parameterraumneutralen" Paradigmenerweiterung abzurücken und statt dessen Erweiterungen vorzunehmen, die nicht nur die Anzahl der Beobachtungsbedingungen, sondern zusätzlich auch die Anzahl der Modellparameter erhöhen. Auch dieser Weg kann jedoch zielführend sein, wenn — grob gesprochen — dem Gewinn an zusätzlichen Beobachtungsbedingungen ein vergleichsweise geringer Anstieg der Parameterzahl gegenübersteht. Ein Lösungsansatz, der praktisch in jedem Fall anwendbar ist, ist die Erweiterung des Ausgangsparadigmas zu einem mehrfaktoriellen Design, verknüpft mit einer Erweiterung des multinomialen Modells zu einem verbundenen Modell für mehrere Bedingungen des Designs. Ein solcher Ansatz führt fast immer zu einem identifizierbaren Gesamtmodell, wenn vorausgesetzt werden kann, daß die Faktoren des Designs selektiv einzelne Modellparameter affizieren (vgl. Abschnitt 4.5). Wenn zwei Parameter des Ausgangsmodells z.B. durch zwei experimentelle Faktoren in m_1 bzw. m_2 Stufen selektiv manipuliert werden, steht dem „Gewinn" an $m_1 \cdot m_2 - 1$ zusätzlichen Beobachtungsbedingungen beispielsweise ein „Verlust" von nur $(m_1 - 1) + (m_2 - 1)$ zusätzlichen

Modellparametern gegenüber. Schon ab $m_1 = m_2 = 2$ Stufen der Faktoren übertrifft die Anzahl zusätzlicher Beobachtungsbedingungen die Anzahl zusätzlicher Modellparameter. Dies ist natürlich noch keine Garantie dafür, daß das Gesamtmodell für den kompletten Versuchsplan identifizierbar ist (Bamber & van Santen, 1985; 2000), macht dies jedoch eher wahrscheinlich.

Der Weg zu einem identifizierbaren Modell ist beim zuletzt diskutierten Vorgehen offenbar unmittelbar an ein Validierungsexperiment gekoppelt. Erst durch Einbettung des Modells in ein mehrfaktorielles Validierungsdesign erhält man ein identifizierbares Gesamtmodell. Genau diesen Weg kann man auch bei der Überführung des Modells zum Loftus-Paradigma in ein identifizierbares Modell wählen. Ein entsprechendes Experiment von Loftus et al. (1978) liefert die notwendigen Daten. Man muß hierzu natürlich voraussetzen, daß die Faktoren des Designs selektiv einzelne Modellparameter beeinflussen. Details hierzu werden in Abschnitt 10.1 berichtet.

Nicht immer ist es erforderlich, eine Erweiterung auf ein mehrfaktorielles Design vorzunehmen, wenn man ein identifizierbares Modell erhalten will. Manchmal reicht es bereits aus, einen einzigen Faktor — z.B. die Antworttendenz zugunsten einer bestimmten Antwort — in zwei oder mehr Stufen zu variieren, sofern vorausgesetzt werden kann, daß diese Manipulation nur bestimmte Modellparameter affiziert und andere untangiert läßt. Ein konkretes Beispiel hierfür ist ein MVB-Modell zum Glaubwürdigkeitseffekt beim syllogistischen Schlußfolgern (Klauer, Musch & Naumer, 1999), welches in Abschnitt 7.3.1.3 ausführlicher besprochen wird.

4.7 Computerprogramme

Computerprogramme sollen nur knapp geschildert werden, da zu erwarten ist, daß eine vergleichende Evaluation der Programme zum gegenwärtigen Zeitpunkt schnell überholt wäre. Es ist absehbar, daß die vorhandene Programme — zumindest teilweise — zunehmend optimiert und im Leistungsumfang verbessert werden. Außerdem werden sicher weitere Programme hinzukommen. Ein Programm, das gegenwärtig im Vergleich sehr gut dasteht, kann daher in naher Zukunft schon suboptimal sein.

Drei Programmklassen können gegenwärtig unterschieden werden: (1) Programme für spezielle MVB-Modelle, (2) Programme zur gesamten Klasse binärer oder allgemeiner MVB-Modelle und (3) Programme für beliebige parametrisierte Modelle. Zur ersten Klasse gehört das DOS-Programm SOURCE von Hu (1990), welches ML-Parameterschätzungen und Log-Likelihood-Quotienten-Tests für die Submodelle des 1HTSM-Quellendiskriminationsmodells von Batchelder und Riefer (1990) ermöglicht. Das Modell kann Modellanpassungstests und Parametertests

bei Mehr-Gruppen-Vergleichen durchführen. In etwa den gleichen Leistungsumfang, jedoch bezogen auf das Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986), hat das DOS-Programm SRM von Erdfelder (1992b). Beide Programme verwenden den EM-Algorithmus der ML-Schätzung, sind sehr leicht zu bedienen und ermöglichen beliebige Vergleiche zwischen mehreren experimentellen Gruppen. Ihr gravierendster Nachteil ist ihre Unflexibilität; Änderungen der Modellgleichungen gegenüber den publizierten Modellen sind nicht möglich. Gerade beim 1HTSM-Quellendiskriminationsmodell ist dies ein Problem, da Validierungsexperimente die Überlegenheit des 2HTSM-Modells im Vergleich zum 1HTSM-Modell anzudeuten scheint (Bayen et al., 1996; Erdfelder et al., 1995).

Wichtiger für die Anwendungspraxis sind Programme der zweiten Klasse, die statistische Analysen für beliebige binäre oder nichtbinäre MVB-Modelle ermöglichen. Drei Modelle sind gegenwärtig verfügbar: Das DOS-Programm MBT (Multinomial Binary Tree) von Hu (1995; 1999b), das Apple-Macintosh Programm „AppleTree" von Rothkegel (1998; 1999) und das Windows-Programm GPT (General Processing Tree, vgl. Hu, 1998; 1999a; Hu & Phillips, 1999). Alle drei Programme bieten beliebige Minimum-PD$^{\lambda}$-Prüfstatistiken und Parameterschätzungen an; sie verwenden den EM- bzw. den ε-korrigierten EM$^{\lambda}$-Algorithmus für binäre (MBT, AppleTree) und auch nichtbinäre MVB-Modelle (GPT). Der Leistungsumfang kann Tabelle 4.3 entnommen werden. Grundprinzip der Programmbenutzung ist die Erstellung zweier ASCII-Eingabedateien, einer EQN-Datei, die das zu analysierende Modell beschreibt, und einer MDT-Datei, in der die zu analysierenden Daten enthalten sind. Im Hauptfenster des betreffenden Programms sind dann methodische Details der Analyse der MDT-Datei auf der Grundlage der EQN-Datei zu spezifizieren (λ-Wert, gewünschte Parameterrestriktionen, gewünschte Elemente der Ausgabedatei etc.). Im Gegensatz zu MBT und AppleTree verfügt GPT alternativ über graphische Eingabemöglichkeiten für das Modell. Die Benutzung der Programme über EQN- und MDT-Dateien ist allerdings zeitsparender. Sie wird im Anhang an zahlreichen Beispielen erläutert.

Keines der oben erwähnten Programme kann parametrisierte multinomiale Modelle analysieren, die nicht zur Familie der MVB-Modelle gehören. Folglich stößt schon die numerische Umsetzung der ML-Analyse des einfachen Signalentdeckungstheoriemodells, angewandt auf eine komplette ROC-Kurve mit mehreren Datenpunkten, auf Schwierigkeiten. Zwar existieren für diesen speziellen Fall Computerprogramme (Dorfman & Alf, 1969; Dorfman & Berbaum, 1986), aber sie sind ähnlich unflexibel wie die oben erwähnten Programme SOURCE und SRM, so daß naheliegende Modellmodifikationen, wie etwa das DPSDM-Modell von Yonelinas und Jacoby (1996), damit nicht analysierbar sind. Ein flexibles Analogon zu MBT, GPT und AppleTree, das auf die *gesamte* Klasse parametrisierter multinomialer Modelle zugeschnitten ist, scheint es tatsächlich noch nicht zu geben.

Auch wenn in dieser Arbeit fast ausschließlich MVB-Modelle besprochen werden, ist das eine ernstzunehmende Einschränkung, denn es gibt keinen erkennbaren Grund, warum Formalisierungen kognitionspsychologischer Theorien als multinomiale Modelle immer in MVB-Modelle einmünden sollten.

Tabelle 4.4:
Übersicht über den Leistungsumfang und die Benutzerfreundlichkeit gegenwärtig verfügbarer flexibler Programme zur statistischen Analyse multinomialer Modelle.

	Programm				
Funktion / Eigenschaft	MBT	AppleTree	GPT	PRAXIS	EXCEL
Betriebssystem	DOS	Mac-OS	Windows	DOS*	fast alle
Leistungsmerkmale:					
geeignet für Modellklasse	bin. MVB	bin. MVB	allg. MVB	alle	alle
G^2 & Min.-PD$^{\lambda}$-Modelltest?	ja	ja	ja	ja	ja
ML- und Min.-PD$^{\lambda}$-Schätzer?	ja	ja	ja	ja	ja
Andere Tests und Schätzer?	nein	nein	nein	ja	ja
Parameterrestriktionen?	ja	ja	ja	ja	ja
Varianz-Kovarianz-Matrix?	ja	ja	ja	nein	nein
Konfidenzintervalle?	ja	ja	ja	nein	nein
Zufallsstartwertoption?	ja	ja	ja	nein	ja**
Identifizierbarkeitscheck?	ja	ja	ja	nein	ja**
Monte-Carlo-Stichproben?	ja	nein	ja	nein	nein
Teststärkeberechnung?	nein	ja	ja	nein	nein
Rang der Jacobischen Matrix?	nein	nein	nein	nein	nein
Test f. nicht ident. Modelle?	nein	nein	nein	nein	nein
Bootstrap-Konfidenzinterv.?	nein	nein	nein	nein	ja**
Bootstrap-Modelltest?	nein	nein	nein	nein	nein
graphische Modelleingabe?	nein	nein	ja	nein	nein
graphische Modellausgabe?	nein	ja	ja	nein	nein
Benutzerfreundlicheit:					
erf. Programmierkenntnisse?	unnötig	unnötig	unnötig	PASCAL	unnötig
Lernbarkeit?	leicht	leicht	aufwendig	aufwendig	leicht
Dokumentation?	mäßig	gut	mäßig	mäßig	gut
Komplikationsrisiko?	gering	gering	mäßig	höher	mäßig
Schnelligkeit der Benutzung?	schnell	schnell	mäßig	aufwendig	mäßig
Max. Modellgröße?	mäßig	sehr gut	sehr gut	mäßig	gut

Anmerkungen:
*: Bei PASCAL-Kenntnissen prinzipiell leicht auf andere Betriebssysteme adaptierbar;
**: Nur bei Benutzung der von Dodson et al. (1998) bereitgestellten EXCEL-Makros.

Einen Ausweg aus dieser Situation ermöglichen Programme, die eine Gradientenmethode der Funktionenminimierung in benutzerfreundlicher Weise implementieren (Brent, 1973; Press et al., 1992). Gegenwärtig scheinen die bereits erwähnten Programme PRAXIS (Gegenfurtner, 1992) und vor allem die SOLVER-Funktion aus EXCEL (Dodson et al., 1998) besonders empfehlenswert zu sein. Während PRAXIS den von Brent (1973, S. 116 - 167) optimierten Gradientenalgorithmus der Funktionenminimierung verwendet, basiert die SOLVER-Funktion von EXCEL auf dem Generalized Reduced Gradient Code GRG2 von Lasdon, Waren, Jain und Ratner (1978). Beide Verfahren finden (mindestens) lokale Minima relativ narrensicher, wenn die zu minimierende Zielfunktion zweimal nach den Modellparametern differenzierbar ist.[56] Das ist u.a. für PD^{λ}-Funktionen zu MVB-Modellen mit Parametern im offenen Intervall (0, 1) der Fall, aber natürlich auch bei vielen anderen parametrisierten multinomialen Modellen. Bei beiden Programmen müssen Benutzer neben den empirischen Kategoriehäufigkeiten $\boldsymbol{y}$ nur die entsprechende PD^{λ}-Funktionen der Daten genau spezifizieren. Als Ergebnis erhält man dann Minimum-PD^{λ}-Schätzer der Parameter und den Wert der zugehörigen Anpassungsstatistik. Technische Details der Programmbenutzung sind bei Gegenfurtner (1992) und Dodson et al. (1998) beschrieben. Ein Nachteil beider Programme ist, daß auf viele nützliche Sonderfunktionen der Programme MBT, GPT und AppleTree verzichtet werden muß (s. Tabelle 4.4). Partiell kann dies allerdings kompensiert werden, wenn das EXCEL-Programm benutzt wird. Dodson et al. (1998) bieten EXCEL-Macros an, mit denen beispielsweise Restarts des Gradientenverfahrens für Zufallsstartwerte (zwecks Identifizierbarkeitsprüfung oder Absicherung eines globalen Maximums der Likelihood-Funktion) sowie Konfidenzintervallbestimmungen mittels Bootstrap möglich werden.

4.8 Zusammenfassung

In diesem Kapitel wurden Probleme der statistischen Analyse parametrisierter multinomialer Modelle unter den Standardannahmen (unabhängige Ziehungen aus der gleichen Grundgesamtheit) behandelt. Besonderes Gewicht wurde dabei auf Modelle gelegt, die zur Familie der multinomialen Verarbeitungsbaummodelle (MVB-Modelle) gehören. Die Abschnitte 4.1 und 4.2 hatten die Bestimmung von Parameterwerten unter Voraussetzung der Modellgültigkeit zum Gegenstand. In Abschnitt 4.1 ging es zunächst um das Identifizierbarkeitsproblem, d.h. um die Frage, ob sich aus bekannten Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ eindeutig der Vektor $\boldsymbol{\theta}$ latenter Parameter ableiten läßt, der die betreffenden Kategoriewahrscheinlichkeiten „erzeugt” hat. Ein stärkerer und ein schwächerer Identifizierbarkeitsbe-

[56] Informationen zu den SOLVER-Algorithmen können dem World Wide Web Dokument http://www.frontsys/ der Firma Frontline Systems Inc. entnommen werden.

griff — globale vs. lokale Identifizierbarkeit — wurden unterschieden, und die Bedeutung des Ranges der Jacobischen Matrix für Entscheidungen über die lokale Identifizierbarkeit wurde herausgearbeitet.

Abschnitt 4.2 behandelte den in der Anwendungspraxis typischen Fall, daß die Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ selbst nicht bekannt sind; vielmehr stehen lediglich Stichprobenhäufigkeiten $\boldsymbol{y}$ für die Beobachtungskategorien zur Verfügung, aus denen möglichst gute Schätzwerte $\hat{\boldsymbol{\theta}}$ der Modellparameter $\boldsymbol{\theta}$ abgeleitet werden müssen. Verschiedenen Kriterien zur Gewinnung von guten Schätzern wurden unterschieden. Empfohlen wurde, ausschließlich von BAN-Schätzern (besten asymptotisch normalverteilten Schätzern) Gebrauch zu machen, wobei innerhalb dieser Gruppe wieder viel für Maximum-Likelihood (ML) Schätzer spricht. Ein allgemeines Verfahren der Gewinnung von BAN-Schätzwerten für MVB-Modelle — der EM^{λ}-Algorithmus von Hu und Batchelder (1994) und Hu (1999a) — wurde vorgestellt. Er ist recht einfach in Computerprogramme implementierbar und funktioniert bei geeigneter Wahl des Schrittweitenparameters für alle BAN-Schätzer der Minimum-PD^{λ}-Familie bis auf $\lambda = -1$. Bei Wahl von $\lambda = 0$ — was der Wahl von ML-Schätzern entspricht — konvergiert der EM-Algorithmus auch ohne Änderung des Schrittweitenparameters immer auf ein Maximum der Likelihood-Funktion.

Parameterwerte, Parameterschätzungen und Konfidenzintervalle bzw. -bereiche sind nur dann sinnvoll interpretierbar, wenn das Modell, in dem die betreffenden Parameter definiert sind, empirisch gilt. In Abschnitt 4.3 wurde dargestellt, wie über die Nullhypothese der Modellgültigkeit anhand von Stichprobendaten entschieden werden kann. Asymptotische χ^2-Standardverfahren wurden ebenso vorgestellt wie Nonstandardverfahren (basierend auf Korrekturformeln und auf dem parametrischen Bootstrap). Nonstandardverfahren sind indiziert, wenn entweder der Stichprobenumfang für asymptotische Verfahren zu klein ist oder wenn einige Modellparameter Werte auf dem Rand des Parameterraumes annehmen, so daß die Regularitätsbedingungen von Birch (1964) verletzt sind. Darüber hinaus wurde dargestellt, wie Parameterhypothesen und spezielle Submodellhypothesen (z.B. Rangordnungshypothesen, stochastische Unabhängigkeitshypothesen) im Rahmen eines empirisch gültigen Rahmenmodells formuliert und durch ΔG^2-Differenzentests optimal getestet werden können. Speziell Rangordnungshypothesen zu den Parametern von MVB-Modellen kommen in der Anwendungspraxis recht häufig vor (vgl. etwa Klauer & Wegener, 1998, Experiment 6, um nur ein konkretes Beispiel zu nennen).

Besonders ausführlich wurde auf die Kontrolle des β-Fehlers bei Modellgeltungstests eingegangen (Abschnitt 4.3.3). Dargestellt wurden klassische und neuere Approximationsmethoden für die nonzentralen Verteilungen der Prüfstatistiken unter H_1. Einige dieser Methoden wurden anhand einer Monte-Carlo-Studie zum

Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986) miteinander verglichen. Als Ergebnis dieser Studie kann festgehalten werden, daß nonzentrale χ^2-Verteilungen bei nicht zu kleinen Stichprobenumfängen durchaus akzeptable Approximationen der exakten Verteilungen ermöglichen. Der Nonzentralitätsparameter ist hierbei idealerweise als Minimum-PD^{λ}-Distanz zwischen den erwarteten Kategoriehäufigkeiten unter H_0 und H_1 zu spezifizieren. Die konkrete Berechnung von Teststärkewerten für bestimmte Alternativhypothesen wurde dargestellt. Außerdem wurde auf verschiedene versuchsplanerische Möglichkeiten der Teststärkeoptimierung hingewiesen, insbesondere auf die Rolle der idealen Aufteilung des Gesamtstichprobenumfangs auf einzelne Teilbäume eines verbundenen MVB-Modells.

Im Anschluß an eine Darstellung von informationstheoretischen Maßen der Anpassungsgüte (AIC und BIC), die speziell beim Vergleich von Modellen aus einer nichthierarchischen Modellfamilie nützlich sein können, wurde Probleme angesprochen, die aus der Äquivalenz, Datenäquivalenz und der „Beinahe-Datenäquivalenz" von Modellen resultieren (Abschnitt 4.4). Lösen kann man sie nur im Rahmen umfassender Validierungsstudien, welche prüfen, ob sich experimentelle Manipulationen hypothesengemäß auf bestimmte Modellparameter auswirken, auf andere Modellparameter dagegen nicht. In Abschnitt 4.5 wurde dargestellt, daß sich auf diese Weise eine Entscheidung zwischen datenäquivalenten Modellen herbeiführen läßt, zwischen äquivalenten Modellen dagegen im allgemeinen nicht.

Abschnitt 4.6 schließlich hatte mögliche Lösungsansätze bei Identifizierbarkeitsproblemen zum Gegenstand. Interessiert nur die Frage nach der Modellgültigkeit und sind die Werte nicht identifizierbarer Parameter nicht oder nur sekundär von Interesse, so kann man u.U. auch nicht identifizierbare Modelle statistisch testen, sofern diese das Kriterium der quantitativen Testbarkeit von Bamber und van Santen (1985) erfüllen und eine Korrektur der Freiheitsgrade durchgeführt wird. Sind dagegen auch Parameter (oder Funktionen von Parametern) nicht identifizierbar, deren Werte psychologisch wichtig sind, so muß man versuchen, Identifizierbarkeit entweder (a) durch inhaltlich vertretbare Restriktionen auf nichtinteressierende Parameter oder (b) durch eine Erweiterungen des empirischen Paradigmas herzustellen. Geht man den erstgenannten Weg, so sollte niemals auf eine kritische Evaluation des möglichen Einflusses der Parameterrestriktion auf das resultierende Ergebnismuster verzichtet werden. Zieht man eine Paradigmenerweiterung vor, so gibt es im allgemeinen mehrere denkbare Modellerweiterungen, zwischen denen anhand von Validierungsexperimenten zu entscheiden ist.

Software zur statistischen Analyse von parametrisierten multinomialen Modellen— speziell von MVB-Modellen — wurde in Abschnitt 4.7 vorgestellt. Gegenwärtig kann kein Programm als „optimales Allzweckwerkzeug" empfohlen

werden. Je nach Fragestellung und Ziel der Analyse muß auf unterschiedliche Programme zurückgegriffen werden (vgl. Tabelle 4.4).

5. Die Robustheit multinomialer Modellierung

Die im letzten Kapitel geschilderten statistischen Analysen basieren wesentlich auf den Annahmen, die dem jeweiligen multinomialen Verteilungsmodell zugrunde liegen. Nominelle BAN-Schätzer der Modellparameter müssen nicht faktische BAN-Schätzer und nominell χ^2-verteilte Prüfstatistiken nicht faktisch χ^2-verteilt sein, wenn diese Annahmen verletzt sind. Praktisch heißt das, daß die statistische Rechtfertigung der geschilderten Analyseverfahren — vorsichtig ausgedrückt — problematisch ist, wenn nicht *allen* Beobachtungen der empirischen Stichprobe die *gleichen* Kategoriewahrscheinlichkeiten $\boldsymbol{\pi}$ zugrunde liegen und wenn die Beobachtungen nicht innerhalb und zwischen den Gruppen eines verbundenen multinomialen Modells *vollständig stochastisch unabhängig* sind. Dies gilt für die asymptotischen Standardverfahren, aber — wie wir gesehen haben — auch für Nonstandardverfahren wie z.B. den nonparametrischen oder den parametrischen Bootstrap.

Die Annahme exakt homogener Kategoriewahrscheinlichkeiten und komplett unabhängiger Beobachtungen ist für Anwendungen multinomialer Modelle in der kognitiven Psychologie nicht sehr realistisch (vgl. z.B. Wickens, 1998a, S. 544f). Typischerweise entstehen die letztlich analysierten Häufigkeitsverteilungen durch Aggregieren über Vpn und Items hinweg: Insgesamt V Individuen bearbeiten die gleichen I Items, z.B. alte und neue Items in einem Rekognitionstest. Die resultierenden $V{\cdot}I$ Beobachtungsresultate entsprechen dann dem Gesamtstichprobenumfang N, der in die Datenanalysen eingeht. Selbst wenn man sich hierbei bemüht, möglichst homogene Vpn-Stichproben und Itempools zugrundezulegen, wird doch eine gewisse (möglicherweise geringe) Restvariabilität zu verzeichnen sein. Folglich sind die Kategoriewahrscheinlichkeiten, die den N Stichprobendaten zugrunde liegen, keine Konstanten, sondern vielmehr selbst Zufallsvariablen, die für unterschiedliche Vp-Item-Kombinationen verschiedene Werte annehmen können. Außerdem ist zu erwarten, daß die Beobachtungsresultate zumindest schwach korreliert sind: Wenn eine Vp z.B. ein bestimmtes Rekognitionstestitem bearbeitet, muß man davon ausgehen, daß die Wahrscheinlichkeit einer korrekten Antwort nicht nur von der über Items und Vpn gemittelten Lösungswahrscheinlichkeit in der jeweiligen Versuchsbedingung abhängt, sondern zusätzlich auch davon, wie viele der anderen Vpn das betreffende Item korrekt beantwortet haben. Dies ist nämlich ein Indiz für die relative Leichtigkeit des Items. Unterscheiden sich die Items hinsichtlich ihrer Leichtigkeit, so sollten die Resultate verschiedener Vpn über die Items hinweg korreliert sein. Unterscheiden sich umgekehrt die Personen hinsichtlich ihrer Fähigkeit, so sollten die Antworten auf verschiedene Items über die Vpn hinweg korreliert sein.

Eine Frage, die sich Anbetracht derartiger Diskrepanzen zwischen statistischer Theorie und experimentalpsychologischer Praxis zwangsläufig stellt, ist die nach der Robustheit multinomialer Datenanalysen: Behalten die Ergebnisse statistischer Analysen parametrisierter multinomialer Modelle — insbesondere die Aussagen zu Punktschätzungen, Konfidenzintervallen und Überschreitungswahrscheinlichkeiten der Teststatistiken — wenigstens annähernd ihre Gültigkeit, wenn die Verteilungsannahmen multinomialer Modelle mehr oder minder stark verletzt sind? Die Hoffnung auf Robustheit ist durchaus berechtigt, weil die Gültigkeit eines angenommenen Verteilungsmodells zwar eine hinreichende, jedoch keinesfalls zugleich auch eine notwendige Bedingung für die Konsistenz eines Schätzers oder eines statistischen Tests ist. Beispielsweise können „Quasi-Maximum-Likelihood-Schätzer" unter bestimmten Bedingungen dennoch konsistent und auch asymptotisch normalverteilt sein (Golden, 1995; White, 1982). (Unter „Quasi-Maximum-Likelihood-Schätzern" versteht man Schätzer, die aus der Anwendung von ML-Schätzformeln für ein nicht zutreffendes Verteilungsmodell resultieren.) Auch die darauf aufbauenden Inferenzverfahren testen bei Verletzungen des Verteilungsmodells u.U. immer noch die korrekten Nullhypothesen, wenn auch möglicherweise nicht so effizient wie im Falle der Gültigkeit des angenommenen Verteilungsmodells (Golden, 1995; White, 1982). Hinweise auf mögliche Verletzungen eines Verteilungsmodells machen also statistische Ergebnisse, die z.B. mit Maximum-Likelihood-Verfahren gewonnen wurden, nicht automatisch wertlos.

Prinzipiell kann man zwei Ansätze verfolgen, um „Quasi-Maximum-Likelihood-Schätzer", „Quasi-Minimum-PD$^{\lambda}$-Schätzer" oder darauf aufbauende Inferenzverfahren statistisch abzusichern:

(1) *Analytischer Ansatz*: Mathematische Untersuchungen zur Frage, auf welche Verletzungen der Modellannahmen die Standardverfahren robust bzw. nicht robust reagieren; eventuell Entwicklung statistischer Tests, die prüfen, ob eine robustheitsgefährdende Verletzung der Verteilungsannahmen im konkreten Anwendungsfall vorliegt oder nicht (White, 1982). Auch derartige Verfahren sind allerdings wieder asymptotische Verfahren und machen bestimmte (schwächere) Verteilungsannahmen, z.B. die Annahme unabhängig und identisch verteilter Beobachtungen auf demselben Stichprobenraum. Falls gerade diese zugrunde gelegten Annahmen verletzt sind oder der vorliegende Stichprobenumfang klein ist, ist der analytische Ansatz nur von begrenztem Wert (Golden, 1995, S. 4).

(2) *Pragmatischer Ansatz*: Untersuchungen zum Verhalten der Standard-Schätzer und Standard-Inferenzverfahren bei verschiedenen Graden der Modellverletzung, beispielsweise mit Hilfe von Monte-Carlo-Studien.

Der analytische Ansatz scheint im Zusammenhang mit parametrisierten multinomialen Modellen noch nicht sehr gut ausgearbeitet zu sein. Zum pragmatischen Ansatz liegen jedoch schon einige Resultate vor. Im folgenden Abschnitt 5.1 werden diese Resultate zunächst kurz zusammengefaßt. Abschnitt 5.2 präsentiert eine Kritik bisheriger Robustheitsstudien, die darauf abzielt, zwischen Verletzungen von Verteilungsannahmen mit und ohne Konsequenzen für die Modellgeltungshypothesen zu unterscheiden. Es wird sich dabei herausstellen, daß die Modellgeltungshypothese für die aggregierten Daten nicht zwangsläufig aus der Gültigkeit eines Modells für jede Vp-Item-Kombination folgt, sondern nur dann, wenn die Modellparameter innerhalb eines Astes des Baummodells über Vp-Item-Kombinationen hinweg nicht korreliert sind. In Abschnitt 5.3 folgt dann eine weitere Monte-Carlo-Studie zum Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986), dieses Mal mit dem Ziel, die Robustheit multinomialer Analysen bei unkorrelierten und korrelierten Modellparametern gegeneinander zu kontrastieren. Im erstgenannten Fall werden sich Modellgeltungstests als tendenziell konservativ, im letzten Fall als antikonservativ erweisen. Abschnitt 5.4 zeigt potentielle Alternativen auf, die Verwendung finden können, wenn Standardanalysen multinomialer Modelle nicht mehr vertretbar erscheinen. Die wichtigsten Resultate dieses Kapitels werden dann in Abschnitt 5.5 noch einmal zusammengefaßt.

5.1 Bisherige Studien zur Robustheit von MVB-Modellanalysen

Batchelder und Riefer (1986, Abschnitt 2.4) haben untersucht, wie stark die Ergebnisse verschiedener Parameterschätzmethoden für das Enkodierung-Abruf-Modell differieren können, wenn (a) das Drei-Parameter-Modell für jede von V untersuchten Vp gilt, (b) die Modellparameter c_v, r_v und u_v für die verschiedenen Vpn $v = 1, ..., V$ unterschiedliche Werte annehmen und (c) die Schätzung der durchschnittlichen Modellparameter

$$\bar{c} = (1/V)\cdot\sum_{v=1}^{V} c_v \;,\; \bar{r} = (1/V)\cdot\sum_{v=1}^{V} r_v \;,\; \bar{u} = (1/V)\cdot\sum_{v=1}^{V} u_v \qquad (5.1\text{-}1)$$

angestrebt wird. Batchelder und Riefer (1986, S. 136) zeigten am Beispiel des mittleren c-Parameters, daß der ML-Schätzer $\hat{c}$ für die über Vpn hinweg aggregierten Daten — im folgenden „Aggregatschätzer" genannt — im allgemeinen nicht identisch ist mit dem Schätzer, der sich durch Mittelwertsbildung über individuelle ML-Schätzwerte $\hat{c}_v$ für jede einzelne Vpn ergibt. Generell gilt, daß die Diskrepanz der Ergebnisse beider Schätzmethoden um so größer wird, je mehr sich die relativen Häufigkeiten in den Beobachtungskategorien $C_{1,2}$, $C_{1,3}$ und $C_{2,1}$ zwischen den Vpn unterscheiden. Interessanterweise erweist sich interindividuelle Variabilität der relativen Häufigkeiten für Kategorie $C_{1,1}$ als unschädlich; es ist also keineswegs

so, daß interindividuelle Unterschiede immer Diskrepanzen zwischen Aggregatschätzern und den gemittelten individuellen Schätzern erzeugen müssen. Ich werde im nächsten Abschnitt genauer untersuchen, welche Modelleigenschaften für die Unschädlichkeit interindividueller Variabilität maßgeblich sind.

Wie ist mit dem Problem möglicher Diskrepanzen zwischen Aggregatschätzern und gemittelten individuellen Schätzern umzugehen? Im Einklang mit der vorherrschenden Forschungspraxis haben Batchelder und Riefer (1986) sich zugunsten von Aggregatschätzern ausgesprochen, weil individuelle ML-Schätzer ihrer Auffassung nach mit gravierenderen Problemen behaftet sind.[57] Zunächst einmal weisen sie wegen des typischerweise sehr kleinen Stichprobenumfanges einen starken mittleren Schätzfehler und exzessiv große Varianz auf. Außerdem muß damit gerechnet werden, daß die Adäquatheitsbedingungen für BAN-Schätzer nach Birch (1964) — insbesondere die Forderung, daß alle Kategoriewahrscheinlichkeiten größer als null sein müssen — bei der Analyse individueller Daten zumindest für einige Vpn verletzt sind. Inwieweit der hierdurch erzeugte mittlere Schätzfehler wirklich größer als bei Aggregatschätzern ist, wurde von Batchelder und Riefer (1986) nicht näher untersucht. Klar ist jedoch, daß die Vorbehalte gegen die Mittelung individueller ML-Schätzwerte mit wachsendem Stichprobenumfang N_V pro Vp zunehmend gegenstandslos werden. Aggregatschätzer müssen also ab einem bestimmten Stichprobenumfang N_V schlechter (d.h. verzerrter) als gemittelte individuelle Schätzer werden, es sei denn, die Variabilität ist von einer Art, die keine Diskrepanzen zwischen beiden Schätzmethoden erzeugt. Ab welchem Stichprobenumfang Aggregatschätzer suboptimal werden, bleibt jedoch offen. Hautus (1997) hat das gleiche Problem für den d' Parameter der Signalentdeckungstheorie mittels Monte-Carlo-Methoden untersucht und schon ab N_V = 16 bis N_V = 64 (abhängig von den Modellparametern) einen Vorteil gemittelter individueller Parameterschätzungen („average d'") gegenüber Aggregatschätzern („pooled d'") gefunden. Es ist nicht unwahrscheinlich, daß sich ein ähnliches Ergebnis auch für die Parameter des Enkodierung-Abruf-Modells ergeben würde.

Die Frage, wie sich Parametervariabilität auf den mittleren Schätzfehler (bias) und die Konfidenzintervalle von ML-Aggregatschätzern auswirkt, wurde für MVB-Modelle erstmals von Riefer und Batchelder (1988) untersucht. Sie führten eine Monte-Carlo-Studie zu einem Modell der proaktiven Hemmung von Greeno, James, DaPolito und Polson (1978) durch. Das Modell bezieht sich auf Paar-Assoziations-Lernexperimente, in denen Vpn der Experimentalbedingung unter anderem A-B-Itempaare und später in der gleichen Liste A-C-Itempaare lernen. Im anschließenden geförderten Reproduktionstest müssen die Vpn auf den Hinweisreiz

[57] Für signalentdeckungstheoretische Analysen haben Macmillan und Kaplan (1985) einen ähnlichen Standpunkt vertreten.

A möglichst beide Zielitems reproduzieren, die mit A in der Lernliste assoziiert waren. Proaktive Hemmung ist demonstriert, wenn die Reproduktion des später gelernten C-Zielitems in der Experimentalbedingung schlechter als in einer Kontrollbedingung ohne vorgeschaltetes A-B-Paar gelingt. Das MVB-Modell von Greeno et al. (1978) stellt die Wahrscheinlichkeiten der vier Ereignisse, die aus der Kreuzklassifikation der B-Reproduktion (erfolgreich / nicht erfolgreich) mit der C-Reproduktion (erfolgreich / nicht erfolgreich) resultieren, als Funktion dreier latenter Wahrscheinlichkeiten dar:

— p (Zugriff auf die Gedächtnisrepräsentation des Hinweisitems A),
— q (Abruf von B, gegeben Zugriff auf die A-Gedächtnisrepräsentation) und
— r (Abruf von C, gegeben Zugriff auf die A-Gedächtnisrepräsentation).

Die Modellgleichungen ergeben sich in kanonischer Weise aus den Annahmen, daß (a) weder Item B noch Item C reproduziert wird, wenn der Zugriff auf die A-Gedächtnisrepräsentation scheitert und (b) der Abruf von B und C bei Zugriff auf die A-Gedächtnisrepräsentation stochastisch unabhängig erfolgt. Die Wahrscheinlichkeit r ist also für Fälle mit und ohne erfolgreiche B-Reproduktion gleich. Eine Ereignisbaumrepräsentation des Modells findet man bei Riefer und Batchelder (1988, S. 327, Figure 1).

Riefer und Batchelder (1988, S. 328) generierten Monte-Carlo-Stichproben zu diesem Modell unter der Annahme, daß (a) das Modell für alle Vpn gilt, (b) grundsätzlich $p = 1$ zutrifft und (c) q_v und r_v Realisationen von Zufallsvariablen sind, die zwischen den Vpn $v = 1, ..., V$ variieren. Als Dichteverteilungen wurden zwei unabhängige Beta-Verteilungen gewählt, da dieser Verteilungstyp psychologisch plausible und statistisch einfach handhabbare Dichten im Intervall [0, 1] erzeugt (Johnson & Kotz, 1970b, Kapitel 24; Wickens, 1982, S. 277). Eine beta-verteilte ZV Θ mit Realisationen θ ($0 \leq \theta \leq 1$) hat die Dichtefunktion

$$f(\theta; p, q) = \frac{1}{\int_{t=0}^{1} t^{p-1}(1-t)^{q-1}\,dt} \cdot \theta^{(p-1)} \cdot (1-\theta)^{(q-1)} \qquad (5.1\text{-}2)$$

und liefert in Abhängigkeit von den Parametern p und q ($p > 0$, $q > 0$) normalverteilungsähnliche symmetrische Verteilungen ($p > 1$, $q > 1$, $p = q$) unimodale schiefe Verteilungen ($p > 1$, $q > 1$, $p \neq q$), Gleichverteilungen ($p = q = 1$) oder auch bimodale Verteilungen im Intervall [0, 1] ($p < 1$, $q < 1$). Riefer und Batchelder (1988) haben sich in ihrer Simulationsstudie auf den psychologisch plausibelsten Fall unimodaler Verteilungen beschränkt. Für einen Gesamtstichprobenumfang von $N = 30$, wie er von einer Vp durchaus erzeugt werden kann (bei 30 A-B- und A-C-Paaren in der Lernliste), resultierte erwartungsgemäß ein starker Bias in den Aggregatschätzern und außerdem eine starke Unterschätzung der faktischen Konfidenzintervallbreite, sofern eine große Streuung der Beta-Verteilungen (d.h. $\sigma \geq .30$) an-

genommen wird. Ab N = 120 ist der Fehler jedoch schon kleiner als 5%, selbst wenn eine derart große interindividuelle Streuung der Parameterwerte vorliegt. Ist die Streuung dagegen klein (d.h. $\sigma \leq .15$), dann reicht auch schon ein Gesamtstichprobenumfang von N = 50 aus, um akzeptable Eigenschaften der Aggregatschätzer sicherzustellen. Riefer und Batchelder (1988) schlußfolgern, daß Parameterschätzer und Konfidenzintervalle für das Modell von Greeno et al. (1978) als „sicher robust" gegenüber Parametervariabilität betrachtet werden können, wenn der Gesamtstichprobenumfang N = 180 erreicht oder überschritten wird. Dies dürfte bei kognitionspsychologischen Anwendungen fast immer der Fall sein.

Läßt sich dieses erfreuliche Ergebnis problemlos auf andere MVB-Modelle übertragen? Das Modell von Greeno et al. (1978) weist im Vergleich zum Enkodierung-Abruf-Modell eine Besonderheit auf, die von Riefer und Batchelder (1988) nicht explizit diskutiert wurde: Aggregatschätzer und gemittelte individuelle Schätzer führen beim Modell von Greeno et al. (1978) zu identischen Ergebnissen, wenn allen Vpn gleich viele A-Items vorgelegt werden.[58] Sei I die Anzahl der A-Items. Unter der Randbedingung p = 1 ist der individuelle ML-Schätzer für q_v einfach der Anteil reproduzierter B-Items für Vp v, d.h. N_{Bv} / I. Der entsprechende Schätzer für r_v ist der Anteil reproduzierter C-Items, d.h. N_{Cv} / I. Für die Aggregatschätzer gilt:

$$\hat{q} = \frac{\sum_{v=1}^{V} N_{B_v}}{V \cdot I} = (1 / V) \cdot \sum_{v=1}^{V} \frac{N_{B_v}}{I} = (1 / V) \cdot \sum_{v=1}^{V} \hat{q}_v$$
$$\hat{r} = \frac{\sum_{v=1}^{V} N_{C_v}}{V \cdot I} = (1 / V) \cdot \sum_{v=1}^{V} \frac{N_{C_v}}{I} = (1 / V) \cdot \sum_{v=1}^{V} \hat{r}_v \qquad (5.1\text{-}3)$$

Die Ausdrücke ganz rechts in den beiden Gleichungen entsprechen gerade den Mittelwerten der individuellen ML-Schätzer, womit deren Äquivalenz zu den Aggregatschätzern gezeigt wäre. Denkbar ist, daß die günstigen Robustheitseigenschaften des Modells von Greeno et al. mit dieser Äquivalenz zusammenhängen und nicht notwendig auf andere Modelle übertragbar sind. Ich komme im nächsten Abschnitt hierauf zurück.

Die erste Arbeit, in der neben der Robustheit von Punkt- und Bereichsschätzern auch die Robustheit des G^2-Modellanpassungstests bei interindividuell unterschiedlichen Parameterwerten untersucht wurde, stammt von Riefer und Batchelder (1991b). Sie führten u.a. Monte-Carlo-Studien zum Enkodierung-Abruf-Modell (EA-Modell) mit beta-verteilten Parametern durch. Zwei Streuungen (σ = .1 und σ = .2) wurden jeweils für alle drei Modellparameter realisiert, abgesehen von dem

[58] Wenn sich die Stichprobenumfänge zwischen den Vpn unterscheiden, entspricht der Aggregatschätzer einem gewichteten Mittel der individuellen ML-Schätzer.

Fall konstanter Parameter, der in Abschnitt 4.2.4 schon besprochen wurde. In bezug auf die Parameterschätzungen zeigte sich der zu erwartende Anstieg des mittleren Schätzfehlers mit zunehmender Parametervariabilität, der aber nur bei einer Streuung von σ = .20 beachtlich ausfiel. Allerdings verschwand der Bias in den ML-Aggregatschätzern von $\bar{c}$ und $\bar{r}$ bei dieser Streuung bis hin zum größten untersuchten Stichprobenumfang *nicht* (N_1 = 150 Itempaare plus N_2 = 75 Einzelitems), sofern als mittlerer Parameter $\bar{c}$ < .30 gewählt wurde. Teilweise wuchs der Bias sogar noch mit dem Stichprobenumfang an (Riefer & Batchelder, 1991b, S. 322, Abbildung 2). Dieses Ergebnis ist in Anbetracht der positiven Robustheitsbefunde zum Modell von Greeno et al. (1978) sicher überraschend und unerfreulich. Weniger Grund zur Sorge ist dagegen in bezug auf die Schätzung der Konfidenzintervallbreite gegeben. Hier wurden natürlich ebenfalls Diskrepanzen zwischen exakten und asymptotischen Konfidenzintervallen bei großer Parametervariabilität beobachtet. Diese Diskrepanzen verschwanden jedoch mit zunehmendem Stichprobenumfang und waren bei N_1 = 150 und N_2 = 75 bereits vernachlässigbar.

Der für die aggregierten Daten durchgeführte G^2-Modellanpassungstest zum Drei-Parameter-EA-Modell (df = 1) hält bei kleinen Streuungen der Modellparameter (σ < .10) bis zu einem Gesamtstichprobenumfang von etwa N_1 = 1000 Itempaaren und N_2 = 500 Einzelitems das α-Niveau halbwegs ein, d.h. bei nominellem α = .05 ist das faktische α < .075. Ist die Parameterstreuung dagegen .20 oder noch größer, so führt dies zu einer exzessiven Erhöhung des faktischen α-Fehlers relativ zum nominellen α-Fehler (Riefer & Batchelder, 1991b, S. 331, Abbildung 9). Der Test ist in diesem Fall statistisch offensichtlich nicht mehr vertretbar, weil eine systematische Voreingenommenheit gegen die Nullhypothese vorliegt. Bei Modellanpassungstests hat dies zu viele fälschliche Modellverwerfungen zur Folge, bei Tests auf Parameterunterschiede (z.B. zwischen experimentellen Gruppen) zu viele fälschliche Annahmen einer Parameterunterschiedshypothese.

Rouder und Batchelder (1998, S. 211ff) haben zu einem Speicher-Abruf-Modell das Paar-Assoziations-Lernens (vgl. Abschnitt 7.1.4) ebenfalls eine Monte-Carlo-Studie durchgeführt, wobei allerdings an Stelle der Beta-Verteilung jeweils eine Gleichverteilung der Parameterwerte in einem bestimmten Intervall zugrunde gelegt wurde. Die Ergebnisse entsprachen im wesentlichen denen von Riefer und Batchelder (1991b): Erst ab N = 1000 ist ein substantielle Erhöhung des α-Fehlers bei moderater Parametervariabilität festzustellen.

Als Fazit kann festgehalten werden, daß die bislang vorliegenden Untersuchungen zum Effekt von Parameterheterogenität auf die Ergebnisse statistischer Analysen von MVB-Modellen keine Hinweise auf bedeutsame Ergebnisverzerrungen liefern, sofern die Streuungen der Parameterwerte σ = .10 oder weniger betragen. Die statistischen Standardanalysen wären demnach bei vergleichsweise homogenen Vpn- und Itemstichproben durchaus noch vertretbar. Diese Interpre-

tation wird auch durch weitere Monte-Carlo-Studien zu anderen MVB-Modellen gestützt, deren Ergebnisse nicht im Detail dokumentiert wurden (vgl. z.B. Riefer et al., 1994). Schon ab Streuungen in der Größenordnung $\sigma \geq .20$ ergeben sich jedoch durchaus heterogene Befunde: Für einige Modelle scheinen Schätzer und Anpassungsstatistiken auch dann noch recht robust zu sein (z.B. Riefer & Batchelder, 1988, S. 328), während es bei anderen Modellen — etwa beim EA-Modell — zu irreführenden statistischen Resultaten kommen kann (vgl. Riefer & Batchelder, 1991b, S. 331). Mögliche Ursachen hierfür sollen im folgenden Abschnitt näher untersucht werden.

5.2 Robustheit bei korrelierten und unkorrelierten Parametern

Ein auf den ersten Blick überraschender Aspekt der Monte-Carlo-Befunde von Riefer und Batchelder (1991b, S. 331) ist der Anstieg des faktischen α-Fehlers mit zunehmender Parametervariabilität. Parametervariabilität führt nämlich zunächst einmal dazu, daß die faktischen Konfidenzintervalle größer ausfallen als die nominellen Konfidenzintervalle, die unter der Annahme konstanter Parameter geschätzt werden. Dies ist eine einfache Konsequenz der Tatsache, daß die Standardschätzfehler bei konstanten Parametern nur durch Stichprobenvariabilität beeinflußt werden, bei heterogenen Parametern dagegen zusätzlich noch durch die Parametervariabilität. Entsprechende Diskrepanzen zwischen faktischen und nominellen Konfidenzintervallen haben Riefer und Batchelder (1988, S. 327) in ihrer Monte-Carlo-Studie zum Modell von Greeno et al. (1978) auch beobachtet.

Der Modellanpassungstest zum Drei-Parameter-EA-Modell prüft die Gleichheit des *u*- und des *a*-Parameters im saturierten Vier-Parameter-EA-Modell. Für einen bestimmten kritischen Wert der Teststatistik sollte die H_0 der Modellgültigkeit um so seltener verworfen werden, je breiter die Konfidenzbereiche für die Schätzer von *u* and *a* ausfallen. Da Parametervariabilität den Standardschätzfehler erhöht, wäre also eigentlich keine Erhöhung, sondern eher eine Verringerung des faktischen α-Fehlers mit zunehmender Parameterstreuung zu erwarten.

Wie ist das dazu im Widerspruch stehende Ergebnis der Monte-Carlo-Studie von Riefer und Batchelder (1991b) zu erklären? Die Antwort auf diese Frage verlangt, daß zwei Aspekte der Robustheitsheitsproblematik, die bislang nicht voneinander unterschieden wurden, separat diskutiert werden: die *Aggregierungsinvarianz der Modellgeltungshypothese* bei Parametervariabilität einerseits und die *Robustheit des statistischen Tests* der Aggregathypothese gegenüber Parametervariabilität andererseits. Der erstgenannte Aspekt betrifft das altbekannte Problem der „Aggregierungsvalidität" statistischer Hypothesen (aggregation validity, vgl. Siemer, 1997). Schon in den fünfziger Jahren fand in der Psychologie eine intensive Diskussion über das Problem der „mittleren Kurve" statt, vor allem im Zusam-

menhang mit quantitativen Lerngesetzen, die eine bestimmte Form der Zunahme von richtigen Antworten (oder der Abnahme von falschen Antworten) über verschiedene Lerndurchgänge hinweg behaupten (z.B. Bakan, 1954; Estes, 1956; Sidman, 1952). Die Lerngesetze sind dabei als allgemeinpsychologische Hypothesen zu verstehen, die zwar für jedes einzelne Individuum Gültigkeit beanspruchen, jedoch interindividuell unterschiedliche Parameter (Lernraten, Lernasymptoten etc.) aufweisen können. Die gängige Praxis war und ist es, nicht die Daten einzelner Vpn zu analysieren, sondern die aggregierten Daten über verschiedene Vpn hinweg. Dann stellt sich aber die Frage, ob ein bestimmtes Lerngesetz aggregierungsinvariant ist, d.h. ob es für die über Vpn gemittelten Daten gelten muß, wenn es erstens für jede einzelne Vp gilt und sich zweitens die Lernparameter zwischen den Individuen unterscheiden. Sidman (1952), Bakan (1954), Estes (1956) und andere Autoren haben Kriterien ausgearbeitet, die eine Entscheidung darüber erlauben, wann bestimmte mathematische Funktionen, die den behaupteten Gesetzmäßigkeiten entsprechen, in diesem Sinne aggregierungsinvariant sind. So sind beispielsweise lineare oder logarithmische Funktionen aggregierungsinvariant, d.h. wenn jede Vp einem linearen (bzw. logarithmischen) Lerngesetz folgt, dann ist auch die mittlere Lernkurve linear (bzw. logarithmisch). Andere Funktionen — z.B. Exponentialfunktionen (Sidman, 1952) oder Gompertz-Funktionen zur Beschreibung von Wachstumsprozessen (vgl. Tanner, 1970) — sind im allgemeinen nicht aggregierungsinvariant. Liegt Aggregierungsinvarianz nicht vor, so ist es offensichtlich nicht sinnvoll, die für individuelle Kurven behauptete Gesetzmäßigkeit anhand interindividuell aggregierter Daten zu überprüfen: Die allgemeinpsychologische Hypothese impliziert ja gerade, daß das entsprechende Gesetz für die aggregierten Daten *nicht* gelten darf, wenn es für jedes einzelne Individuum gilt und die Parameter interindividuell variieren. In diesem Fall bleibt oftmals[59] keine Alternative zur Überprüfung des Gesetzes anhand der Daten einzelner Vpn bzw. anhand der Daten einer Vpn-Gruppe, die als sicher homogen bezüglich der Parameter gelten kann (vgl. Bredenkamp, 1972, Kapitel 4).

Bei der statistischen Prüfung parametrisierter multinomialer Modelle stellt sich genau das gleiche Problem. Multinomiale Modelle sind in kognitionspsychologischen Anwendungen eigentlich immer als allgemeinpsychologische Modelle aufzufassen, die für jedes einzelne Individuum gelten sollen, wenn auch mit mög-

[59]Die Einschränkung „oftmals" ist nötig, weil die korrekte Mittelwertskurve auch bei Funktionen, die nicht aggregierungsinvariant sind, u.U. dennoch hergeleitet werden kann (vgl. Bredenkamp, 1972; Sidman, 1952). In einem solchen Fall kann es sinnvoll sein, die implizierte Mittelwertskurve, die von einem anderen Typ als die individuellen Kurven ist, anhand der aggregierten Daten zu überprüfen. Muß nämlich die Hypothese über mittlere Kurve verworfen werden, so impliziert dies auch die Verwerfung der übergeordneten allgemeinpsychologischen Hypothese (Bredenkamp, 1972, S. 164).

licherweise unterschiedlichen Parametern.[60] Ehe man ein solches Modell anhand interindividuell aggregierter Daten überprüft, ist die Aggregierungsinvarianz der Modellgeltungshypothese zu überprüfen.

Ich möchte dieses Problem anhand der in dieser Arbeit primär interessierenden Modellklasse der binären MVB-Modelle untersuchen. Die Übertragung auf nichtbinäre Modelle bereitet keinerlei Probleme. Um dem Problem interindividuell variabler Parameter gerecht zu werden, muß das einfache und das verbundene multinomiale Stichprobenmodell aus den Abschnitten 3.1 und 3.2 zunächst durch ein zweistufiges Zufallsexperiment pro Stichprobendatum ersetzt werden. In der ersten Stufe wird ein bestimmter Wert $\boldsymbol{\theta}$ des „Parametervektors" aus einer Grundgesamtheit möglicher Werte gezogen. Der „Parametervektor" ist in diesem Kontext also eigentlich gar kein Vektor von Parametern mehr, sondern vielmehr selbst ein (mehrdimensionaler) Zufallsvektor $\boldsymbol{\Theta}$. Ich werde $\boldsymbol{\Theta}$ deshalb im folgenden als Zufallsparametervektor bezeichnen und jeden konkreten Wert $\boldsymbol{\Theta} = \boldsymbol{\theta}$ als Realisation von $\boldsymbol{\Theta}$. In typischen Anwendungen korrespondiert die Ziehung einer Realisation von $\boldsymbol{\Theta}$ der Ziehung einer bestimmten Vp oder einer Vp-Item-Kombination aus einer Grundgesamtheit. In der zweiten Stufe des Zufallsexperiments wird — wie in einem gewöhnlichen multinomialen oder verbundenen multinomialen Zufallsexperiment auch — zu gegebenem $\boldsymbol{\Theta} = \boldsymbol{\theta}$ ein Beobachtungsresultat gemäß Kategoriewahrscheinlichkeitsvektor $\boldsymbol{\pi}(\boldsymbol{\Theta} = \boldsymbol{\theta})$ gewonnen. Von dem Zufallsvektor $\boldsymbol{\Theta}$ muß die Verteilung nicht notwendig bekannt sein. Es sei lediglich unterstellt, daß für alle Zufallsparameter $0 \leq \Theta_s \leq 1$, $s = 1, ..., S$, gilt; ferner sei vorausgesetzt, daß Erwartungswerte $E(\Theta_s)$, Varianzen $Var(\Theta_s)$ und Kovarianzen $Cov(\Theta_s, \Theta_{s'})$ aller S Zufallsparameter existieren.

Das zweistufige Zufallsexperiment wird dann mehrfach unabhängig wiederholt, bis insgesamt N Beobachtungen vorliegen bzw. bei verbundenen Modellen $\boldsymbol{N} = (N_1, ..., N_K)$ Beobachtungen in den verschiedenen Gruppen. Die Modellgleichung für ein binäres MVB-Modell zu diesem zweistufigen Zufallsexperiment entspricht im wesentlichen der eines „normalen" binären MVB-Modells gemäß Gleichung (3.5-1); lediglich die Modellparameter sind durch Zufallsparameter zu ersetzen. Dadurch wird natürlich auch jede Kategoriewahrscheinlichkeit zu einer Zufallsvariablen $P_{k,j}(\boldsymbol{\Theta})$:

$$\forall k, \forall j\colon P_{k,j}(\boldsymbol{\Theta}) = \sum_{i=1}^{I_{k,j}} c_{k,j,i} \cdot \prod_{s=1}^{S} \left[(1-\Theta_s)^{b_{k,j,i,s}} \cdot (\Theta_s)^{a_{k,j,i,s}} \right] \qquad (5.2\text{-}1)$$

[60] Es gibt selbstverständlich Ausnahmen. Beispielsweise ist die Hypothese, daß die mittlere Assoziation zweier dichotomer Variablen in einer Population von Individuen null ist, vermutlich häufiger von Interesse als die Hypothese, daß die Assoziation für jedes Individuum exakt null ist (Wickens, 1993). Derartige Anwendungen sind jedoch nicht unbedingt typisch für die Kognitionspsychologie.

Man beachte, daß bei Gültigkeit eines MVB-Modells mit Zufallsparametern gemäß Gleichung (5.2-1) die Gültigkeit eines strukturgleichen „normalen" MVB-Modells für jede konkrete Realisation $\boldsymbol{\Theta} = \boldsymbol{\theta}$ impliziert ist. „Strukturgleichheit" bedeutet, daß die $a_{k,j,i,s}$- und die $b_{k,j,i,s}$-Matrizen des normalen MVB-Modells gemäß Gleichung (3.5-1) denen des Modells für Zufallsparameter gemäß Gleichung (5.2-1) entsprechen. Ein solches MVB-Modell mit Zufallsparameters soll nun „aggregierungsinvariant" heißen, wenn die mittleren Kategoriewahrscheinlichkeiten — also die Erwartungswerte von $P_{k,j}(\boldsymbol{\Theta})$ — ebenfalls einem strukturgleichen normalen MVB-Modell gemäß Gleichung (3.5-1) entsprechen. Darüber hinaus wird für ein aggregierungsinvariantes Modell gefordert, daß die Parameter, welche die mittleren Kategoriewahrscheinlichkeiten beschreiben, gerade die Erwartungswerte der Zufallsparameter aus Gleichung (5.2-1) sind. Kurz: Ein binäres MVB-Modell mit Zufallsparametern gemäß Gleichung (5.2-1) heißt „aggregierungsinvariant" genau dann, wenn für die mittleren Kategoriewahrscheinlichkeiten gilt:

$$\forall k, \forall j\colon E(P_{k,j}(\boldsymbol{\Theta})) = \sum_{i=1}^{I_{k,j}} c_{k,j,i} \cdot \prod_{s=1}^{S} \left[\left(1 - E(\Theta_s)\right)^{b_{k,j,i,s}} \cdot E(\Theta_s)^{a_{k,j,i,s}} \right] \qquad (5.2\text{-}2)$$

Wie man aus den Rechenregel für Erwartungswerte (siehe z.B. Hays, 1973, Appendix B) leicht herleiten kann, folgt die für Aggregierungsinvarianz charakteristische Gleichung (5.2-2) aus Gleichung (5.2-1) dann und nur dann, wenn

(a) keiner der Exponenten $b_{k,j,is}$ und $a_{k,j,i,s}$ größer als 1 ist. Das ist genau dann der Fall, wenn kein Parameter in einem Ast der Ereignisbaumrepräsentation des Modells mehr als einmal vorkommt. Ferner muß gelten, daß

(b) ein Parameter Θ_S und dessen Komplement $1-\Theta_S$ niemals in demselben Ast (d.h. in demselben Produktterm) des Modells vorkommen, so daß
$a_{k,j,i,s} = 1 \Rightarrow b_{k,j,i,s} = 0$ sowie $b_{k,j,i,s} = 1 \Rightarrow a_{k,j,i,s} = 0$.
Schließlich ist zu fordern, daß

(c) alle Zufallsparameter (oder deren Komplemente), die in einem Ast eines Baumes vorkommen, paarweise unkorreliert sind, so daß für $s \in \{1, ..., S\}$, $s' \in \{1, ..., S\}$, $s \neq s'$, gilt:
$Cov(\Theta_s, \Theta_{s'}) = 0$, sofern $a_{k,j,i,s} = a_{k,j,i,s'} = 1$,
$b_{k,j,i,s} = b_{k,j,i,s'} = 1$,
$a_{k,j,i,s} = b_{k,j,i,s'} = 1$ oder $a_{k,j,i,s'} = b_{k,j,i,s} = 1$ gilt.

Bedingung (a) garantiert, daß $E(\Theta_s^{a_{k,j,i,s}}) = E(\Theta_s)^{a_{k,j,i,s}}$ und $E((1-\Theta_s)^{b_{k,j,i,s}}) = E(1-\Theta_s)^{b_{k,j,i,s}}$ gilt. Nimmt man Bedingung (b) hinzu, so folgt

$$E\left(\left(1 - \Theta_s\right)^{b_{k,j,i,s}} \cdot \Theta_s^{a_{k,j,i,s}} \right) = \left(1 - E(\Theta_s)\right)^{b_{k,j,i,s}} \cdot E(\Theta_s)^{a_{k,j,i,s}} \qquad (5.2\text{-}3)$$

Zusammen mit Bedingung (c) ergibt sich dann der Schluß, daß auch

$$E\left(\prod_{s=1}^{S}\left[\left(1-\Theta_s\right)^{b_{k,j,i,s}}\cdot\left(\Theta_s\right)^{a_{k,j,i,s}}\right]\right)=\prod_{s=1}^{S}\left[\left(1-E(\Theta_s)\right)^{b_{k,j,i,s}}\cdot E(\Theta_s)^{a_{k,j,i,s}}\right] \quad (5.2\text{-}4)$$

gelten muß. Alle drei Bedingungen zusammen sichern daher die Ableitbarkeit von Gleichung (5.2-2) aus Gleichung (5.2-1).

Binäre MVB-Modelle mit Zufallsparametern sind folglich aggregierungsinvariant, wenn in keinem „Ast" des entsprechenden Baummodells — bzw. in keinem Produktterm der Modellgleichung (5.2-1) — Zufallsparameter auftauchen, die korreliert sind. Wie man leicht nachvollziehen kann, gilt das gleiche Resultat auch für allgemeine MVB-Modelle mit nichtbinären Zufallsparametern. Eine hinreichende, aber nicht notwendige Bedingung für die Aggregierungsinvarianz von MVB-Modellen ist offenbar, daß (a) jeder Zufallsparameter oder dessen Komplement in jedem Produktterm maximal einmal vorkommt und (b) alle S Zufallsparameter stochastisch unabhängig sind.[61]

Die Forderung, daß kein Parameter mehr als einmal in einem Ast des Baummodells vorkommen darf, ist beim EA-Modell gemäß Abbildung 3.4 verletzt, wenn man es ohne weitere Modifikation auf ein Modell mit Zufallsparametern verallgemeinert. Der Zufallsparameter U oder dessen Komplement $(1-U)$ kämen dann nämlich in vier Ästen mehrfach vor, so daß perfekt positive Korrelationen (vgl. Äste 3 und 6 im Teilbaum zu den Itempaaren) oder perfekt negative Korrelationen vorliegen (vgl. Äste 4 und 5 im Teilbaum zu den Itempaaren). Genau dieses Modell mit Zufallsparametern haben Riefer und Batchelder (1991b) in ihrer Monte-Carlo-Studie untersucht. Da dieses Modell nicht aggregierungsinvariant ist, ist die mit zunehmendem Stichprobenumfang wachsende Rejektionsrate der Modellgeltungshypothese bei konstantem nominellen α-Fehler nicht überraschend: Hierin drückt sich lediglich aus, daß die Modellgeltungshypothese für die aggregierten Daten aufgrund der korrelierten Parameter nicht gilt. Es kann daher nicht erwartet werden, daß sich der Modellgeltungstest in diesem Fall gegenüber Parametervariabilität sehr robust verhält, es sei denn, die Varianz der Parameter ist so gering, daß die Abweichung der aggregierten Daten vom Modell nur minimal ausfällt.

Man kann das EA-Modell jedoch problemlos auf ein Modell mit Zufallsparametern erweitern, das die Forderung nach Aggregierungsinvarianz erfüllt. Hierzu werden für das erste Wort eines Itempaares und das zweite Wort eines Itempaares zwei *verschiedene* Abrufparameter U_1 und U_2 eingeführt, die zwar den gleichen Mittelwert und die gleiche Streuung aufweisen, jedoch unkorreliert sind. Im Durchschnitt werden die beiden Wörter also gleich gut abgerufen, jedoch ist in

[61] Ich danke Reinhard Niederée für eine Reihe von äußerst fruchtbaren Diskussionen zu diesem Punkt. Die Einsicht, daß stochastische Unabhängigkeit aller Zufallsparameter innerhalb eines MVB-Astes hinreichend für Aggregierungsinvarianz des Modells ist, geht auf seine Anregung zurück.

jedem Einzelfall die Realisation von U_2 aus der von U_1 nicht vorhersagbar und vice versa. Bezüglich dieses Modells ist die Frage nach der Robustheit des Tests der Aggregathypothese gegenüber Parametervariabilität sinnvoll: Wir wissen, daß die Modellgeltungshypothese des EA-Modells für die aggregierten Daten gelten muß, wenn das Modell mit Zufallsparametern gilt. Lediglich die im Modellgeltungstest gemachte Voraussetzung konstanter Parameter über alle Stichprobendaten hinweg ist verletzt.

Welche Ergebnisse sind in bezug auf die Robustheit des Modellgeltungstests gegenüber Parametervariabilität bei aggregierungsinvarianten Modellen zu erwarten? Tendenziell sollte der Modelltest nicht mehr, sondern weniger α-Fehler produzieren als nominell vorgeben, da die Konfidenzintervalle mit zunehmender Parametervariabilität immer größer werden. Als Konsequenz daraus ist ein konservatives Verhalten des Modelltests zu erwarten: Es wird häufiger zugunsten der Aggregat-Modellgeltungshypothese (H_0) entschieden, als es aufgrund des nominellen α-Risikos zu erwarten ist. Zu vermuten ist, daß sich dies auch negativ auf die Teststärke des Modelltests auswirkt. Die nominelle Teststärke zum EA-Modelltest bei konstanten Parametern und einer Parameterdifferenz von z.B. $\Delta = |a - u| = |.60 - .40| = .20$ ist also vermutlich größer als die faktische Teststärke zum Aggregat-Modelltest bei Zufallsparametern und einer Parametermittelwertsdifferenz von $|E(A) - E(U)| = |.60 - .40| = .20$. Das Ausmaß des zu erwartenden konservativen Verhaltens ist unklar und analytisch nur schwer zu klären. Um zu untersuchen, ob und gegebenenfalls in welchem Ausmaß sich der Modelltest bei aggregierungsinvarianten Modellen mit Zufallsparametern konservativ verhält, wurde daher eine weitere Monte-Carlo-Studie zum EA-Modell von Batchelder und Riefer durchgeführt.

5.3 Eine Monte-Carlo-Studie zur Robustheit bei EA-Modellanalysen

Die in diesem Abschnitt zu berichtende Monte-Carlo-Studie stellt teilweise eine Replikation der Studie von Riefer und Batchelder (1991b, S. 330 ff) zum Enkodierung-Abruf-Modell (EA-Modell) mit korrelierten Zufallsparametern dar. Mit dieser Replikation soll gezeigt werden, daß die hier verwendeten Simulationsmethoden Ergebnisse liefern, die bei Gleichheit des Simulationsmodells mit denen von Riefer und Batchelder (1991b) übereinstimmen. Zusätzlich wurde der Modelltest zum EA-Modell mit unkorrelierten Zufallsparametern C, R, U_1 (für Item 1 eines Itempaares), U_2 (für Item 2 eines Itempaares) und A (für Einzelitems) untersucht, um Informationen über das Verhalten des Tests bei aggregierungsinvarianten Modellen zu erhalten. Für beide Modelle — mit und ohne korrelierte Zufallsparameter — wurde darüber hinaus der mittlere Bias in den Parameterschätzungen relativ zum zugrundeliegenden Erwartungswert des Zufallsparameters geschätzt.

Für Daten, die nach dem Modell mit korrelierten Zufallsparametern erzeugt werden, ist ein mehr oder minder starker Bias auch bei großem N zu erwarten, da das gewöhnliche EA-Modell für die aggregierten Daten fehlspezifiziert ist. Werden die Daten dagegen nach dem Modell unkorrelierter Zufallsparameter erzeugt, so sollte unabhängig von der Größe der Parameterstreuungen zumindest bei großem N kein mittlerer Schätzfehler resultieren, da das gewöhnliche EA-Modell für die aggregierten Daten adäquat ist und dessen Parameter gerade den Erwartungswerten der Zufallsparameter entsprechen müssen (vgl. Gleichung (5.2-2)).

Wie bei Riefer und Batchelder wurden unabhängige Beta-Verteilungen aller Zufallsparameter simuliert.[62] Die Streuungen der simulierten Beta-Verteilungen betrugen $\sigma = .10$ oder $\sigma = .20$. Als Mittelwert wurde für alle Zufallsparameter $\mu = .50$ festgesetzt, abgesehen von den Untersuchungen zur Teststärke, auf die weiter unten einzugehen ist. Die Erzeugung beta-verteilter Zufallsvariablen mit Streuung $\sigma = .20$ erfolgte unter Benutzung der Rejektionsmethode von Jöhnk (1964), die von unabhängigen, gleichverteilten Zufallszahlen X und Y im Intervall [0, 1] ausgeht (vgl. auch Johnson & Kotz, 1970b, S. 39). X und Y wurden — wie schon im Falle der Monte-Carlo-Studie zur Teststärke in Abschnitt 4.3.3.5 — mit der RND-Funktion des GW-BASIC-Interpreters für das DOS-Betriebssystem erzeugt (vgl. Fußnote 46). Jöhnk (1964) hat gezeigt, daß für beliebige positiv-reellwertige Parameter n und r die bedingte Verteilung von $X^{(1/n)}$, gegeben $X^{(1/n)} + Y^{(1/r)} \leq 1$, eine Beta-Verteilung B(p, q) mit Parametern $p = n$ und $q = r + 1$ ist. Diese Parameter p und q stehen wie folgt zum Mittelwert und zur Streuung der Beta-Verteilung in Beziehung (vgl. Johnson & Kotz, 1970b, S. 40):

$$\mu_{p,q} = p / (p + q) \text{ und}$$
$$\sigma_{p,q} = \sqrt{\frac{pq}{(p+q)^2(p+q+1)}}. \qquad (5.3\text{-}1)$$

Eine beta-verteilte Zufallsvariable mit Mittelwert .5 und Streuung .2 entspricht also beispielsweise einer B(p, q)-Verteilung mit $p = q = 2.625$. Sie kann erzeugt werden, indem zunächst zwei gleichverteilte Zufallsvariablen X und Y generiert und dann nur die Fälle berücksichtigt werden, in denen $X^{(1/2.625)} + Y^{(1/(2.625-1))} \leq 1$ erfüllt ist. Die Zufallsvariable $Z := X^{(1/2.625)}$ weist für die verbleibenden Fälle die gewünschte Beta-Verteilung auf.

Prinzipiell kann auf diese Weise jede beliebige beta-verteilte ZV erzeugt werden. Allerdings ist die geschilderte Rejektionsmethode für manche Werte von p und q ineffizient, weil zu viele Realisationen gleichverteilter ZVn X und Y der

[62] Prinzipiell lassen sich Simulationsstudien zu MVB-Modellen mit beta-verteilten Zufallsparametern mit den Computerprogrammen MBT (Hu, 1995) und GPT (Hu, 1999; Hu & Phillips, 1999) durchführen. Aus den bereits in Fußnote 46 erläuterten Gründen wurde jedoch ein eigenes Programm zur Erzeugung der Monte-Carlo-Stichproben erstellt.

geforderten Ungleichheitsrelation nicht genügen. Deshalb wurden beta-verteilte Zufallszahlen mit Mittelwert .5 und Streuung .1 ($p = q = 12$) dadurch generiert, daß zunächst $p + q - 1 = 23$ unabhängige gleichverteilte ZVn im Intervall [0, 1] gewonnen wurden. Man kann zeigten, daß eine B(p, q)-verteilte ZV resultiert, wenn jeweils die p-tkleinste dieser $p + q - 1$ Zufallszahlen herausgegriffen wird, in diesem Fall also die zwölfte (vgl. Ross, 1997, S. 577). Diese Selektionsmethode funktioniert natürlich nur für ganzzahlige Parameter p und q und ist bei großen Werten für p und / oder q ebenfalls recht ineffizient. Für $p = q = 12$ ist sie jedoch zeitsparender als die oben geschilderte Rejektionsmethode von Jöhnk (1964).

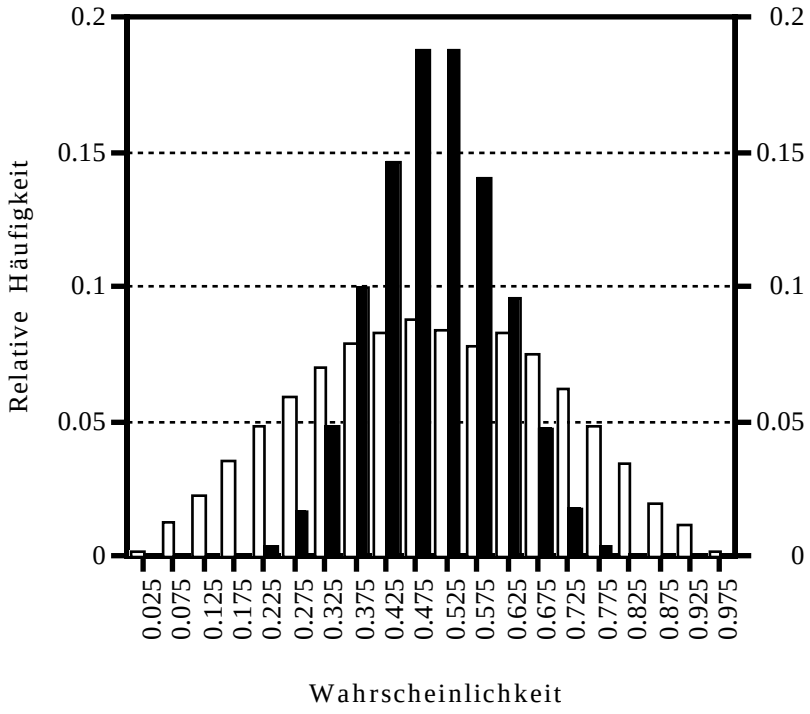

Abbildung 5.1:
Histogramme zu Stichproben des Umfangs $N = 10000$, die (a) mit der Rejektionsmethode von Jöhnk (1964) zu einer Beta-Verteilung mit $\mu = .5$ und $\sigma = .2$ (weiße Balken) und (b) mit der Selektionsmethode nach Ross (1997) zu einer Beta-Verteilung mit $\mu = .5$ und $\sigma = .1$ erzeugt wurden (schwarze Balken).

Abbildung 5.1 illustriert Histogramme zu Stichproben des Umfangs $N = 10000$ aus Beta-Verteilungen mit Mittelwert $\mu = .5$ und Streuung $\sigma = .1$ (schwarze Balken) sowie Mittelwert $\mu = .5$ und Streuung $\sigma = .2$ (weiße Balken), die mit den oben geschilderten Simulationsmethoden gewonnen wurden. Die Stichprobenmittelwerte und -streuungen betragen $\hat{\mu} = .4993$ bzw. $\hat{\sigma} = .1994$ für die Stichprobe aus der Population mit $\sigma = .20$ und $\hat{\mu} = .4995$ bzw. $\hat{\sigma} = .1009$ für die Stichprobe aus der Population mit $\sigma = .10$. Diese Illustration und weitere vorgenommene Tests zeigen, daß die Simulationsmethoden korrekt implementiert wurden und zufriedenstellend funktionieren. Zugleich machen die Histogramme deutlich, daß eine Beta-Streuung von $\sigma = .20$ durchaus schon eine erhebliche Variabilität der Parame-

ter mit sich bringt. In praktischen Anwendungskontexten ist wohl kaum damit zu rechnen, daß die Parameterstreuung wesentlich größer als $\sigma = .20$ ausfällt (vgl. Riefer & Batchelder, 1991b).

In der Monte-Carlo-Studie zum EA-Modell mit korrelierten Zufallsparametern wurden pro Stichprobendatum Realisationen von drei unabhängigen beta-verteilten ZVn C, R und U erzeugt; die zugehörigen Kategoriewahrscheinlichkeiten (in diesem Kontext ebenfalls Realisationen von Zufallsvariablen) wurden sodann gemäß

$$\begin{aligned} P_{1,1}(C,R,U) &= C \cdot R \\ P_{1,2}(C,R,U) &= (1-C) \cdot U^2 \\ P_{1,3}(C,R,U) &= 2 \cdot (1-C) \cdot U \cdot (1-U) \\ P_{1,3}(C,R,U) &= C \cdot (1-R) + (1-C) \cdot (1-U)^2 \\ P_{2,1}(C,R,U) &= U \\ P_{2,2}(C,R,U) &= (1-U) \end{aligned} \qquad (5.3\text{-}2)$$

aus den jeweiligen Realisationen der Zufallsparameter errechnet. Im Modell mit unkorrelierten Zufallsparametern wurden ganz analog Realisationen von fünf unabhängigen beta-verteilten ZVn C, R, U_1, U_2 und A pro Stichprobendatum erzeugt; die Kategoriewahrscheinlichkeiten wurden dann gemäß

$$\begin{aligned} P_{1,1}(C,R,U) &= C \cdot R \\ P_{1,2}(C,R,U) &= (1-C) \cdot U_1 \cdot U_2 \\ P_{1,3}(C,R,U) &= (1-C) \cdot (U_1 \cdot (1-U_2) + (1-U_1) \cdot U_2) \\ P_{1,4}(C,R,U) &= C \cdot (1-R) + (1-C) \cdot (1-U_1) \cdot (1-U_2) \\ P_{2,1}(C,R,U) &= A \\ P_{2,2}(C,R,U) &= (1-A) \end{aligned} \qquad (5.3\text{-}3)$$

bestimmt. Man beachte, daß die Modellgleichungen (5.3-3) zu denen in (5.3-2) äquivalent wären, wenn bei jedem einzelnen Stichprobenelement $U_1 = U_2 = A$ gelten würde. Allein durch die Unabhängigkeit der drei Zufallsvariablen — erkennbar nur an der gemeinsamen Verteilung der drei ZVn U_1, U_2 und A , nicht an den univariaten Randverteilungen — ergibt sich also ein Unterschied.

Im Einklang mit den gemäß (5.3-2) bzw. (5.3-3) für jedes Stichprobenelement getrennt bestimmten Kategoriewahrscheinlichkeiten wurden Beobachtungsresultate für die 4 + 2 Beobachtungskategorien auf die gleiche Weise generiert, die auch in der Monte-Carlo-Studie zur Teststärke des EA-Modelltests in Kapitel 4.3.3.5 Verwendung fand. Die realisierten Gesamtstichprobenumfänge betrugen N = 180, 360, 720, 1440, 2880, 5760 und 11520, wobei wie schon in Kapitel 4.3.3.5 zwei Drittel des Gesamtstichprobenumfangs auf Itempaare und ein Drittel auf Einzelitems entfielen.

Pro Stichprobenumfang N wurden 500 Monte-Carlo-Stichproben erzeugt. An jede dieser Monte-Carlo-Stichproben wurde das gewöhnliche Drei-Parameter-EA-Modell (vgl. Abschnitt 3.5.1) mittels AppleTree (Rothkegel, 1998; 1999) unter Verwendung des EM-Algorithmus angepaßt, und es wurden jeweils die G^2-Anpassungsstatistiken und die ML-Schätzwerte der Parameter c, r und u errechnet. Ein Großteil der Analysen wurde auch für die Pearson- und die Cressie-Read-Statistik repliziert; die Ergebnisse unterschieden sich jedoch von denen zur G^2-Statistik nur so minimal, daß hier ausschließlich die Resultate zur ML-Methode berichtet werden.

Tabelle 51:
Mittelwerte der ML-Parameterschätzungen zum EA-Modell bei jeweils 500 Monte-Carlo-Datensätzen, die gemäß EA-Modellen mit korrelierten (Gleichungen (5.3-2)) vs. unkorrelierten Zufallsparametern (Gleichungen (5.3-3)) erzeugt wurden; in der vorletzten Zeile (∅) ist der mittlere, in der letzten Zeile (max) der maximale Bias der Schätzer über die verschiedenen Stichprobenumfänge hinweg angegeben.

	Parameterstreuung σ = .10						Parameterstreuung σ = .20					
	korrelierte P.			unkorrelierte P.			korrelierte P.			unkorrelierte P.		
N	$\bar{\hat{c}}$	$\bar{\hat{r}}$	$\bar{\hat{u}}$	$\bar{\hat{c}}$	$\bar{\hat{r}}$	$\bar{\hat{u}}$	$\bar{\hat{c}}$	$\bar{\hat{r}}$	$\bar{\hat{u}}$	$\bar{\hat{c}}$	$\bar{\hat{r}}$	$\bar{\hat{u}}$
180	.5077	.4948	.5058	.4942	.5142	.4988	.5394	.4667	.5302	.4912	.5128	.4991
360	.5116	.4906	.5091	.4998	.5046	.4984	.5442	.4648	.5304	.4980	.5026	.5006
720	.5101	.4912	.5082	.4986	.5020	.5005	.5462	.4580	.5311	.4989	.5014	.5001
1440	.5103	.4890	.5087	.5003	.5010	.4999	.5458	.4593	.5311	.4993	.5000	.5002
2880	.5115	.4893	.5079	.5000	.4994	.4998	.5461	.4591	.5310	.4992	.4995	.5002
5760	.5118	.4880	.5075	.4998	.5014	.4995	.5463	.4584	.5313	.4994	.4992	.5002
11520	.5114	.4892	.5080	.4999	.5007	.4998	.5463	.4585	.5312	.4995	.4991	.5003
∅	+.011	-.010	+.008	-.001	+.003	.000	+.044	-.039	+.031	-.002	+.002	.000
max	+.012	-.012	+.009	-.006	+.014	-.002	+.046	-.042	+.031	-.009	+.013	+.001

Tabelle 5.1 stellt zunächst die Ergebnisse zum mittleren Schätzfehler der drei Parameter zusammen. Die Ergebnisse fallen erwartungsgemäß aus. Die Schätzer zu Daten, die nach dem Modell korrelierter Zufallsparameter erzeugt wurden, weichen im Durchschnitt vom zugrundeliegenden Mittelwert der Zufallsparameter (μ = .5) ab, und zwar um so mehr, je größer die Streuung der Zufallsparameter ausfällt. Besonders ausgeprägt ist dies beim Schätzer des Parameters c, der einen Bias von etwa +.01 bei der Parameterstreuung .10 und von fast +.05 bei der Parameterstreuung .20 aufweist. Auffällig ist auch, daß der Bias mit zunehmendem Stichprobenumfang nicht kleiner wird; tendenziell finden sich die größeren Abweichungen eher bei den größeren Stichprobenumfängen.

Genau umgekehrt sieht das Bild aus, wenn das EA-Modell an die nach dem Modell unkorrelierter Zufallsparameter erzeugten Daten angepaßt wird: Abweichungen vom Erwartungswert .50 liegen eher bei kleineren als bei größeren Stichprobenumfängen vor. Im Durchschnitt über alle Stichprobenumfänge hinweg ist der Bias bei allen Schätzern nahezu 0.00. Mit anderen Worten: Die ML-Punktschätzungen sind robust gegenüber Parametervariabilität in MVB-Modellen, wenn die Parameter innerhalb der Äste des Modells unkorreliert sind. Sind die Parameter dagegen korreliert, ist Robustheit nicht generell garantiert, sondern nur bei kleinen Parameterstreuungen ($\sigma \leq .10$).

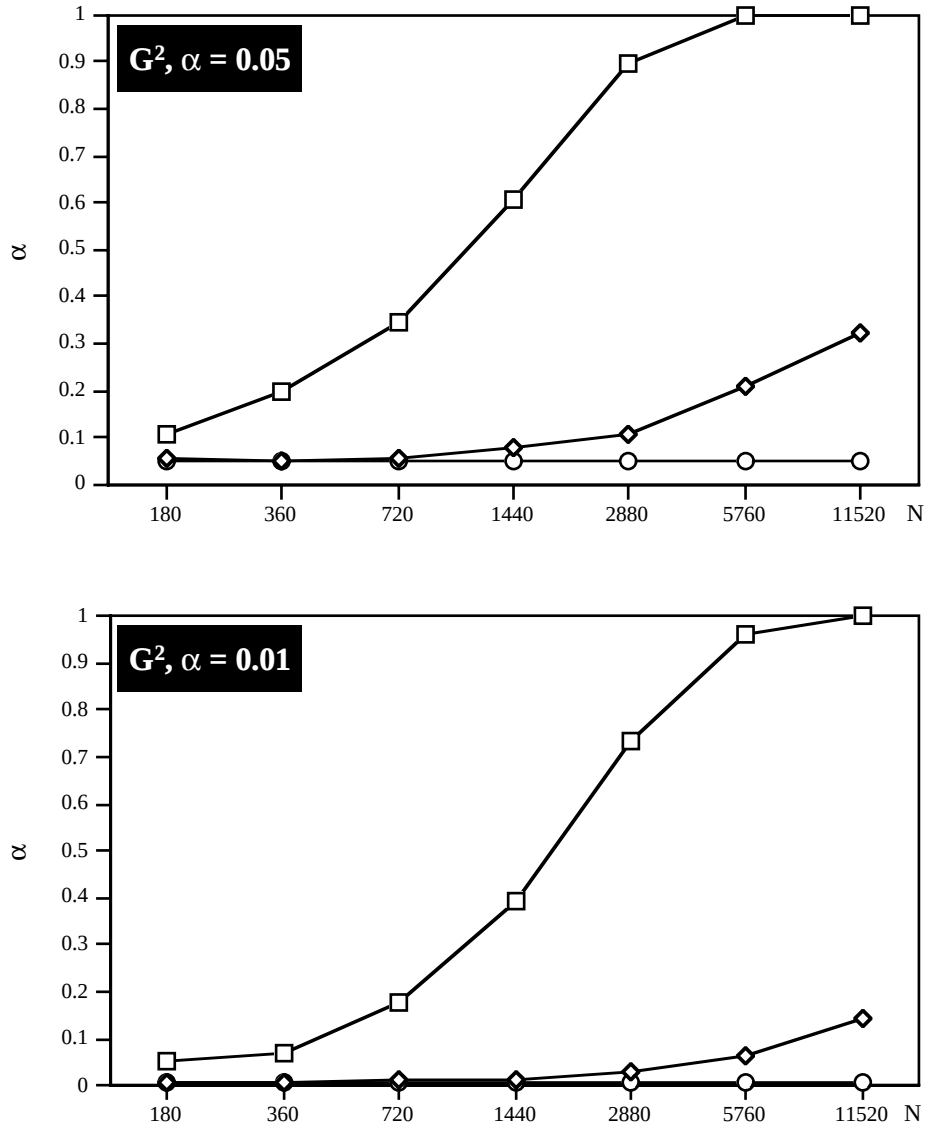

Abbildung 5.2:
Veranschaulichung des faktischen Alpha-Fehlers in Abhängigkeit vom Gesamtstichprobenumfang *N* bei Anwendung des Log-Likelihood-Quotienten-EA-Modelltests auf Daten, die einem Modell mit korrelierten Zufallsparametern mit Streuung .20 (¤) bzw. .10 (◊) folgen; getrennt für die nominellen Alpha-Niveaus $\alpha = .05$ (oben) und $\alpha = .01$ (unten). Das nominelle Signifikanzniveau ist jeweils durch einen Kreis (o) markiert.

Abbildung 5.2 veranschaulicht die Ergebnisse zum G^2-Modellanpassungstest an die aggregierten Daten, die nach dem Modell korrelierter Zufallsparameter mit Erwartungswert $E(C) = E(R) = E(U) = .50$ und Streuung $\sigma = .20$ bzw. $\sigma = .10$ erzeugt wurden. Pro Stichprobenumfang wurde jeweils ausgezählt, welcher Anteil der G^2-Anpassungsstatistiken zu den 500 Monte-Carlo-Stichproben den nominellen kritischen $\chi^2_{(1)}$-Wert zum Niveau $\alpha = .05$ und zum Niveau $\alpha = .01$ überschreitet. Die Ergebnisse spiegeln die Monte-Carlo-Resultate von Riefer und Batchelder (1991b, Abbildung 9) nahezu perfekt wider: Obwohl das EA-Modell für jedes einzelne Stichprobendatum gilt, ist der auf die aggregierten Daten angewandte Modelltest bei einer Parameterstreuung von $\sigma = .20$ durchweg deutlich antikonservativ. Bei sehr großen Stichprobenumfängen ist die Verwerfung der Modellgeltungshypothese nahezu sicher. Beträgt die Streuung dagegen nur $\sigma = .10$, so ist der Anpassungstest bis zu einem N von ca. 1500 halbwegs robust und wird erst bei größerem N antikonservativ. Die hohe Übereinstimmung der Resultate zum nominellen $\alpha = .05$ mit den entsprechenden Ergebnisse von Riefer und Batchelder (1991b) belegt die Korrektheit der hier verwendeten Simulationsmethoden. Zudem offenbaren die Ergebnisse analoge antikonservative Trends des Modelltests auch bei $\alpha = .01$, was aufgrund der Tatsache, daß das zugrundeliegende Modell nicht aggregierungsinvariant ist, nicht überraschen kann.

Abbildung 5.3 veranschaulicht die zu Abbildung 5.2 analogen Ergebnisse bei Anwendung des G^2-Modellanpassungstest auf Daten, die nach dem Modell unkorrelierter Zufallsparameter erzeugt wurden. Man sieht, daß eine scheinbar minimale Änderung des Simulationsmodells — der Übergang von perfekt korrelierten zu unkorrelierten U_1-U_2-Parametern drastische Konsequenzen nach sich zieht: Der Modelltest ist nun tendenziell konservativ (zumindest ab $N = 2000$) und nicht mehr antikonservativ. Dies ist genau das, was man bei aggregierungsinvarianten Modellen mit zunehmender Parameterstreuung erwarten muß: Der Modelltest verwirft das für die aggregierten Daten nunmehr gültige Modell relativ zum spezifizierten α-Niveau zu selten, weil die Konfidenzintervalle der relevanten Parameter aufgrund der Parametervariabilität größer sind als im Standard-EA-Modell mit fixen Parametern angenommen.

Das konservative Verhalten des Modelltests bei unkorrelierten Zufallsparametern ist per se nicht negativ zu beurteilen. Wenn man bei Modelltests weniger α-Fehler macht als erwartet, so ist das natürlich zunächst erfreulicher, als wenn man zu viele α-Fehler macht. Grund zur Sorge gibt das in Abbildung 5.3 veranschaulichte Ergebnis allein deshalb, weil es die Vermutung nahelegt, daß sich der Modelltest der H_0: $u = a$ bei Anwendung auf aggregierte Daten auch dann noch konservativ verhält, wenn die Gleichheit der Erwartungswerte $E(U_1) = E(U_2) = E(A)$ nicht mehr vorausgesetzt werden kann. Praktisch hieße das, daß zwar die α-

Fehlerrate des Modelltest bei unkorrelierten Zufallsparametern eingehalten wird, nicht jedoch die nominelle Teststärke (1–β) bei Ungültigkeit des Modells.

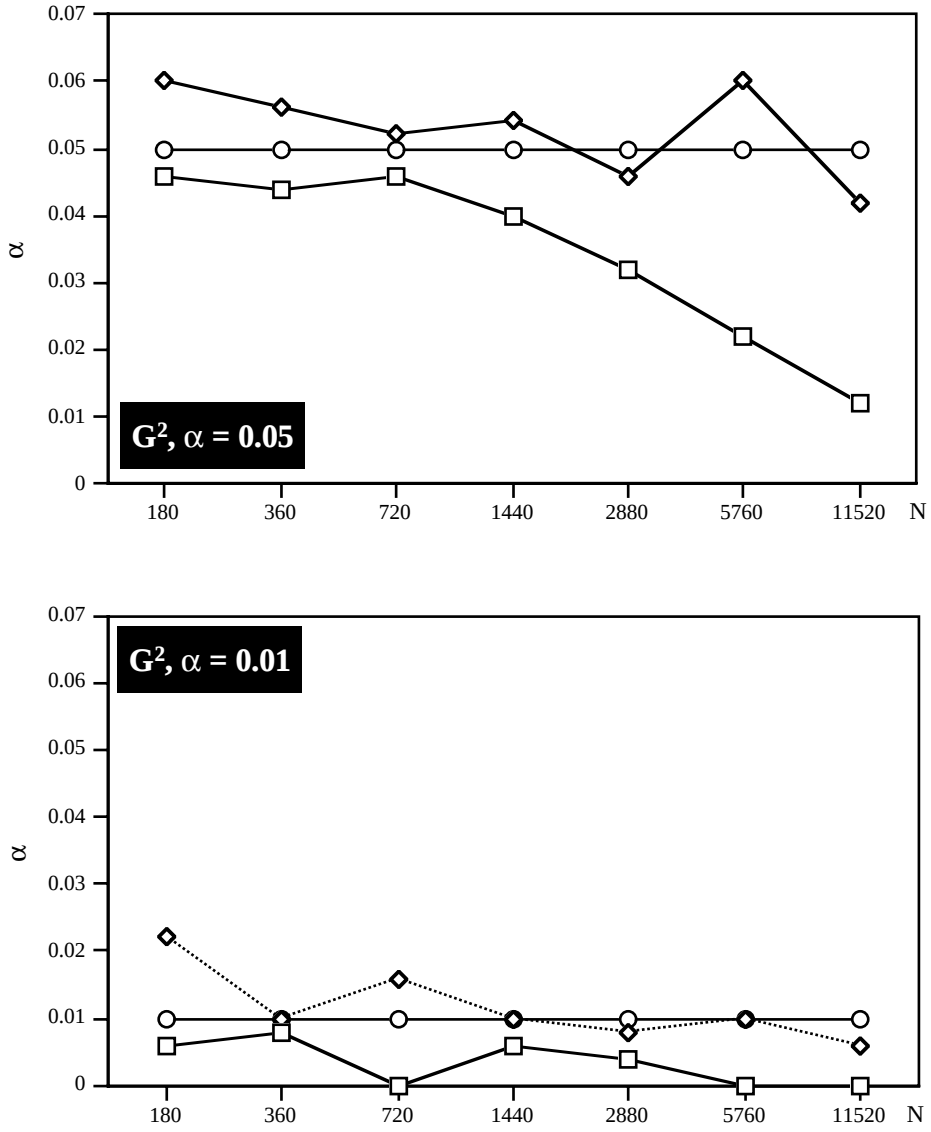

Abbildung 5.3:
Veranschaulichung des faktischen Alpha-Fehlers in Abhängigkeit vom Gesamtstichprobenumfang *N* bei Anwendung des Log-Likelihood-Quotienten-EA-Modelltests auf Daten, die einem Modell mit fünf unkorrelierten Zufallsparametern mit Streuung .20 (¤) bzw. 10 (◊) folgen; getrennt für die nominellen Alpha-Niveaus α = .05 (oben) und α = .01 (unten). Das nominelle Signifikanzniveau ist jeweils durch einen Kreis (o) markiert.

Um zu überprüfen, ob und gegebenenfalls wie stark sich das konservative Verhalten des Modelltests bei Modellgültigkeit auf die Teststärke bei Modellungültigkeit auswirkt, wurde eine weitere Monte-Carlo-Studie zum Modell mit unkorrelierten Zufallsparametern durchgeführt. Drei Stufen der Modellverletzung wurden simuliert: Δ = .10 ($E(U_1) = E(U_2) = .45$ und $E(A) = .55$), Δ = .20 ($E(U_1) = E(U_2) = .40$ und $E(A) = .60$) und schließlich Δ = .40 ($E(U_1) = E(U_2) = .30$ und $E(A) = .70$). Generell war $E(C) = E(R) = .50$. Es wurde ausschließlich die Streuung σ = .20 der Zu-

fallsparameter realisiert, bei der in der zuvor geschilderten Monte-Carlo-Studie unter H_0 das konservativste Verhalten zu beobachten war.

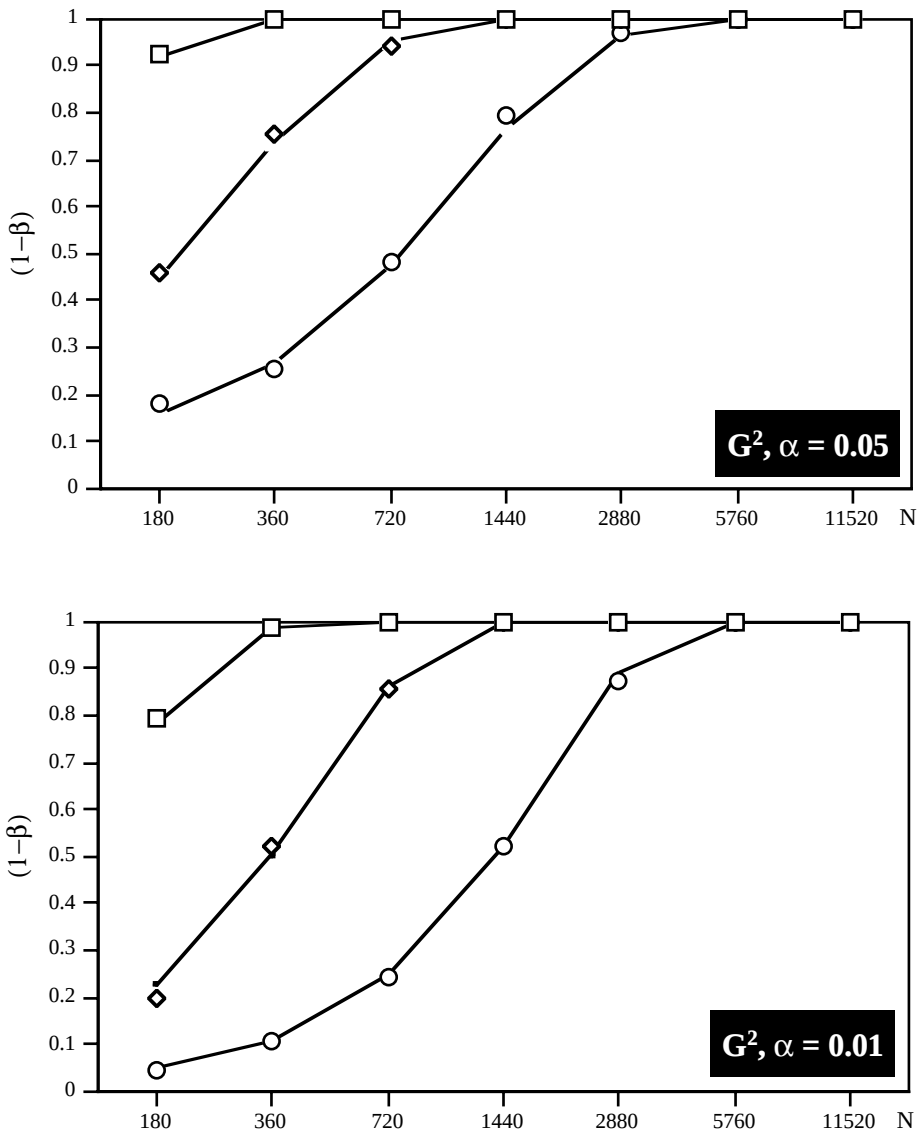

Abbildung 5.4:
Veranschaulichung der Beziehung zwischen empirisch geschätzten Teststärkewerten (Symbole o, ◊ und ¤) und nominellen Teststärkewerten (durchgezogene Linien) in Abhängigkeit vom Gesamtstichprobenumfang *N* bei Anwendung des Log-Likelihood-Quotienten-EA-Modelltests auf Daten, die einem Modell mit fünf unkorrelierten Zufallsparametern mit Streuung .20 folgen; getrennt für verschiedene Differenzen Δ zwischen $E(U_1) = E(U_2)$ und $E(A)$ (o: Δ = .10, ◊: Δ = .20, ¤: Δ = .40) sowie nominelle Alpha-Niveaus von α = .05 (oben) und α = .01 (unten).

Abbildung 5.4 veranschaulicht die empirisch ermittelten Überschreitungsraten der kritischen $\chi^2_{(1)}$-Werte („faktische Teststärke") als Symbole und die nominellen Teststärkewerte als durchgezogene Linien. Die nominellen Teststärkewerte wurden mittels der nonzentralen χ^2-Approximation mit Nonzentralitätsparameter $\gamma^{(\lambda=0)}$ bestimmt, die auch in der Monte-Carlo-Studie zur Teststärke in Abschnitt 4.3.3.5 verwendet worden war. Zugrunde gelegt wurden bei diesen Berech-

nungen also H_0- und H_1-Modelle mit fixierten Parametern. Wie Abbildung 5.4 zeigt, stimmen die nominellen und die faktischen Teststärkewerte trotz der Abweichung vom Simulationsmodell und trotz der recht großen Parameterstreuung von $\sigma = .20$ sehr gut überein. Bei Stichprobenumfängen bis $N = 2000$ ist das nicht überraschend, weil der Modelltest sich in diesem Bereich auch bei $E(U_1) = E(U_2) = E(A) = .50$ nicht konservativ verhält. Die hohe Übereinstimmung bei noch größeren Stichprobenumfängen war jedoch nicht unbedingt zu erwarten. Zurückzuführen ist sie vermutlich darauf, daß in diesem Bereich die Teststärke bereits so groß ist, daß Abweichungen zwischen nominellen und faktischen α-Niveaus diesbezüglich kaum noch ins Gewicht fallen.

Selbstverständlich kann aus diesen Ergebnissen nicht der Schluß gezogen werden, daß der EA-Modelltest bei aggregierten Daten, die einem Modell mit unkorrelierten Zufallsparametern folgen, *immer* robust ist. Wenn noch kleinere Modellverletzungen als $\Delta = .10$ vorliegen und/oder größere Streuungen der Parameter als $\sigma = .20$ untersucht werden, ist ein konservatives Verhalten des Modelltests auch unter H_1 zu erwarten. Für die Anwendungspraxis dürfte dies aber kaum relevant sein, weil winzig kleine Diskrepanzen zwischen den durchschnittlichen *U*- und *A*-Parametern nicht von Interesse und Parameterstreuungen größer als .20 nicht zu erwarten sind.

Als Fazit kann festgehalten werden, daß die statistischen Standardanalysen zum EA-Modell (ML-Parameterschätzungen und G^2-Tests) gegenüber Parametervariabilität weitgehend robust sind, sofern die betreffenden Zufallsparameter innerhalb der Äste eines MVB-Modells unkorreliert sind. Nur bei sehr großen Parameterstreuungen und/oder exzessiv großen Stichprobenumfängen muß mit einem konservativen Verhalten des Modelltests zugunsten der Nullhypothese gerechnet werden. ML-Parameterschätzungen sind zumindest bei großen Stichprobenumfängen erwartungstreu. Anders sieht es bei korrelierten Zufallsparametern, d.h. bei nicht aggregierungsinvarianten Modellen aus: Die Parameterschätzungen weisen typischerweise einen Bias auf, der auch mit wachsendem *N* nicht verschwindet. Die Modelltests sind tendenziell antikonservativ. Die geschilderten Verzerrungen sind nicht gravierend, wenn die Parameterstreuungen klein und die Stichprobenumfänge nicht exzessiv groß sind. Bei Parameterstreuung im Bereich $\sigma = .20$ und größer kann jedoch nicht mehr von einem robusten Verhalten der Standardanalyseverfahren gesprochen werden (vgl. auch Riefer & Batchelder, 1991b; Rouder & Batchelder, 1998).

In praktischer Hinsicht ergibt sich aus den obigen Ergebnissen die Empfehlung, in Anwendungskontexten alle möglichen versuchsplanerischen Maßnahmen zu ergreifen, welche geeignet sind, (a) die Korrelationen zwischen den Modellparametern und (b) deren Streuungen zu reduzieren. Strenge Selektionskriterien für Vpn, die in eine Untersuchung eingehen, und Zufallssampling des Item-

materials aus einem großen Itempool — für jede Vp getrennt — können Schritte in die richtige Richtung sein. Zufallssampling des Itemmaterials schließt Korrelationen zwischen den Modellparametern aufgrund von Itemeffekten weitgehend aus. Homogene Vpn-Stichproben machen darüber hinaus Korrelationen zwischen den Parametern aufgrund von Personeneffekten unwahrscheinlich. Ein Gewähr für die Robustheit der statistischen Analysen bieten die genannten Maßnahmen allerdings nicht.

5.4 Alternativen zum Standardvorgehen bei Parametervariabilität

Sofern klar ist, daß die zu analysierenden Vpn- und Itemstichproben heterogen sind und daß Parameter innerhalb der Äste einer MVB-Modells über Vp-Item-Kombinationen hinweg stark korreliert sind, muß man sich nach Alternativen zur statistischen Standardanalyse der Daten umschauen oder zumindest Vorsichtsmaßnahmen gegen mögliche Fehlschlüsse ergreifen. Gleiches gilt, wenn bereits durchgeführte Monte-Carlo-Studien Hinweise auf die mangelnde Robustheit statistischer Analysen zum interessierenden Modell geliefert haben. Denkbar ist, daß es um die Robustheit von MVB-Modellen nicht immer so günstig wie im Falle des EA-Modells bestellt ist.

Welche Möglichkeiten stehen Anwendern offen, wenn nicht von der Robustheit statistischer Standardanalysen ausgegangen werden kann? Grob kann man zwei Lösungsansätze zum Variabilitätsproblem unterscheiden (vgl. auch Wickens, 1993; 1998a): (1) Änderung der Analyseebene vom Aggregat auf Teildatensätze, für die konstante Parameter angenommen werden können; (2) Übergang zu allgemeineren Modellen für die aggregierten Daten, in denen die Parameterheterogenität explizit mitmodelliert wird. Wie deutlich werden wird, können derartige Modelle durchaus selbst wieder MVB-Modelle sein und daher mit den bereits besprochenen Standardtechniken analysiert werden. Zunächst sollen jedoch Verfahren zum Ansatz (1) besprochen werden. Anschließend gehe ich auf verschiedene Möglichkeiten ein, MVB-Modelle zu Modellen für heterogene Grundgesamtheiten zu verallgemeinern.

5.4.1 Zerlegung von Aggregatstichproben

Der naheliegendste Lösungsansatz zum Variabilitätsproblem besteht darin, die Analyseebene vom Vpn-Item-Aggregat auf Subebenen zu verlagern. Man modelliert dann nicht mehr die aggregierten Daten, sondern eine Aufsplittung der aggregierten Daten in homogene Teildatensätze, so daß sich die Parameter nur noch zwischen den Teildatensätzen unterscheiden (vgl. Wickens, 1993). Wenn jede Vp hinreichend viele Daten produziert (z.B. $N_V \geq 50$), könnte man beispielsweise die Daten der Vpn separat untersuchen und individuelle Parameterschätzungen gege-

benenfalls über die Vpn hinweg mitteln, um einen Schätzer für den mittleren Parameter in der betreffenden Versuchsgruppe zu erhalten. Dies ist ein recht häufig anzutreffendes Vorgehen. Zu den Arbeiten, die individuelle Parameterschätzungen berechnet und dann gemittelt bzw. mittels Varianz- oder Korrelationsanalysen weiteranalysiert haben, zählen die in Kapitel 7 zusammengefaßten Arbeiten von Chechile und Mitarbeitern (z.B. Chechile & Meyer, 1976), Henkel, Johnson und de Leonardis (1998), Jacoby und Yonelinas (z.B. Jacoby, 1991; Yonelinas et al., 1995), Prinzmetal und Mitarbeitern (Ashby, Prinzmetal, Ivry und Maddox, 1996; Prinzmetal, Henderson & Ivry, 1995), Ross und Bower (1981) sowie Scholz, Ott und Müller-Sinik (1997).

Auf diesen Ansatz und damit verbundene Probleme werden ich in den nächsten beiden Abschnitten eingehen (vgl. auch Abschnitt 7.1.2). Da alle vorgestellten Lösungsansätze zu diesen Problemen recht aufwendig sind, werde ich anschließend noch ein gröberes Verfahren vorstellen: Die Analyse mutmaßlich homogener Teilstichproben, in denen Daten verschiedener, hinsichtlich bestimmter Kriterien „ähnlicher" Vpn zusammengefaßt werden.

5.4.1.1 Vp-spezifische Analyse: Bootstrap- und Jackknife-Verfahren

Eine triviale Voraussetzung für die Analyse individueller Datensätze ist, daß jede Vp Daten für alle relevanten Versuchsbedingungen liefert, die in dem betreffenden Modell berücksichtigt werden. Ist das nicht der Fall, so ist die Modellprüfung und im Regelfall auch die Parameterschätzung für eine einzelne Vp nicht durchführbar, da sie strukturell nicht die Datenmenge liefert, die für eine Modellprüfung bzw. Parameterschätzung erforderlich wäre. Leider ist es aber nicht immer möglich, Vpn unter allen erforderlichen Versuchsbedingungen zu untersuchen, weil sonst die interne Validität eines Experiments durch künstliche Abhängigkeiten der verschiedenen Beobachtungsresultate (z.B. als Folge des Strebens der Vpn, sich über verschiedene Bedingungen hinweg konsistent zu verhalten) gefährdet ist.[63] Liegt ein solcher Fall vor, gibt es prinzipiell keine Alternative zur Analyse interindividuell aggregierter Daten (Bredenkamp, 1972; 1980). Man muß sich dann den Ansätzen der Behandlung von Parameterheterogenitätsproblemen zuwenden, die in den Abschnitten 5.4.1.3ff besprochen werden.

Im folgenden sei vorausgesetzt, daß jede Vp eine strukturell ausreichende Datenmenge für individuelle Datenanalysen liefert. Das verbleibende Problem ist dann der typischerweise kleine oder bestenfalls mäßige Stichprobenumfang N_V pro

[63] Beispielsweise ist im Zusammenhang mit der Prozeßdissoziationsprozedur (Jacoby, 1991) diskutiert worden, ob es vertretbar ist, eine Vp sowohl unter Inklusions- als auch unter Exklusionsbedingungen zu beobachten (Buchner et al., 1995; Graf & Komatsu, 1994). Beides ist erforderlich, um Parameterschätzungen auf der Ebene einzelner Individuen ableiten zu können. Viele andere Beispiele lassen sich ergänzen.

Vp. Parameterschätzungen dürfen beispielsweise auch bei $N_v \approx 50$ nur mit Vorsicht interpretiert werden, da sie aufgrund des wahrscheinlich hohen Standardschätzfehlers (der mit Bootstrap-Verfahren für jede einzelne Vp ermittelt werden sollte, vgl. Abschnitt 4.2.5) recht unreliabel sind. Besondere Vorsicht ist geboten, wenn man Korrelationen zwischen Schätzungen verschiedener Parameter über die Vpn hinweg berechnet und inhaltlich interpretieren will. Eine signifikante Korrelation der Schätzer muß nicht unbedingt auf einer Korrelation der zugrundeliegenden Parameter beruhen. Schätzer können aufgrund von Eigenschaften der gemeinsamen Stichprobenverteilung eine Korrelation aufweisen, auch wenn die zugrundeliegenden Parameter unkorreliert sind. Ob die gemeinsame Stichprobenverteilung von Schätzern eine Korrelation aufweist, kann ebenfalls nur über Bootstrap-Verfahren ermittelt werden; von einer Schätzung der Varianz-Kovarianz-Matrix über die beobachtete Fisher-Informationsmatrix ist wegen des typischerweise kleinen Stichprobenumfangs pro Vp abzuraten.

Ein Test der globalen H_0, daß das zu prüfende Modell für jede einzelne Vp gilt, gegen die H_1, daß es für mindestens eine Vp verletzt ist, kann vom Ansatz her erreicht werden, indem (a) das Modell an die Daten jeder einzelnen Vp angepaßt wird, (b) eine individuelle Modellanpassungsstatistik für Vp v berechnet wird — z.B. die Log-Likelihood-Statistik G^2_v— und dann (c) die G^2_v-Werte über die Vpn hinweg zum Gesamt-G^2 aufsummiert werden. Die unter den üblichen Regularitätsbedingungen abgeleitete asymptotische Verteilung des Gesamt-G^2 unter H_0 ist eine χ^2-Verteilung mit $df = V \cdot df_v$, wobei df_v die Freiheitsgrade für den individuellen Modelltest sind.[64] Allerdings wird N_v selten so groß sein, daß man die Anwendung des asymptotischen χ^2-Tests rechtfertigen kann. Im Regelfall kommt daher wohl nur ein parametrisches Bootstrap-Verfahren in Frage. Hierzu generiert man wiederholt Monte-Carlo-Stichproben für alle V Vpn aus Grundgesamtheiten, die sich jeweils durch Einsetzen der individuellen ML-Schätzer in die Modellgleichungen ergeben. Man berechnet dann jeweils ein Bootstrap-Gesamt-G^2 für alle V Monte-Carlo-Stichproben und fährt so fort, bis man eine hinreichend genaue Schätzung erhält, ob das Gesamt-G^2 für die empirische Stichprobe zu den oberen $\alpha \cdot 100\%$ der Werte der Bootstrap-Verteilung der simulierten Gesamt-G^2-Werte gehört (H_0-Verwerfung) oder nicht (H_0-Beibehaltung). Ob bei dieser Bootstrap-Methode die Log-Likelihood-Quotienten-Statistik als Modellanpassungsstatistik op-

[64] Ist sowohl N_v als auch die Anzahl V der Versuchspersonen sehr groß, kommt prinzipiell auch die Anwendung eines anderen asymptotischen Tests in Frage. Jede einzelne G^2_v--Statistik ist unter H_0 asymptotisch $\chi^2_{(df_v)}$-verteilt, und die verschiedenen Statistiken sind unabhängig. Folglich kann man mit einem Kolmogoroff-Smirnoff-Anpassungstests prüfen, ob die Verteilung der V verschiedenen G^2_v--Statistiken einer $\chi^2_{(df_v)}$-Verteilung genügt (H_0) oder nicht (H_1) (vgl. Bredenkamp, 1984). Ob bzw. ab welchem V und N_v und für welche Punkt-Alternativhypothesen dieser Test Verwendung finden kann, sollte von systematischen Teststärkevergleichen mit dem im folgenden dargestellten parametrischen Bootstrap-Verfahren zum Gesamt-G^2 abhängig gemacht werden. Ergebnisse hierzu liegen noch nicht vor.

timal ist, muß vorab durch Simulationsstudien geklärt werden. Von Davier (1997) hat Simulationsstudien zu Latent-Class-Modellen vorgelegt, die vermuten lassen, daß bei kleinem N_v die Cressie-Read- oder die Pearson-Statistik in Verbindung mit dem parametrischen Bootstrap günstigere Eigenschaften haben als die G^2-Statistik. Obwohl dieser Befund nicht ohne weiteres auf die hier gegebene Situation übertragbar ist, deutet er zumindest an, daß man bei „sparse data" nicht blind auf die G^2-Statistik vertrauen sollte.

Ist der Stichprobenumfang pro Vp sogar kleiner als $N_v = 50$ — was bei kognitionspsychologischen Anwendungen häufig der Fall sein dürfte — muß bei der Berechnung individueller Parameterschätzwerte nicht nur mit einem hohen Standardschätzfehler, sondern darüber hinaus auch mit einem erheblichen mittleren Schätzfehler (Bias) gerechnet werden. Dann ist es offenkundig nicht sinnvoll, die individuellen Schätzwerte zu mitteln, um einen Schätzer für den mittleren Parameter in der betreffenden Versuchsbedingung zu erhalten: Auch dieser Schätzer hätte nämlich einen starken mittleren Bias. Es kann in solchen Fällen deutlich günstiger sein, statt gemittelter individueller Schätzwerte den Schätzer für die aggregierten Daten zu verwenden, selbst wenn dieser Schätzer auf inadäquaten Modellannahmen beruht (Macmillan & Kaplan, 1985). Ein bessere Alternative haben Dorfman und Berbaum (1986) am Beispiel der Schätzung mittlerer d' Parameter aus Konfidenzratings verschiedener Vpn illustriert. Sie haben die auf Quenouille (1949) und Tukey (z.B. 1969) zurückgehende Jackknife-Methode der Biasreduktion verwendet (Efron, 1982; Shao & Tu, 1995). Konkret sieht das Vorgehen so aus, daß ein interessierender Modellparameter θ zunächst aus den aggregierten Daten aller Vpn geschätzt wird. Das Ergebnis sei $\hat{\theta}_a$. Dann eliminiert man die Daten der v-ten Vp ($v = 1, ..., V$) aus dem Aggregat und schätzt für die Daten der restlichen $V-1$ Vpn ebenfalls den entsprechenden Parameter. Das Ergebnis sei $\hat{\theta}_v$. Entsprechend verfährt man für alle V Vpn. Anschließend werden die sogenannten „Pseudowerte"

$$\hat{\theta}^*_v = V \cdot \hat{\theta}_a - (V-1) \cdot \hat{\theta}_v, \quad v = 1, \ldots, V \tag{5.4.1.1-1}$$

für jede Vp berechnet und über alle Vpn gemittelt. Der Mittelwert der Pseudowerte

$$\hat{\theta}_J = \sum_{v=1}^{V} \frac{\hat{\theta}^*_v}{V} \tag{5.4.1.1-2}$$

ist der Jackknife-Schätzer des interessierenden Parameters θ für das Aggregat. Viele Untersuchungen haben die hervorragenden Eigenschaften von Jackknife-Schätzern im Vergleich zu gemittelten individuellen Schätzern aufgezeigt (Dorfman & Berbaum, 1986; Rey, 1978). Jackknife-Schätzer weisen typischerweise geringeren

Bias und größere Robustheit auf (Rey, 1978). Eine systematische Evaluation von Jackknife-Schätzern im Kontext von MVB-Modellen steht allerdings noch aus.

5.4.1.2 Vp-spezifische Analyse: Bayessche Verfahren

Bootstrap- und Jackknife-Verfahren helfen dabei, Modelltests zu konstruieren, die auch für kleine Stichproben das α-Niveau einhalten, und Parameterschätzungen abzuleiten, die auch bei kleinem N_V noch annähernd erwartungstreu sind. Ein Problem, das beide Verfahrensklassen jedoch nicht lösen können, sind die exzessiv großen Standardschätzfehler und Konfidenzintervalle, die mit Analysen individueller Datensätze fast zwangsläufig verbunden sind. Kann auf Vorwissen bezüglich der zu erwartenden Parameterwerte einer Vp zurückgegriffen werden, so können Bayessche Verfahren der Parameterschätzung hier weiterhelfen (vgl. Batchelder, 1998; Chechile & Meyer, 1976; zur Einführung: Molenaar & Lewis, 1996). Ist beispielsweise bekannt, in welchem Bereich die Parameterwerte einer Referenzgruppe, der eine Vp entstammt, typischerweise liegen, so kann man die entsprechende Verteilung des Parameters in der Referenzgruppe als Apriori-Verteilung verwenden. Die empirischen Daten einer Vp dienen dann nur noch dem Zweck, eine versuchspersonenspezifische Korrektur der Apriori-Verteilung zu erhalten. Batchelder (1998) schlägt vor, bei der Analyse binärer MVB-Modelle unabhängige Beta-Verteilungen der Parameter als Apriori-Verteilungen anzunehmen. Damit sind mehrere Vorteile verbunden:

1) Beta-Verteilungen sind sehr flexibel und erlauben die Beschreibung einer Vielzahl unterschiedlicher Dichten im Intervall [0, 1]. Alle in der Praxis interessierenden Vorwissensverteilungen dürften sich damit abbilden lassen (vgl. Abschnitt 5.1).
2) Die Parameter der Beta-Verteilungen lassen sich sehr leicht schätzen (Johnson & Kotz, 1970b, Kap. 24). Wenn man etwa ein bestimmtes multinomiales Modell mit den oben geschilderten Methoden an die Daten einzelner Vpn aus einer Referenzpopulation angepaßt hat, erhält man Verteilungen von Schätzwerten pro Parameter, die teilweise genuine Parametervariabilität in der Population, teilweise auch Stichprobenfehler reflektieren. Diese empirisch ermittelten Verteilungen können als Apriori-Verteilungen („best guess”) zur Schätzung der Parameter für eine weitere Vp aus der gleichen Population eingesetzt werden. Unter der (plausiblen) Voraussetzung, daß sie sich angemessen durch Beta-Verteilung beschreiben lassen, berechnet man einfach den Mittelwert und die (biaskorrigierte) Streuung der vorliegenden Schätzwerte und setzt diese in Gleichung (5.3-1) für den Erwartungswert und Streuung einer beta-verteilten ZV ein. Durch Auflösung der

Gleichungen nach p und q erhält man dann Momentschätzer der beiden Parameter der Beta-Verteilung (Batchelder, 1998, vgl. auch Abschnitt 4.2.3).

3) Der größte Vorteil der Beta-Verteilung besteht darin, daß bei einer Beta-Apriori-Verteilung auch die Aposteriori-Verteilung zu gegebenen Daten einer Vp eine Beta-Verteilung sein muß, wenn ein bestimmtes MVB-Modell zutrifft. Das macht die Implementierung des Bayesschen Verfahrens in den EM-Algorithmus der Parameterschätzung für MVB-Modelle sehr einfach (vgl. Batchelder, 1998).

Ein geeignetes Bayessches Pendant zu Punktschätzern ist der Mittelwert der Aposteriori-Verteilung des betreffenden MVB-Parameters. Dieser ergibt sich bei Verwendung der Beta-Verteilung erfreulicherweise einfach als gewichtetes arithmetisches Mittel des gewöhnlichen ML-Schätzers auf der Basis der vorliegenden Daten einerseits und des Mittelwerts der Apriori-Verteilung andererseits. Das Gewicht des Apriori-Mittelwerts sinkt mit der Größe der Datenstichprobe, die für die betreffende Vp beobachtet wurde. Je größer also die vorliegende empirische Evidenz für die Parameterschätzung ist, desto geringer wird der Einfluß der zugrunde gelegten Vorwissensverteilung.

Bayessche Verfahren der Parameterschätzung liefern folglich Ergebnisse, die als Kompromiß zwischen den Vorwissenserwartungen und den konventionellen Schätzwerten für eine Datenstichprobe betrachtet werden können, wobei das Gewicht der Daten vernünftigerweise vom Stichprobenumfang abhängig gemacht wird. Der Gewinn Bayesscher Verfahren relativ zu konventionellen Punktschätzungen der Parameter wird deutlich, wenn man die Streuung der Aposteriori-Verteilung betrachtet: Diese ist deutlich kleiner als der Standardschätzfehler konventioneller ML-Schätzer (Batchelder, 1998). Auch die Generalisierung des Ansatzes auf allgemeine MVB-Modelle mit nichtbinären Parametern bereitet prinzipiell keine Probleme: Für Parametergruppen, die mehr als einen freien Parameter umfassen, wählt man dann zweckmäßigerweise eine Dirichlet-Verteilung als Apriori-Verteilung (zur Dirichlet-Verteilung vgl. Abschnitt 5.4.2.2).

Die Nachteile Bayesscher Verfahren liegen ebenfalls auf der Hand: Ein Gewinn an Effizienz durch Vorwissensnutzung kann trügerisch sein, wenn die zugrunde gelegte Apriori-Verteilung des Parameters nicht gut begründet ist. Falsch bzw. willkürlich gewählte Apriori-Verteilungen können das Bild der Parameterschätzung grob verzerren. „Neutrale” oder „uninformative” Vorwissensverteilungen — z.B. die Gleichverteilung im Intervall [0, 1] bzw. die Beta-Verteilung mit $p = q = 1$ — führen zwar nicht zu groben Verzerrungen, ändern auf der anderen Seite aber auch kaum etwas an den Ergebnissen, die sich bei Verwendung konventioneller ML-Schätzer ergeben würden: Der Mittelwert der Aposteriori-Verteilung wird unter diesen Umständen sehr nahe am ML-Schätzer und deren Streuung sehr nahe beim Standardschätzfehler liegen (Batchelder, 1998). Als Fazit muß da-

her festgehalten werden, daß Bayessche Verfahren der Parameterschätzung nur dann empfohlen werden können, wenn (a) auf zahlreiche vorherige Parameterschätzungen zum gleichen Modell zurückgegriffen werden kann, die (b) als Zufallsziehungen aus der gleichen Referenzpopulation aufgefaßt werden können. Die Parameterschätzungen müssen zudem mit Verfahren durchgeführt worden sein, die auch bei kleinen Datenstichproben möglichst biasfrei sind (Bootstrap- oder Jackknife-Schätzer).

5.4.1.3 Analyse von Teilstichproben basierend auf mehreren Vpn

Alle bislang geschilderten Analysemethoden für parameterheterogene Datensätze sind ausgesprochen aufwendig und rechenintensiv. Somit sind sie für Routineanwendungen kaum geeignet. Einfacher anwendbare approximative Verfahren sind daher wünschenswert, die trotz ihrer Einfachheit möglicher Parameterheterogenität Rechnung tragen. Darüber hinaus ist manchmal die Analyse der Daten auf der Ebene einzelner Vpn aus Gründen der Validitätskontrolle nicht möglich (vgl. Abschnitt 5.4.1.2). Auch dies erzwingt den Übergang zu alternativen Verfahren der Parametervariabilitätskontrolle.

Ein Ansatz hierzu ist folgender: Hat man a priori Hypothesen darüber, wodurch die Heterogenität in einem Aggregatdatensatz hervorgerufen wird, so kann man das Variabilitätsproblem häufig durch Aufsplittung des Gesamtdatensatzes in wenige Teildatensätze lösen, ohne auf die Ebene individueller Datensätze zurückzugehen. Erdfelder und Bredenkamp (1998, Experiment 1) haben beispielsweise Versuchsbedingungen und Itemmaterial über vier Vpn-Gruppen hinweg ausbalanciert, so daß in jeder Versuchsbedingung jedes Item gleich häufig vertreten war. Denkbar ist, daß das unterschiedliche Itemmaterial innerhalb einer Versuchsbedingung Parameterheterogenität erzeugt. Erdfelder und Bredenkamp haben deshalb die Daten der vier Ausbalancierungsgruppen getrennt analysiert und die Ergebnisse mit denen zum Aggregat aller vier Gruppen verglichen. Es ergaben sich keine bedeutsamen Unterschiede, so daß Parameterheterogenität als potentielle Verfälschungsquelle unwahrscheinlich ist. Komplett ausgeschlossen sind Artefakte allerdings nicht, denn auch die Ergebnisse zu den vier Ausbalancierungsgruppen können durch heterogene Vpn oder Items innerhalb der Ausbalancierungsgruppen verfälscht sein.

In ganz ähnlicher Weise haben Riefer, Hu und Batchelder (1994) die Quellendiskriminationsdaten zu ihrem Experiment 1 doppelt ausgewertet: Einmal für das Aggregat aller Vpn und einmal getrennt für die 11 Vpn mit den besten und die 11 Vpn mit den schlechtesten Gedächtnistestergebnissen. Da die Ergebnisse — von Unterschieden in den Antworttendenzparametern abgesehen — ähnlich ausfielen, konnte Heterogenität zwischen „guten" und „schlechten" Vpn als Artefaktquelle

weitgehend ausgeschlossen werden. Eine absolute Garantie, daß Parametervariabilität keine Verzerrungen bewirkt hat, hat man allerdings in dieser Studie ebensowenig wie im Falle der zuvor geschilderten Arbeit von Erdfelder und Bredenkamp (1998).

Die Aufsplittung des Gesamtdatensatzes in mutmaßlich homogene Teildatensätze ist die Minimum-Vorsichtsmaßnahme, die man beim Verdacht heterogener Stichproben einleiten sollte. Sie ist relativ leicht durchführbar und macht kaum zusätzlichen Aufwand bei der Datenanalyse, so daß man sie durchaus routinemäßig einsetzen kann.

5.4.2 Modelle mit variablen Parametern

Eleganter und auch effizienter als die Analyse von Teildatensätzen oder individuellen Datensätzen ist der Übergang zu sophistizierteren Modellen, die Heterogenität zwischen Items und/oder Vpn explizit in das stochastische Modell integrieren. Drei Möglichkeiten der Verallgemeinerung von MVB-Modellen auf Modelle mit heterogenen Parametern sollen im folgenden kurz vorgestellt werden: Latent-Class-MVB-Modelle, Beta- und Dirichlet-MVB-Modelle sowie Rasch-MVB-Modelle. Bei allen drei Modellen handelt es sich um spezielle Mischverteilungsmodelle (vgl. Rost & Erdfelder, 1996; Sixtl, 1993).

5.4.2.1 Latent-Class-MVB-Modelle

Latent-Class-MVB-Modelle beruhen auf der Idee, daß die vorliegende Stichprobe einer Population entnommen wurde, die in zwei oder mehr Subpopulationen („latente Klassen") zerfällt. Innerhalb jeder Subpopulation l ($l = 1, ..., L$) gilt ein bestimmtes Standard-MVB-Modell, jedoch mit einem klassenspezifischem Parametervektor $\boldsymbol{\theta}_l$. Die Parametervariabilität ist demnach kategorialer bzw. diskreter Natur: Sie besteht *zwischen* den latenten Klassen, die der betreffenden Aggregatstichprobe zugrunde liegen, nicht jedoch *innerhalb* der latenten Klassen. Hätte man ein beobachtbares empirisches Kriterium für die Klassenzugehörigkeit eines Stichprobenelements, so könnte man die Gesamtstichprobe also in homogene Teilstichproben zerlegen und diese getrennt analysieren — wie im letzten Abschnitt empfohlen. Der Vorteil eines Latent-Class-MVB-Modells besteht darin, daß der Heterogenität der Gesamtstichprobe auch dann Rechnung getragen werden kann, wenn ein Kriterium für die Klassenzugehörigkeit eines Stichprobenelements nicht vorhanden ist — wenn die Klassen also „latent" sind.

Wie man sich leicht klarmachen kann, ist ein Latent-Class-MVB-Modell selbst wieder ein MVB-Modell: Man kann die Aufsplittung der Gesamtpopulation in L latente Klassen einfach als erste Ebene in einem Verarbeitungsbaumdiagramm auffassen. Die latenten Ereignisse auf der ersten Ebene sind also die laten-

ten Klassen, und die Wahrscheinlichkeiten dieser Ereignisse sind die latenten Klassenwahrscheinlichkeiten p_l. Jedem Ereignis auf der ersten Ebene folgt dann das gleiche Teilbaummodell, nämlich das MVB-Modell für die betreffende latente Klasse.

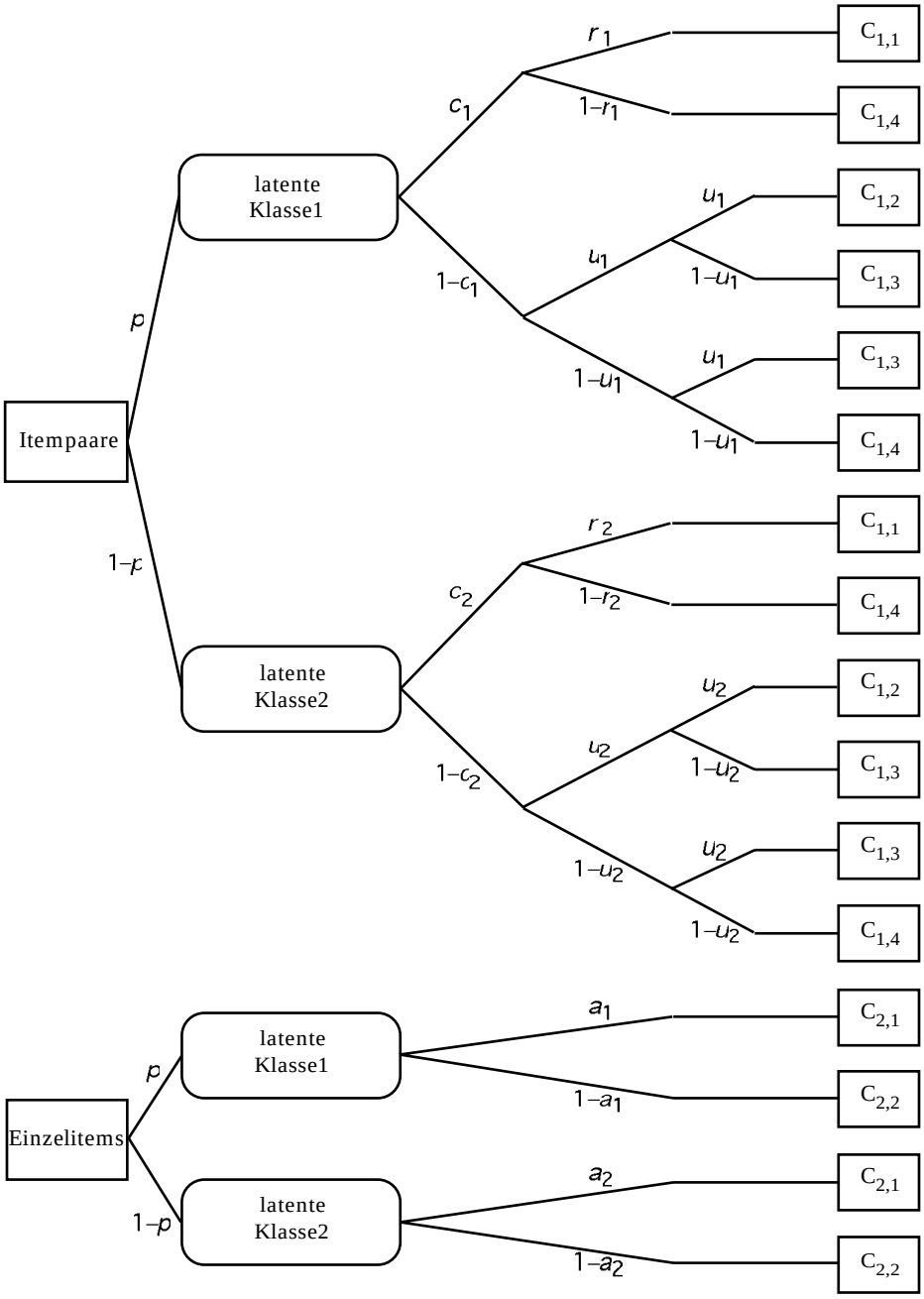

Abbildung 5.5:
Eine Verallgemeinerung des Enkodierung-Abruf-Modells von Batchelder und Riefer (1980; 1986) auf ein Modell mit zwei latenten Klassen.

Ein MVB-Modell mit Zufallsparametern wird also bei der Latent-Class-MVB-Modellierung durch einen einfachen Trick in ein Standard-MVB-Modell mit fixen Parametern transformiert: Das MVB-Modell mit fixen Parametern wird jeweils für Teilgruppen der Gesamtstichprobe formuliert, bis alle Parameterunter-

schiede in der Gesamtstichprobe durch Teilstichprobenmodelle mit teilgruppenspezifischen Parametern abgebildet werden können.

Das einfachste Latent-Class-MVB-Modell zum Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986) ist das Zwei-Klassen-Modell. Die Veranschaulichung dieses Modells in Form eines binären Verarbeitungsbaums (unter Fortlassung der Prozeßbeschreibungen an den Knoten des Modells) illustriert Abbildung 5.5. Offensichtlich ist dieses Modell auch dann überparametrisiert, wenn man die Gleichsetzung der Parameter $u_l = a_l$ innerhalb jeder Klasse vornimmt. Wir haben dann immer noch drei Modellparameter c_l, r_l und u_l innerhalb jeder Klasse (insgesamt also 6 Parameter) und zusätzlich noch eine freie latente Klassenwahrscheinlichkeit p. Insgesamt sind das 7 Parameter bei nur (4–1) + (2–1) = 4 funktional unabhängigen Kategoriewahrscheinlichkeiten.

Wie läßt sich das hieraus resultierende Identifizierbarkeitsproblem lösen? Eine Möglichkeit ist die Erweiterung des Modells auf einen dreifaktoriellen Versuchsplan, in dem die Parameter c_l, r_l und u_l jeweils unabhängig voneinander in zwei Stufen manipuliert werden (vgl. Abschnitt 4.6). Nehmen wir etwa an, daß acht Versuchsgruppen eine Liste von Wörtern sequentiell in einer Zufallsreihenfolge dargeboten bekommen. Die Wortliste bestehe teilweise aus semantisch zusammengehörigen Wörtern (z.B. Geige — Klavier, Stiefmütterchen — Tulpe etc.) und teilweise aus Einzelwörtern. Ein späterer Gedächtnistest wird vor und während der Lernphase nicht angekündigt (inzidentelle Lernbedingung), so daß die Art der Verarbeitung der Wörter wesentlich von der jeweiligen Versuchsbedingung abhängt. Die acht Versuchsbedingungen entstehen aus der kompletten Kreuzklassifikation der zweistufigen unabhängigen Variablen A (Angenehmheitsrating des Items in der Enkodierungsphase: kein Rating vs. Rating), B (Abrufhilfe in der Abrufphase: Oberkategorien der Zielwörter beim Reproduktionstest nicht präsent vs. präsent) und C (semantische Klassifizierungsaufgabe in der Enkodierungsphase: Keine Klassifizierung vs. Klassifizierung). Die UV A sollte die Intensität itemspezifischer Verarbeitungsprozesse, die UV C dagegen semantisch-relationale Verarbeitungsprozesse in der Enkodierungsphase affizieren (Einstein & Hunt, 1980; Engelkamp, Biegelmann & McDaniel, 1998; Hodge & Otani, 1996; Hunt & Einstein, 1981). Folglich ist zu erwarten, daß die UV A selektiv den u-Parameter für nicht geclusterte Wörter und die UV C ebenfalls selektiv den Cluster-Enkodierungsparameter c des EA-Modells beeinflußt. Die Präsentation von Abrufhilfen (UV B) sollte aus naheliegenden Gründen den Abruf geclusterter Wörter fördern und damit den r-Parameter des Modells erhöhen (Baddeley, 1997). Ein zusätzlicher Effekt dieser Maßnahme auf den u-Parameter kann beispielsweise dadurch verhindert werden, daß nur Oberkategorien für diejenigen Wörter der Liste präsentiert werden, die zuvor als Wortpaare präsentiert wurden.

Tabelle 5.2:
Parameter eines Enkodierung-Abruf-Modells mit zwei latenten Klassen für die simultane Auswertung eines Versuchsplans mit acht Versuchsgruppen (VG 1 bis VG 8), die aus der Kreuzklassifikation von drei unabhängigen Variablen (UV A, UV B und UV C) resultieren. Der erste Index der *c*-, *r*- und *u*-Parameter gibt jeweils die latente Klasse (1 oder 2), der zweite Index die Ausprägung des Parameters (schwach = 0, stark = 1) an. Der Parameter *p* für die Größe der ersten latenten Klasse ist in der Tabelle nicht aufgeführt.

UV B:	UV C:	UV A: itemspezifische Verarbeitungsprozesse	
Relation	Abruf	schwach	stark
schwach	schwach	VG 1: Klasse 1 / Klasse 2 $c_{1,0}, r_{1,0}, u_{1,0}$ / $c_{2,0}, r_{2,0}, u_{2,0}$	VG 2: Klasse 1 / Klasse 2 $c_{1,0}, r_{1,0}, u_{1,1}$ / $c_{2,0}, r_{2,0}, u_{2,1}$
schwach	stark	VG 3: Klasse 1 / Klasse 2 $c_{1,0}, r_{1,1}, u_{1,0}$ / $c_{2,0}, r_{2,1}, u_{2,0}$	VG 4: Klasse 1 / Klasse 2 $c_{1,0}, r_{1,1}, u_{1,1}$ / $c_{2,0}, r_{2,1}, u_{2,1}$
stark	schwach	VG 5: Klasse 1 / Klasse 2 $c_{1,1}, r_{1,0}, u_{1,0}$ / $c_{2,1}, r_{2,0}, u_{2,0}$	VG 6: Klasse 1 / Klasse 2 $c_{1,1}, r_{1,0}, u_{1,1}$ / $c_{2,1}, r_{2,0}, u_{2,1}$
stark	stark	VG 7: Klasse 1 / Klasse 2 $c_{1,1}, r_{1,1}, u_{1,0}$ / $c_{2,1}, r_{2,1}, u_{2,0}$	VG 8: Klasse 1 / Klasse 2 $c_{1,1}, r_{1,1}, u_{1,1}$ / $c_{2,1}, r_{2,1}, u_{2,1}$

Legt man die geschilderten Hypothesen für die simultane Auswertung der Kategorienhäufigkeiten für alle acht Versuchsbedingungen zugrunde, so werden bei Zugrundelegung eines EA-Modells mit zwei latenten Klassen nur 6 Modellparameter pro Klasse (jeweils zwei *u*- , *c*- und *r*-Parameter für die Bedingungen mit schwacher versus starker Ausprägung des betreffenden Parameters) und zusätzlich ein Parameter *p* für die Wahrscheinlichkeit der ersten latenten Klasse benötigt. Tabelle 5.2 veranschaulicht das Parameter-Restriktionsmuster über die Versuchsgruppen hinweg, das aus den o.g. Hypothesen folgt. Der Klassengrößenparameter *p* ist in der Tabelle 5.2 nicht berücksichtigt. Da für jede Versuchsbedingung vier Beobachtungshäufigkeiten für Wortpaare und zwei für Einzelwörter erhoben werden, stehen den $S = 13$ Modellparametern $M = 8 \cdot ((4–1) + (2–1)) = 32$ funktional unabhängige Kategoriewahrscheinlichkeiten gegenüber. Das Modell ist unterparametrisiert und lokal identifizierbar. Die globale Identifizierbarkeit scheitert — wie bei allen Latent-Class-Modellen — daran, daß sich durch Vertauschung der Klassenindizes in Tabelle 5.2 ein datenäquivalentes Modell ergibt. Die Etikettierung der Klassen als ‚Klasse 1" und „Klasse 2" ist also beliebig. Man könnte die globale Identifizierbarkeit des Modells — wie in Abschnitt 4.3.2.1 erläutert — durch eine Ordnungsrestriktion zu den Klassengrößenparametern erzwingen. Dies ist jedoch nicht unbedingt nötig, da lokale Identifizierbarkeit für die BAN-Parameterschätzung und die statistische Rechtfertigung des asymptotischen Modelltests ausreicht. Jede Minimum-PD$^{\lambda}$-Modellanpassungsstatistik zu diesem Modell und diesem Versuchsdesign ist unter den üblichen Regularitätsbedingungen bei Modellgültigkeit $\chi^2_{(df)}$-verteilt mit $df = M – S = 32 – 13 = 19$ Freiheitsgraden. Auch ein EA-

Modell mit drei latenten Klassen wäre für diesen Versuchsplan noch lokal identifizierbar und testbar (bei $df = 12$).

Tabelle 5.3:
Ergebnisse der Parameterschätzung zum Zwei-Klassen-Enkodierung-Abruf-Modell, das auf einen nahezu idealen Datensatz zu dem in Tabelle 5.2 dargestellten 8-Gruppen-Experiment angewandt wird. Die Daten sind in der Datei „LCAEABSP.MDT" im vierten Abschnitt des Anhangs wiedergegeben.

Klasse:	Latente Klasse 1 ($\hat{p} = .579$)		Latente Klasse 2 ($1-\hat{p} = .421$)	
Ausprägung:	schwach (0)	stark (1)	schwach (0)	stark (1)
$\hat{c}$	.088	.384	.496	.795
$\hat{r}$	.093	.418	.487	.774
$\hat{u}$	.096	.392	.486	.791

Kann das Zwei-Klassen-EA-Modell in Anwendungskontexten nützlich sein? Diese Frage ist unbedingt zu bejahen. Im dritten Abschnitt des Anhangs ist zur Veranschaulichung ein künstlich konstruierter Datensatz aufgeführt (Datei „LCAEABSP.MDT"), der nahezu ideal zu dem geschilderten Zwei-Klassen-EA-Modell paßt. Die Anpassungsstatistik zum oben dargestellten 8-Gruppen-Modell nimmt für diesen Datensatz den fast perfekten Wert $G^2_{(19)} = 0.03$ an, und die Parameterschätzungen, die in Tabelle 5.3 wiedergegeben sind, passen sehr gut zu den oben dargestellten Prognosen: Jede der drei UVn hat die erwarteten Effekte, und zwar innerhalb jeder der beiden latenten Klassen. Insgesamt ist die latente Klasse 1 eher durch schwächere Gedächtnisleistungen bei der Cluster-Enkodierung ($\hat{c}$), dem Cluster-Abruf ($\hat{r}$) und dem Einzelitemgedächtnis ($\hat{u}$) gekennzeichnet, die latente Klasse 2 dagegen eher durch bessere Gedächtnisleistungen. In der Gesamtpopulation und der daraus gewonnenen Aggregatstichprobe sind die Parameter innerhalb der Äste des EA-Verarbeitungsbaummodells folglich stark korreliert: Hohe Leistungen bei der Cluster-Enkodierung gehen mit hohen Abrufleistungen einerseits und gutem Einzelitemgedächtnis andererseits einher. Nach den Ergebnissen des letzten Abschnitts ist aufgrund dieser Korrelationen zu erwarten, daß das einfache EA-Modell, welches konstante Parameter über alle Elemente der Stichprobe annimmt, nicht gut auf die Daten passen kann. Tatsächlich ergibt sich eine inakzeptabel hohe Anpassungsstatistik $G^2_{(26)} = 47.95$ für das einfache Acht-Gruppen-EA-Modell, wenn man die Parameter über die acht Gruppen genau so restringiert, wie das beim Zwei-Klassen-EA-Modell geschah (vgl. Tabelle 5.1). Selbst wenn man keinerlei Parameterrestriktion bis auf die Gleichsetzung $u = a$ innerhalb der Gruppen vornimmt, resultiert immer noch eine recht hohe (auf dem Niveau $\alpha = .10$ signifikante) Anpassungsstatistik von $G^2_{(8)} = 15.23$. Darüber hinaus folgen die Parameterschätzungen zu diesem Modell nicht mehr dem prognostizierten Muster

über die Versuchsbedingungen hinweg, was die psychologische Validität des Modells fragwürdig erscheinen läßt.

Dieses fiktive Beispiel zeigt, daß man bei der Analyse aggregierter Daten mit Standard-MVB-Modellen gehörig in die Irre geführt werden kann, wenn man blind auf die Robustheit der statistischen Analysen gegenüber Parameterheterogenität vertraut. Im Beispielfall würde man für den Acht-Gruppen-Versuchsplan das EA-Modell verwerfen oder zumindest die psychologische Validität der Modellparameter anzweifeln. Die gleichen Beispieldaten können dagegen perfekt und hypothesengemäß mit einem Latent-Class-EA-Modell beschrieben werden, welches unterschiedliche EA-Modellparameter in zwei latenten Klassen annimmt, die über die Klassen hinweg korreliert sind. Die Ursache für das Scheitern des EA-Modells besteht als ausschließlich in der Parameterheterogenität (in Verbindung mit Parameterkorrelationen), nicht etwa in grundlegenderen psychologischen Fehlannahmen des EA-Modellansatzes.

Latent-Class-MVB-Modelle können offensichtlich sehr hilfreich sein. Sie sind zudem auch noch statistisch leicht analysierbar, da es sich um Spezialfälle der MVB-Modellfamilie handelt. Soweit vorliegende Daten bzw. Versuchspläne es zulassen, sollten man daher im Falle des Scheiterns eines Standard-MVB-Modells auf jeden Fall Latent-Class-MVB-Modelle als Alternative ausprobieren. Dies ist die einfachste Möglichkeit, um herauszufinden, ob das Goodness-of-fit- oder Validitätsproblem lediglich auf Parametervariabilität in Verbindung mit Parameterkorrelationen zurückzuführen ist oder ob fundamentalere Schwächen im Modellansatz vorliegen.

5.4.2.2 Beta-MVB-Modelle und Dirichlet-MVB-Modelle

Latent-Class-MVB-Modelle basieren auf der Annahme, daß die Parameterheterogenität diskreter Natur ist: Jeder MVB-Modellparameter nimmt in der zugrundeliegenden Population nur wenige (z.B. zwei oder drei) mögliche Werte an. Diese Einschränkung mag in vielen Anwendungskontexten unplausibel erscheinen. Daher liegt die Frage nahe, ob sich ein zu Latent-Class-MVB-Modellen analoger Ansatz auch für den Fall stetig verteilter Zufallsparameter ableiten läßt.

Prinzipiell ist dies natürlich möglich. Das entsprechende Mischverteilungsmodell läßt sich in allgemeiner Form recht leicht formulieren (vgl. Rost & Erdfelder, 1996, Gleichung (3)). Nehmen wir an, wir haben es mit einem binären MVB-Modell (BVB-Modell) mit S Zufallsparametern zu tun. Die Zufallsparameter Θ_1, ..., Θ_S mit Werten θ_1, ..., θ_S seien stetig verteilt mit gemeinsamer Dichtefunktion $f(\theta_1, ..., \theta_S; \boldsymbol{\kappa})$, die von einem (mehrdimensionalen) Parameter $\boldsymbol{\kappa}$ abhängt. Nehmen wir weiter an, für jede Ausprägung der Zufallsparameter gelte ein bestimmtes BVB-Modell, das der Kategorie $C_{k,j}$ die Wahrscheinlichkeit $p_{k,j}(\theta_1, ..., \theta_S)$ zuweist, wel-

che gemäß Gleichung (3.5-1) von den Parameterwerten abhängt. Die „lokale" Wahrscheinlichkeitsfunktion für gegebene Werte $\theta_1, ..., \theta_S$ folgt also einer verbundenen Multinomialverteilung gemäß Gleichung (3.2-3). Dann ist die Wahrscheinlichkeitsfunktion der Kategoriehäufigkeiten $\boldsymbol{y}$ in der Mischpopulation, die der Aggregatstichprobe zugrunde liegt, gegeben durch

$$Pr(\boldsymbol{y};\boldsymbol{\kappa}) = \int_{\theta_1=0}^{1} \ldots \int_{\theta_S=0}^{1} Pr(\boldsymbol{y};\theta_1, \ldots,\theta_S)\cdot f(\theta_1, \ldots,\theta_S;\boldsymbol{\kappa})\, d\theta_S \ldots d\theta_1 =$$
$$= \int_{\theta_1=0}^{1} \ldots \int_{\theta_S=0}^{1} \prod_{k=1}^{K} N_k! \prod_{j=1}^{J_k} \frac{p_{k,j}(\theta_1, \ldots,\theta_S)^{y_{k,j}}}{y_{k,j}!} \cdot f(\theta_1, \ldots,\theta_S;\boldsymbol{\kappa})\, d\theta_S \ldots d\theta_1 . \qquad (5.4.1.2\text{-}1)$$

Die Schwierigkeit besteht allein darin, eine geeignete Wahl für die Dichtefunktion $f(\theta_1, ..., \theta_S; \boldsymbol{\kappa})$ zu treffen. „Geeignet" bedeutet hierbei, daß die Dichtefunktion einerseits psychologisch plausibel und andererseits analytisch leicht handhabbar ist, so daß die Integrale in Gleichung (5.4.1.2-1) problemlos bestimmt werden können. Schließlich sollte der Parametervektor $\boldsymbol{\kappa}$ möglichst auch noch von geringer Dimensionalität sein, damit eine gewisse Chance besteht, daß das Gesamtmodell für einen überschaubaren Versuchsplan – wie er etwa im letzten Abschnitt exemplarisch vorgestellt wurde – identifizierbar und testbar ist.

Für BVB-Modelle mit Zufallsparametern wurde bereits die Beta-Verteilung als naheliegende Apriori-Verteilung vorgestellt. Nimmt man Unabhängigkeit der S Zufallsparameter an, ergibt sich aus diesem Ansatz die Spezifikation (vgl. Gleichung 5.1-2):

$$f(\theta_1, \ldots,\theta_S;\boldsymbol{\kappa} = (p_1, q_1, \ldots, p_S, q_S)) = \prod_{s=1}^{S} \frac{\theta_s^{(p_s-1)}\cdot(1-\theta_s)^{(q_s-1)}}{\int_{t=0}^{1} t^{(p_s-1)}(1-t)^{(q_s-1)}\,dt} \qquad (5.4.1.2\text{-}2)$$

Da jede der S univariaten Beta-Verteilungen $B(p_j, q_j)$ durch nur zwei Parameter p_j und q_j gekennzeichnet ist, ergibt sich durch den Übergang zu einem Mischverteilungsmodell lediglich eine Verdoppelung der Parameterzahl, ähnlich wie etwa bei Zwei-Klassen-MVB-Modellen. Dies ist erfreulich.

Wie ist bei allgemeinen MVB-Modellen zu verfahren, bei denen auch nichtbinäre Parametergruppen vorkommen können? Hierfür bieten sich sogenannte Dirichlet-Verteilungen als Verteilungen der Zufallsparameter an (vgl. Johnson & Kotz, 1972, Kapitel 40, Abschnitt 5). Dirichlet-Verteilungen können als multivariate Verallgemeinerung von Beta-Verteilungen aufgefaßt werden: Sind $\Theta_1, ..., \Theta_i, ..., \Theta_m$ positiv-reellwertige ZVn mit Werten $\theta_1, ..., \theta_i, ..., \theta_m$, deren Summe niemals größer als 1 wird, so folgenden sie einer gemeinsamen Dirichlet-Verteilung, wenn sie die Dichtefunktion

$$f(\theta_1, \ldots, \theta_m; \kappa_0, \ldots, \kappa_m) = \frac{\Gamma\left(\sum_{j=0}^{m} \kappa_j\right)}{\prod_{j=0}^{m} \Gamma\left(\kappa_j\right)} \left(1 - \sum_{j=1}^{m} \theta_j\right)^{(\kappa_0 - 1)} \prod_{j=1}^{m} \theta_j^{(\kappa_j - 1)} \quad (5.4.1.2\text{-}3)$$

aufweisen. Hierbei bezeichnet Γ(.) die Gamma-Funktion (vgl. Johnson & Kotz, 1970a, Kapitel 17)

$$\Gamma(x) := \int_{t=0}^{\infty} e^{-t} \cdot t^{(x-1)} dt \,. \quad (5.4.1.2\text{-}4)$$

Für m = 1 reduziert sich die Dichtefunktion — wie man durch Vergleich mit Gleichung (5.1-2) leicht zeigen kann — auf eine einfache Beta-Verteilung $B(\kappa_1, \kappa_0)$. Die Momente einer Dirichlet-Verteilung hängen von den m+1 Parametern $\kappa_0, \kappa_1, \ldots, \kappa_m$ wie folgt ab (Johnson & Kotz, 1972, S. 233f.): Jede einzelne der ZVn Θ_i einer Dirichlet-Verteilung, i = 1, ..., m, ist beta-verteilt mit Erwartungswert

$$E(\Theta_i) = \frac{\kappa_i}{\sum_{j=0}^{m} \kappa_j} \quad (5.4.1.2\text{-}5)$$

und Varianz

$$Var(\Theta_i) = \frac{\kappa_i \cdot (\sum_{j=0}^{m} \kappa_j - \kappa_i)}{\left(\sum_{j=0}^{m} \kappa_j\right)^2 \left(\sum_{j=0}^{m} \kappa_j + 1\right)} \,. \quad (5.4.1.2\text{-}6)$$

Zwei ZVn einer gemeinsamen Dirichlet-Verteilung haben die Korrelation

$$\rho(\Theta_i, \Theta_{i'}) = -\sqrt{\frac{\kappa_i \cdot \kappa_{i'}}{\left(\sum_{j=0}^{m} \kappa_j - \kappa_i\right) \cdot \left(\sum_{j=0}^{m} \kappa_j - \kappa_{i'}\right)}} \,. \quad (5.4.1.2\text{-}7)$$

Auch die Dirichlet-Verteilung ist einfach handhabbar und „parameterarm", da sie zur Beschreibung der gemeinsamen Verteilung einer Gruppe von m Zufallsvariablen nur m+1 Parameter benötigt. Im Endeffekt kann also ein Dirichlet-MVB-Modell leicht mit weniger als der doppelten Parameteranzahl – verglichen mit einem Standard-MVB-Modell – auskommen.

Anders als Latent-Class-MVB-Modelle gehören Beta- und Dirichlet-MVB-Modelle selbst nicht zur Klasse der MVB-Modelle. Dennoch stößt die Herleitung von Schätz- und Testverfahren auf keine besonderen Schwierigkeiten. Ich werde die Details hier jedoch nicht weiter ausführen, da Beta- und Dirichlet-MVB-Modelle in der Praxis kaum Bedeutung erlangen dürften. Der Grund ist die Unabhängigkeitsannahme zwischen verschiedenen Parametergruppen, der in diese Modellklasse eingeht (vgl. Gleichung (5.4.1.2-2)). Abhängigkeiten zwischen den Zufallsparametern innerhalb eines Astes können demnach nur noch dadurch zu-

stande kommen, daß Zufallsparameter aus einer Parametergruppe in einem Ast mehrfach vorkommen. Kommt eine Parametergruppe pro Ast eines MVB-Modells dagegen maximal einmal vor – was bei recht vielen MVB-Modellen der Fall ist –, so implizieren Beta- und Dirichlet-MVB-Modelle die Annahme der kompletten Unkorreliertheit aller Zufallsparameter innerhalb der Äste des Modells. Wie wir gesehen haben, sind derartige Modelle aber aggregierungsinvariant. Die statistischen Standardanalysen für MVB-Modelle mit fixen Parametern sind im Falle aggregierungsinvarianter Modellen gegenüber Parameterheterogenität weitgehend robust (vgl. Abschnitt 5.3). Es bedarf also eigentlich keiner komplizierten Mischverteilungsmodelle, um die Verteilungsklasse angemessen analysieren zu können, die durch Beta- und Dirichlet-MVB-Modelle beschrieben wird. Lediglich bei Aussagen zu Konfidenzintervallen und bei Modellen, in denen sich Parameter einer Parametergruppe in mindestens einem Ast des Modells wiederholen, ist Vorsicht geboten. Derartige Fälle lassen sich aber problemlos über Latent-Class-MVB-Modelle behandeln, die im letzten Abschnitt besprochen wurden.

Als Fazit bleibt festzuhalten, daß Beta- und Dirichlet-MVB-Modelle zwar prinzipiell denkbare Modelle für stetig verteilte Zufallsparameter in MVB-Modellen sind; wegen der involvierten Unabhängigkeitsannahme sind sie jedoch für die Praxis nicht so nützlich wie Latent-Class-MVB-Modelle, die Korrelationen zwischen allen Parametern des Modells zulassen. Die Beschränkung auf diskrete Verteilungen der Zufallsparameter bei Latent-Class-MVB-Modellen wiegt demgegenüber nicht schwer: Selbst wenn diese Annahme falsch sein sollte, sind diskrete Verteilungen doch immerhin als Approximationen an stetige Verteilungen brauchbar. Solange keine praktikablen Modelle für stetig verteilte Zufallsparameter verfügbar sind, die ebenfalls Korrelationen zwischen allen Paaren von Modellparametern erlauben, werden Latent-Class-MVB-Modelle daher die wichtigere Modellklasse bilden.

5.4.2.3 Rasch-MVB-Modelle

Die typische Datenstruktur innerhalb einer Versuchsbedingung ist in der Kognitionspsychologie und in der psychologischen Diagnostik sehr ähnlich: Wir haben es fast immer mit einer Vpn × Items-Matrix zu tun, d.h. eine Menge von Vpn bearbeitet eine Menge von Items, so daß für jede Vp-Item-Kombination ein singuläres, im Regelfall kategoriales Datum vorliegt. In der psychologischen Diagnostik sind die Items z.B. Intelligenztestaufgaben, die ein Proband löst oder nicht. In der Kognitionspsychologie kann es sich z.B. um alte Items im Rahmen eines Rekognitionstests handeln, die eine Vp korrekt als alt erkennt (Treffer) oder nicht (Verpasser). Denkbar ist auch, daß pro Vp-Item-Kombination ein mehrkategoriales Datum vorliegt: In der psychologischen Diagnostik etwa ein Punktwert in einer In-

telligenztestaufgabe oder ein Rating in einem Fragebogen, in der Kognitionspsychologie beispielsweise die Klassifikation eines gelernten Wortpaares als „beide Wörter nacheinander reproduziert" ($C_{1,1}$), „beide Wörter reproduziert, aber nicht nacheinander" ($C_{1,2}$), „nur eines der beiden Wörter reproduziert" ($C_{1,3}$), „keines der beiden Wörter reproduziert" ($C_{1,4}$), die dem EA-Modell zugrunde liegt (vgl. Batchelder & Riefer, 1980; 1986). Erst durch Aggregierung derartiger singulärer Daten über Items und/oder Personen hinweg entstehen die Stichprobenhäufigkeitsverteilungen, welche die Grundlage statistischer Analysen von MVB-Modellen im Regelfall bilden.

Die Strukturähnlichkeit der Datensätze in der psychologischen Diagnostik und in der Kognitionspsychologie läßt den Versuch naheliegend erscheinen, die Methoden, welche in der psychologischen Testtheorie zur Repräsentation und Messung von Person- und Itemunterschieden entwickelt wurden, in die MVB-Modellbildung zu integrieren, um auf diese Weise kognitionspsychologische Modelle zu erhalten, die Parameterheterogenität aufgrund von Vp- und Itemeffekten nicht nur explizit berücksichtigen, sondern auch messen können. Dies wäre vor allem unter Anwendungsgesichtspunkten von großer Bedeutung, denn in der klinischen, neuropsychologischen oder pädagogischen Praxis interessieren im Regelfall nicht nur testtheoretisch, sondern auch substanzwissenschaftlich fundierte Maße für individuelle kognitive Fertigkeiten, weniger dagegen Charakteristika von Probandengruppen. Derartige Messungen können die Grundlage für therapeutische oder pädagogische Interventionen bilden, die im konkreten Einzelfall auf die Förderung spezieller Fertigkeiten bzw. die Kompensation vorhandener Defiziten abzielen. Wo individuums- und itemspezifische Maße nicht auf besonderes Interesse stoßen — beispielsweise in der kognitionspsychologischen Grundlagenforschung — können Modelle mit Person- und Itemparametern im Sinne einer Erhöhung der Robustheit statistischer Analysen gegenüber Person- und Itemeffekten dennoch von Bedeutung sein; falls erwünscht, kann man individuumsspezifische Parameterschätzungen ja anschließend wieder mitteln, um Kennwerte für Gruppen zu erhalten.

Unter den psychometrischen Modellen der Testtheorie nehmen das Rasch-Modell und seine Varianten aus wohlbekannten Gründen, die hier nicht rekapituliert zu werden brauchen, eine herausragende Stellung ein (vgl. Fischer, 1974; Rasch, 1960/1980; Roskam, 1996b; Rost, 1988; 1999). Rasch-Modellierung und MVB-Modellierung lassen sich erfreulicherweise zwanglos zusammenführen, indem man (a) MVB-Modelle zunächst explizit für die Personen-Item-Ebene formuliert und dann (b) für jeden binären Modellparameter ein separates Rasch-Modell postuliert. Die Modellparameter sind ja nichts anderes als Wahrscheinlichkeiten latenter Ereignisse (analog zu Lösungswahrscheinlichkeiten von Intelligenztestitems), so daß deren Parametrisierung im Sinne von Rasch-Modellen auf keinerlei

konzeptuelle Probleme stößt. Ist $v = 1, ..., V$ der Index für die V Vpn eines Experiments und $t = 1, ..., T$ der Index für die T Items, so erhält man folglich eine zu Gleichung (3.5-1) analoge Modellgleichung für ein Rasch-MVB-Modell durch

$$\forall k,j,v,t\colon p_{k,j,v,t}(\boldsymbol{\theta}_{v,t}) = \\ = \sum_{i=1}^{I_{k,j}} c_{k,j,i} \cdot \prod_{s=1}^{S} \left[(1-\theta_{s,v,t})^{b_{k,j,i,s}} \cdot \left(\theta_{s,v,t}\right)^{a_{k,j,i,s}} \right], \qquad (5.4.2.3\text{-}1)$$

wobei für jeden latenten Parameter ein Rasch-Modell

$$\theta_{s,v,t} = \frac{e^{(\xi_{s,v}-\varepsilon_{s,t})}}{1+e^{(\xi_{s,v}-\varepsilon_{s,t})}} \qquad (5.4.2.3\text{-}2)$$

gilt. Sofern sich die in verschiedenen Versuchsbedingungen des Modells ($k = 1, ..., K$) verwendeten Vpn oder Items unterscheiden, sollten natürlich auch die zugehörigen Vp-„Fähigkeitsparameter" $\xi_{s,v}$ oder Item-„Schwierigkeitsparameter" $\varepsilon_{s,t}$ zwischen den Bedingungen unterschiedliche Werte annehmen können. Wenn dagegen verschiedenen Versuchsbedingungen partiell oder generell dieselben Vpn oder Items zugrunde liegen, so sollten selbstverständlich auch die betreffenden Person- bzw. Itemparameter über die Versuchsbedingungen hinweg Gleichheitsrestriktionen unterzogen werden. Kurz: Die Indizes *s,v* und *t* laufen über *alle* Parameter, Vpn und Items des Gesamtmodells, egal in welchen Versuchsbedingungen sie auftauchen bzw. Verwendung finden.

Die Verallgemeinerung des Ansatzes auf nichtbinäre MVB-Modelle stößt auf keine prinzipiellen Probleme. Man formuliert einfach für jede nichtbinäre Parametergruppe ein mehrkategoriales bzw. polytomes Rasch-Modell (vgl. z.B. Roskam, 1996b, S. 451f). Ist also $\theta_{g,l}$ ein Parameter der nichtbinären Parametergruppe g ($g = 1, ..., G$), so setzt man

$$\theta_{g,l,v,t} = \frac{e^{-(\xi_{g,l,v}-\varepsilon_{g,l,t})}}{\sum_{r=1}^{L_g} e^{-(\xi_{g,r,v}-\varepsilon_{g,r,t})}} \qquad (5.4.2.3\text{-}3)$$

fest. Für jede Stufe l der Parametergruppe g wird also eine separate Menge von Personenparametern und Itemparametern – wenn man so will: ein separater Latent Trait – postuliert.

Die Idee, Parameter von MVB-Modellen durch separate Rasch-Modelle zu ersetzen, um auf diese Weise Parameterheterogenitätsprobleme in den Griff zu bekommen, ist in der Literatur zu MVB-Modellen zwar in jüngster Zeit angedeutet worden (z.B. 1998; Batchelder, 1999; Batchelder & Crowther, 1997, S. 48; Batchelder, Kumbasar & Boyd, 1997, S. 37), wurde aber in der MVB-Literatur meines Wissens noch niemals praktisch umgesetzt. Interessanterweise ist der gleiche Gedanke schon viele Jahre zuvor in der psychometrischen Literatur aufgetaucht, zu einem

Zeitpunkt, als die Klasse der multinomialen Verarbeitungsbaummodelle noch gar nicht formal abgegrenzt bzw. eindeutig definiert war. Das wohl bekannteste nichttriviale[65] Rasch-MVB-Modell ist ein logistisches Item-Response-Modell der Art

$$p_{v,t}(c_t,\xi_v,\varepsilon_t) = c_t + (1-c_t)\cdot\frac{e^{(\xi_v-\varepsilon_t)}}{1+e^{(\xi_v-\varepsilon_t)}}\ . \qquad (5.4.2.3\text{-}4)$$

Die Lösungswahrscheinlichkeit $p_{v,t}$ eines dichotomen Items t für Person v hängt hierbei von zwei Itemparametern c_t und ε_t sowie einem Fähigkeitsparameter ξ_v ab. Es handelt sich um einen Spezialfall des sogenannten dreiparametrigen logistischen Modells von Birnbaum (1968, vgl. auch Roskam, 1996b, S. 454, Gleichung (34)). Als Rasch-MVB-Modell läßt es sich wie folgt motivieren: Man nimmt an, daß ein Item auf zwei Wegen gelöst werden kann, nämlich einerseits durch Nachdenken und andererseits durch erfolgreiches Raten. Durch Nachdenken wird die Lösung mit einer Wahrscheinlichkeit gefunden, die dem Rasch-Modell mit Personenparameter ξ_v und Schwierigkeitsparameter ε_t folgt. Führt Nachdenken mit der komplementären Wahrscheinlichkeit nicht zum Erfolg, so wird geraten. Für erfolgreiches Raten gilt ein „degeneriertes Rasch-Modell", bei dem alle Personenparameter gleich sind; die Wahrscheinlichkeit hängt dann nur noch vom Item ab (c_t). Nur wenn beide Wege — Nachdenken und Raten — scheitern, wird das Item nicht gelöst. Insgesamt resultiert ein Baummodell mit drei Ästen, wovon zwei zu einer korrekten Antwort führen.

Etwas komplexere Rasch-MVB-Modelle wurden Anfang der achtziger Jahre von Susan Embretson-Whitely unter der Bezeichnung „Mehrkomponenten-Latent-Trait-Modelle" vorgestellt (Embretson, 1984; Whitely, 1980).[66] Eines dieser Modelle ist das „independent components model" (Whitely, 1980), welches Maris (1992; 1995) später als „konjunktives Rasch-Modell" bezeichnet hat. Angenommen wird hierbei, daß eine Aufgabe nur dann erfolgreich bewältigt wird, wenn mehrere involvierte Teilaufgaben erfolgreich bewältigt werden; wird auch nur eine Teilaufgabe nicht bewältigt, scheitert die gesamte Aufgabenbearbeitung. Für jede Teilaufgabe wird dabei ein separates Rasch-Modell angenommen. Wie man sich leicht klarmachen kann, korrespondiert diesem Ansatz ein einfaches binäres Verarbeitungsbaummodell auf der Vp-Item-Ebene, das in Abbildung 5.6 dargestellt ist.

---

[65] Der Zusatz „nichttrivial" wird gewählt, weil natürlich auch das Modell von Rasch (1960) selbst als Rasch-MVB-Modell für einen einfachen binären Ereignisbaum mit zwei „beobachteten Ästen" (Item gelöst versus Item nicht gelöst) aufgefaßt werden kann. Umgekehrt sind auch alle MVB-Modelle zugleich Rasch-MVB-Modelle (mit identischen Itemparametern und identischen Personenparametern pro Knoten).

[66] Whitely (1980, S. 486ff.) hat zusätzlich die Beziehungen zu anderen Verallgemeinerungen des klassischen Rasch-Modells (Rasch, 1960) diskutiert. Obwohl einige dieser Modelle der Klasse der Rasch-MVB-Modelle recht ähnlich sind, sind allein die Mehrkomponentenmodelle von Embretson-Whitely mit Rasch-MVB-Modelle formal voll vereinbar.

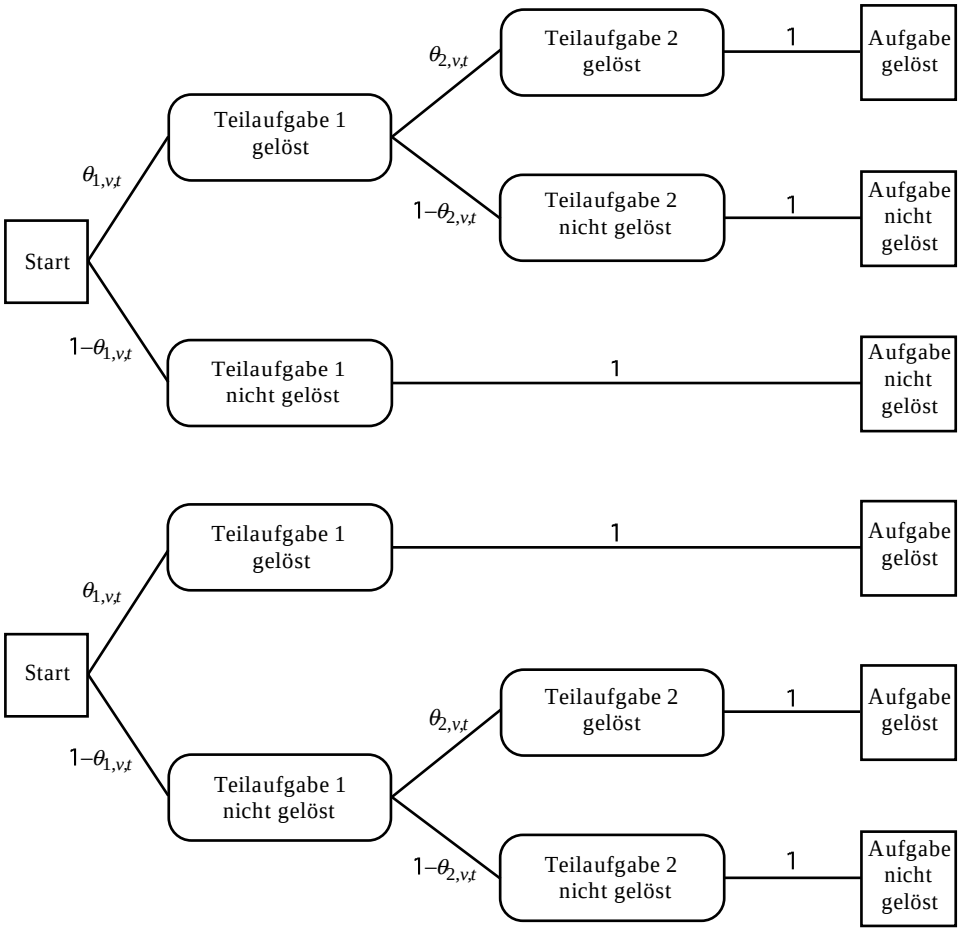

Abbildung 5.6:
Das konjunktive Rasch-Modell (oben) und das disjunktive Rasch-Modell (unten) für zwei kognitive Teilprozesse bei der Aufgabenbearbeitung (vgl. Maris, 1992; 1995). Jeder Modellparameter $\theta_{s,v,t}$ (s = 1, 2) ist als Rasch-Modell gemäß Gleichung (5.4.2.3-2) zu parametrisieren.

Eine naheliegende Alternative zum konjunktiven Rasch-Modelle ist das „disjunktive Rasch-Modell" (Maris, 1992; 1995): Die Lösung einer von mehreren Teilaufgaben ist nach diesem Modell hinreichend für die Lösung der Gesamtaufgabe. Offensichtlich entspricht das disjunktive Rasch-Modell u.a. den Strukturannahmen des 1HT- und des 2HT-Rekognitionsmodells für alte Items: Wenn entweder ein Item als alt erkannt (Teilprozeß 1) oder „ja" geraten wird (Teilprozeß 2), kommt es zu einer „Ja"-Antwort auf ein Zielitem im Rekognitionstest, d.h. zu einem Treffer. Nur wenn beide Teilprozesse (Rekognition und Raten) scheitern, resultiert eine „Nein"-Antwort (Verpasser).

Maris (1992; 1995) nennt Modelle wie das konjunktive und das disjunktive Rasch-Modell „psychometric latent response models". Die von Maris aufgeführte Beispiele für diese Modellklasse zeigen, daß er damit im Kern die Rasch-MVB-Modellklasse meint, auch wenn die Rahmenmodellgleichung (5.4.2.3-1) nicht explizit aufgeführt wird. Jedenfalls lassen sich alle von ihm aufgeführten Beispiele für „psychometric latent response models" unter die o.g. Modellgleichung subsumieren.

Hinsichtlich der statistischen Analyse von Rasch-MVB-Modellen sind bereits beträchtliche Fortschritte erzielt worden. Embretson (1980; 1984) hat sich noch darauf beschränkt, beobachtbare Indikatoren (Tests) für die involvierten Teilaufgaben zu suchen, auf die dann gewöhnliche Rasch-Modelle angepaßt wurden. Sie hat die latenten Teilprozesse also zunächst „manifest" gemacht, bevor beispielsweise das konjunktive Rasch-Modell für eine komplexe Gesamtaufgabe überprüft wurde. In statistischer Hinsicht kommt man dabei mit den Mitteln aus, die bei der Analyse von Rasch-Modellen Standard sind (Fischer, 1974; Rost, 1988). Die Modellgleichungen des konjunktiven Rasch-Modells wurden von ihr nur benutzt, um die Lösungswahrscheinlichkeiten einer Gesamtaufgabe aus den Rasch-Parameterschätzungen der untersuchten Teilaufgaben *vorherzusagen*. Stimmen die vorhergesagten Lösungswahrscheinlichkeiten und die tatsächlichen Lösungen bzw. Fehler bei der Gesamtaufgabe für die verschiedenen Vp-Item-Kombinationen gut überein, so schließt sie auf die Gültigkeit des konjunktiven Rasch-Modells. Dieser Ansatz hängt also ganz wesentlich von der Verfügbarkeit spezieller Tests für die involvierten Teilaufgaben ab. Ohne solche Tests ist der Ansatz von Embretson nicht umsetzbar.

Einen erheblichen Fortschritt bedeuten die Arbeiten von Maris (1992; 1995). Er stellte allgemein anwendbare Parameterschätzverfahren für Rasch-MVB-Modelle (bzw. psychometric latent response models) vor, die auch dann funktionieren, wenn keine Indikatoren für die involvierten Teilprozesse zur Verfügung stehen, wenn diese also – wie in dieser Arbeit durchweg angenommen – vollständig latent sind. Hierzu zählt einerseits ein EM-Algorithmus der ML-Parameterschätzung für „latente Rasch-Modelle". Da unbedingte ML-Schätzer bei diesen Modellen nicht immer existieren müssen und außerdem die unbedingten ML-Itemparameterschätzungen wegen der Existenz inzidenteller Parameter (hier: der Personenparameter) nicht konsistent sind (vgl. Neyman & Scott, 1948), hat er darüber hinaus alternative (marginal maximum likelihood) Schätzmethoden vorgestellt, die diese Probleme nicht aufweisen. Die praktische Umsetzung der Schätzmethoden hat er am Beispiel des konjunktiven Rasch-Modells demonstriert. Aus Platzgründen kann hier nicht näher darauf eingegangen werden.

Ein Problem, das die Analysen von Maris (1992; 1995) noch offenlassen, sind Gleichheitsrestriktionen von Item- und/oder Personenparametern über verschiedene Versuchsbedingungen bzw. Aufgaben hinweg. Zumindest die unbedingte ML-Schätzung für derartige Problemstellungen ist aber leicht umsetzbar. Die Likelihood-Funktion für binäre Rasch-MVB-Modelle ergibt sich unmittelbar aus der Modellgleichung (5.4.2.3-1) in Verbindung mit (5.4.2.3-2), indem man einfach das Produkt der Wahrscheinlichkeiten für die Beobachtungsresultate bei den verschiedenen Vp-Item-Kombinationen bildet. Maxima dieser Likelihood-Funktion kann man mit Gradientenverfahren suchen, für die in Abschnitt 4.7 Computer-

programme vorgestellt wurden. Um die Identifizierbarkeit des Modells sicherzustellen, sind – wie anderen Rasch-Modellen auch – geeignete Identifizierbarkeitsrestriktionen einzelner Person- oder Itemparameter (z.B. $\varepsilon_{s,1} = 0$, $s = 1, ..., S$) vorzunehmen (Maris, 1992; 1995).

Eine Modellprüfung kann – wie bei Rasch-Modelltests üblich – über Tests auf Homogenität der Personenparameter in verschiedenen Itemsets und/oder Tests auf Homogenität der Itemparameter in verschiedenen Personengruppen erfolgen. Auf Probleme und Lösungsmöglichkeiten der Prüfung von Rasch-Modellen kann hier nicht näher eingegangen werden (Klauer, 1988; 1991; Rost, 1988). In jedem Fall lassen sich die bei einfachen Rasch-Modellen praktikablen Ansätze auch auf Rasch-MVB-Modelle übertragen.

5.5 Zusammenfassung

In diesem Kapitel wurden Probleme der Robustheit asymptotischer Standardverfahren (Punktschätzungen, Bereichsschätzungen, Modelltests) bei der statistischen Analyse von multinomialen Modellen diskutiert. Speziell ging es um die Frage, wie sich Verletzungen der Annahme konstanter (fixer) Parameterwerte, die qua Modell allen Elementen einer Aggregatstichprobe in gleicher Weise zugrundeliegen müssen, auf die Güte der asymptotischen Verfahren auswirken.

Der erste Abschnitt 5.1 gab einen Überblick über bislang publizierte Ergebnisse zum Effekt von Parametervariabilität in MVB-Modellen auf mittlere Parameterschätzungen und Konfidenzintervalle einerseits und auf α-Fehler bei Modellgeltungstests andererseits. Sind die Parameter eines MVB-Modells nicht mehr konstant, sondern selbst Zufallsvariablen, so wird aus dem MVB-Modell ein MVB-Modell mit Zufallsparametern. Die in der Literatur berichteten Ergebnisse lassen vermuten, daß MVB-Modelle mit Zufallsparametern wie gewöhnliche MVB-Modelle mit fixen Parametern analysiert werden können, sofern geringe Grade der Parametervariabilität vorliegen. Demgegenüber führen Parameterstreuungen ab $\sigma = .20$ unter gewissen Randbedingungen, die z.B. beim EA-Modell erfüllt sind, durchaus zu verzerrten Ergebnissen. Dies gilt dann um so mehr, je größer die zugrundeliegenden Stichproben sind. In Abschnitt 5.2 wurde herausgearbeitet, daß derartige Probleme immer dann zu erwarten sind, wenn korrelierte Zufallsparameter innerhalb der Äste eines Modells vorliegen. Demgegenüber folgen die mittleren Kategoriewahrscheinlichkeiten bei Modellen mit unkorrelierten Zufallsparametern unabhängig von der Höhe der Parameterstreuungen immer dem MVB-Modell, das sich durch Fixierung der Zufallsparameter auf ihre Erwartungswerte ergibt. Mit anderen Worten: Bei unkorrelierten Zufallsparametern entspricht das Modell für die durchschnittlichen aggregierten Daten gerade dem strukturgleichen MVB-Mo-

dell mit den durchschnittlichen individuellen Parametern. Solche MVB-Modelle mit Zufallsparametern wurden als „aggregierungsinvariant" bezeichnet.

Eine in Abschnitt 5.3 berichtete Monte-Carlo-Studie zum EA-Modell zeigte erwartungsgemäß, daß statistische Analysen von MVB-Modelle mit korrelierten Parametern nur gegenüber kleinen Parameterstreuungen ($\sigma \leq .10$) robust sind, während bei aggregierungsinvarianten Modellen generell von robustem Verhalten ausgegangen werden kann. Dies gilt zumindest für Punktschätzungen und für das Verhalten des Modellgeltungstests bei kleinen und mittleren Stichprobenumfängen. Im Falle sehr großer Stichproben ist ein konservatives Verhalten des Modelltests bei aggregierungsinvarianten Modellen zu beobachten, das sich jedoch erfreulicherweise kaum nachteilig auf die Teststärke auswirkt. Als erfreuliches Fazit kann festgehalten werden, daß statistische Standardanalysen von MVB-Modellen bei aggregierungsinvarianten Modellen im allgemeinen gerechtfertigt zu sein scheinen, und zwar auch dann, wenn erhebliche Parameterstreuungen vorliegen. Daß die Annahme eines aggregierungsinvarianten Modells für die Anwendungspraxis durchaus realistisch ist, illustriert das Speicher-Abruf-Modell von Chechile (1987; Chechile & Meyer, 1976; vgl. Abschnitt 7.6.1), für das häufig versuchspersonenspezifische Parameterschätzungen abgeleitet wurden. Sowohl für Kinder als auch für Erwachsene erwiesen sich die Speicher- und Abrufparameter dieses Modells immer wieder als unkorreliert (zusammenfassend: Brainerd, 1985, S. 162). Dieses Ergebnis ist natürlich nicht ohne weiteres auf andere Modelle generalisierbar; dennoch zeigt es, daß die Annahme eines aggregierungsinvarianten Modells nicht aus der Luft gegriffen ist.

So vielversprechend die in diesem Kapitel berichteten Monte-Carlo-Ergebnisse zur Robustness von Hypothesentests bei Modellgültigkeit und Modellungültigkeit sind, so wenig können sie ohne weiteres vom Enkodierung-Abruf-Modell auf andere MVB-Modelle oder parametrisierte multinomiale Modelle im allgemeinen verallgemeinert werden. Denkbar ist, daß die Robustheitsthese auch für den Fall unkorrelierter Parameter für bestimmte Modelle relativiert werden muß, speziell im Hinblick auf die Teststärke bei Modellverletzungen. Ich schließe mich daher der Empfehlung von Riefer und Batchelder (1991b) sowie Garcia-Pérez (1994) an, im Zweifelsfall für das konkret zu betrachtende Modell eine weitere Monte-Carlo-Studie zur Robustness der Tests unter H_0 und H_1 durchzuführen. Mit den in diesem Kapitel beschriebenen Techniken der Erzeugung beta-verteilter Zufallsparameter und den Computerprogrammen AppleTree oder GPT ist eine solche Monte-Carlo-Studie immer relativ leicht umsetzbar.

In Abschnitt 5.4 wurden alternative statistische Analyseverfahren vorgestellt, die bei nicht aggregierungsinvarianten Modellen mit großen Parameterstreuungen angewandt werden können. Ein Ansatz besteht darin, auf die Ebene einzelner Vpn (soweit sie strukturell ausreichende Datensätze liefern) oder auf die

Ebene mutmaßlich homogener Teilstichproben zurückzugehen. Geeignete Verfahren für die Analyse kleiner Stichproben (Bootstrap, Jackknife, Bayessche Verfahren) wurden vorgestellt. Befriedigender ist ein alternativer Ansatz, der darauf abzielt, die Parameterheterogenität im Modell explizit zu berücksichtigen. Drei derartige Modellklassen wurden vorgestellt, wovon Latent-Class-MVB-Modelle und Rasch-MVB-Modelle am vielversprechendsten erscheinen. Latent-Class-MVB-Modelle haben den Vorteil, statistisch sehr leicht analysierbar zu sein; Rasch-MVB-Modelle stellen größere Ansprüche an die statistische Analyse, liefern aber Schätzer für Person- und Itemparameter der einzelnen kognitiven Teilprozesse und sind damit in individualdiagnostischen Kontexten besonders nützlich.

6. MVB-Modelle als Rahmenmodelle

Am Ende des letzten Kapitels wurden *Verallgemeinerungen* von MVB-Modellen behandelt, die Parametervariabilität zu repräsentieren erlauben. In diesem Kapitel soll — mit einer Ausnahme — genau die umgekehrte Perspektive eingenommen werden: Gefragt werden soll, welche bekannten und in der kognitiven Psychologie teilweise schon gebräuchlichen Modellklassen sich als *Spezialfälle* unter das „Dach" der MVB-Modelle subsumieren lassen.

Mehrere Klassen von Spezialfällen werden der Reihe nach vorgestellt und an Beispielen veranschaulicht: Modelle der Genetik (Abschnitt 6.2), Latent-Class-Modelle (Abschnitt 6.3), Binomialmischverteilungsmodelle (Abschnitt 6.4), Markoff-Ketten-Modelle (Abschnitt 6.5), gemischte Markoff-Ketten-Modelle (Abschnitt 6.6) und verallgemeinerte Signalentdeckungsmodelle (Abschnitt 6.7). Eine erschöpfende Auflistung aller Spezialfälle ist damit nicht intendiert. Es soll lediglich an einigen Beispielen illustriert werden, daß Modelle, die man vielleicht nicht spontan mit MVB-Modellen in Verbindung bringen würde, durchaus Instanzen dieser Modellklasse sein können. Darüber hinaus gibt es bekannte Modelle für kategoriale Daten, die zwar nicht zur MVB-Modellklasse gehören, jedoch zu bestimmten MVB-Modellen datenäquivalent sind (vgl. die Behandlung log-linearer Modelle in Abschnitt 6.1). Der praktische Gewinn, der mit derartigen Einsichten verbunden ist, besteht darin, daß das gesamte statistische Instrumentarium, welches für MVB-Modelle entwickelt und in dieser Arbeit ausführlich dargestellt wurde, auf die vorzustellenden Modelle übertragbar ist. Dies ermöglicht potentiell neue Einsichten in Datenstrukturen, die gewöhnlich mit anderen methodischen Mitteln analysiert werden.

Die angedeutete Ausnahme betrifft eine Erweiterung von MVB-Modellen auf Modelle für die simultane Analyse von Reaktionshäufigkeiten und Reaktionszeiten. Zwei Modellklassen für Reaktionshäufigkeiten und Reaktionszeiten werden in Abschnitt 6.8 kurz skizziert, jedoch nicht voll ausgearbeitet. Obwohl es sich hierbei teilweise nicht um Submodelle der MVB-Modellfamilie handeln kann, und obwohl sie auch nicht zu bestimmten MVB-Modellen datenäquivalent sind, fungieren MVB-Modelle doch insofern als Rahmenmodelle, als deren Struktureigenschaften für die Modellierung von Antwort- bzw. Reaktionszeiten herangezogen werden. Der Nutzen dieser Modellklasse besteht insbesondere darin, zwischen substanzwissenschaftlichen Theorien, die äquivalenten oder datenäquivalenten MVB-Modellen entsprechen (vgl. etwa Abbildung 3.5 in Abschnitt 3.6), im Rahmen der erweiterten Modellklasse potentiell empirisch diskriminiert werden kann.

Eine Zusammenfassung schließt das Kapitel ab (Abschnitt 6.9).

6.1 Log-lineare Modelle

Log-lineare Modelle sind die wohl am häufigsten verwendeten strukturellen Modelle für Kontingenztafeln bzw. multivariate kategoriale Datensätze ganz allgemein (Agresti, 1990; Andersen, 1990; 1996; Bishop et al., 1975; Fienberg, 1977; Haberman, 1978; 1979; Wickens, 1998a). Für eine dreidimensionale Kontingenztafel, die durch die Kreuzklassifikation dreier Variablen A, B und C mit I, J und K Kategorien aufgespannt wird, lautet das saturierte log-lineare Modell[67]

$$\ln(p_{i,j,k}) = \ln(1/z) + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,j}^{AB} + \lambda_{i,k}^{AC} + \lambda_{j,k}^{BC} + \lambda_{i,j,k}^{ABC} . \qquad (6.1\text{-}1)$$

Hierbei ist $p_{i,j,k}$ die Wahrscheinlichkeit der Zelle (i, j, k) in der Kontingenztafel, d.h. die Wahrscheinlichkeit, daß Variable A die Kategorie a_i, $i = 1, ..., I$, Variable B die Kategorie b_j, $j = 1, ..., J$, und zugleich Variable C die Kategorie c_j', $k = 1, ..., K$, annimmt. Der Konstante z ist die Anzahl der Zellen in der Kontingenztafel. Hat jede der drei Variablen A, B und C beispielsweise zwei Kategorien ($I = J = K = 2$), so ist $z = 2 \cdot 2 \cdot 2 = 8$.

In Gleichung (6.1-1) sind insgesamt $I + J + K + I{\cdot}J + I{\cdot}K + J{\cdot}K + I{\cdot}J{\cdot}K$ Modellparameter enthalten, nämlich die λ-Parameter für alle möglichen Wertekombinationen der Indizes i, j und k. Dieser Vielzahl an Parametern stehen nur $I{\cdot}J{\cdot}K - 1$ freie Kategoriewahrscheinlichkeiten $p_{i,j,k}$ gegenüber. Damit das Modell (6.1-1) identifizierbar wird, müssen Nebenbedingungen zugrunde gelegt werden, welche die Anzahl freier Parameter reduzieren. Üblich ist die Forderung, daß die durchschnittlichen λ-Parameter über die Stufen jeder Variablen der Kontingenztafel null sind:

$$\begin{aligned}
&\sum_{i=1}^{I} \lambda_i^A = \sum_{i=1}^{I} \lambda_{i,j}^{AB} = \sum_{i=1}^{I} \lambda_{i,k}^{AC} = \sum_{i=1}^{I} \lambda_{i,j,k}^{ABC} = 0 \\
&\sum_{j=1}^{J} \lambda_j^B = \sum_{j=1}^{J} \lambda_{i,j}^{AB} = \sum_{j=1}^{J} \lambda_{j,k}^{BC} = \sum_{j=1}^{J} \lambda_{i,j,k}^{ABC} = 0 \\
&\sum_{k=1}^{K} \lambda_k^C = \sum_{k=1}^{K} \lambda_{i,k}^{AC} = \sum_{k=1}^{K} \lambda_{j,k}^{BC} = \sum_{k=1}^{K} \lambda_{i,j,k}^{ABC} = 0 .
\end{aligned} \qquad (6.1\text{-}2)$$

Diese Nebenbedingungen stellen sicher, daß genau $I{\cdot}J{\cdot}K - 1$ freie (und identifizierbare) Modellparameter verbleiben. Bei einer 2×2×2-Kontingenztafel gibt es folglich nur noch einen freien λ-Parameter in jeder der in Gleichung (6.1-1) aufgeführten Parametergruppen, insgesamt also sieben freie Parameter.

---

[67] Gewöhnlich werden log-lineare Modelle nicht für die Zellwahrscheinlichkeiten, sondern für die erwarteten Häufigkeiten (die mit N multiplizierten Zellwahrscheinlichkeiten) formuliert. Da in diesem Fall unterschiedlichen Stichprobenumfängen in der Modellgleichung Rechnung getragen werden muß, ist die Konstante $\ln(1/z)$ durch den freien Parameter λ_0 zu ersetzen.

Sind alle λ-Parameter gleich null, so entspricht das log-lineare Modell dem Gleichverteilungsmodell über alle z Zellen der Kontingenztafel (mit $p_{i,j,k} = 1/z$). Die λ-Parameter mit einem Index werden gewöhnlich als „Haupteffekte", die mit zwei Indizes als „Zweifachwechselwirkungen" und die mit drei Indizes als „Dreifachwechselwirkungen" der Variablen auf die Kategoriehäufigkeiten interpretiert. Hierbei ist allerdings zu beachten, daß sich diese Begriffe auf die *logarithmierten* Kategoriewahrscheinlichkeiten beziehen und somit eine andere Bedeutung haben als die Haupt- und Wechselwirkungsbegriffe im linearen Modell der mehrfaktoriellen Varianzanalyse. Im Kern stehen Wechselwirkungen in log-linearen Modellen für bestimmte stochastische Abhängigkeiten zwischen den Variablen der Kontingenztafel. Fehlen alle Wechselwirkungseffekte, entspricht das dem Modell der totalen stochastischen Unabhängigkeit von A, B und C. Fehlen nur einige Wechselwirkungseffekte, während andere im Modell verbleiben, so entspricht das in der Regel Modellen der Unabhängigkeit oder bedingten Unabhängigkeit einer Teilmenge von Variablen für festgehaltene Ausprägungen der anderen Variablen (vgl. z.B. Andersen, 1996, S. 308f).

Gewöhnlich geht man bei der Analyse von Kontingenztafeln so vor, daß nach dem sparsamsten passenden hierarchischen log-linearen Modell gesucht wird: Ausgehend vom saturierten Modell (6.1-1) werden sukzessive alle Wechselwirkungs- und Haupteffekte eliminiert, die für eine gute Datenanpassung verzichtbar sind. „Hierarchisch" ist ein log-lineares Modell, wenn zu jedem beliebigen Modellparameter höherer Ordnung (z.B. zur $A{\times}B$-Wechselwirkung) immer auch alle Modellparameter niedrigerer Ordnung der involvierten Variablen in der Modellgleichung auftauchen (hier: die A- und B-Haupteffekte). Die Einschränkung auf hierarchische Modelle ist sinnvoll, weil die Interpretation der Parameter nichthierarchischer Modelle Schwierigkeiten bereitet, auf die hier nicht näher eingegangen werden soll (vgl. dazu z.B. Fienberg, 1977). Insbesondere ist eine Interpretation in Termini von Unabhängigkeits- bzw. konditionalen Unabhängigkeitsmodellen nicht immer möglich (Haberman, 1978, Kapitel 3). Alle hierarchischen log-linearen Modelle, die sich zu einer dreidimensionalen Kontingenztafel konstruieren lassen, sind in Tabelle 6.1 (linke Seite) aufgeführt.

Für die Erhellung der Beziehungen zwischen MVB-Modellen und log-linearen Modellen ist es nützlich, die Modellgleichung (6.1-1) in antilogarithmierter Form zu schreiben:

$$p_{i,j,k} = (1/z) \cdot \gamma_i^A \cdot \gamma_j^B \cdot \gamma_k^C \cdot \gamma_{i,j}^{AB} \cdot \gamma_{i,k}^{AC} \cdot \gamma_{j,k}^{BC} \cdot \gamma_{i,j,k}^{ABC} \; . \qquad (6.1\text{-}3)$$

Hierbei stehen die γ-Parameter für die Antilogarithmen der entsprechenden λ-Parameter aus Gleichung (6.1-1). Die Modellgleichung (6.1-3) ist nunmehr ein einfaches Produkt von Parametern. Auf den ersten Blick sieht sie also aus wie die eines MVB-Modells, bei dem alle Äste in unterschiedliche Kategorien einmünden

(„beobachtete Äste"), so daß zur Berechnung der Kategoriewahrscheinlichkeiten nicht über verschiedene Äste summiert werden muß. Diese oberflächliche Ähnlichkeit von MVB-Modellgleichungen und log-linearen Modellgleichungen bedeutet aber nicht, daß log-lineare Modelle zwangsläufig immer einfache Spezialfälle von MVB-Modellen mit beobachteten Ästen sind. Die γ-Parameter in (6.1-3) können nämlich beliebige positive Werte annehmen, während die Parameter in MVB-Modellen auf das [0, 1]-Intervall eingeschränkt sind. Daraus folgt zunächst, daß log-lineare Modelle und MVB-Modelle nicht äquivalent im Sinne des Äquivalenzbegriffs aus Abschnitt 4.4 sein können, weil hierzu die numerische Identität der Parameterwerte erforderlich wäre. Möglich ist jedoch Datenäquivalenz zwischen bestimmten MVB-Modellen einerseits und log-linearen Modellen andererseits, d.h. die Äquivalenz der zugehörigen Modellgeltungshypothesen. Die Parameterwerte datenäquivalenter Modelle sind zwar nicht gleich, lassen sich jedoch durch geeignete Transformationen ineinander überführen. Voraussetzung hierfür ist ein MVB-Modell, das sich (a) auf die gleiche Kontingenztafel wie das entsprechende log-lineare Modell bezieht und (b) genau so viele Modellparameter umfaßt.

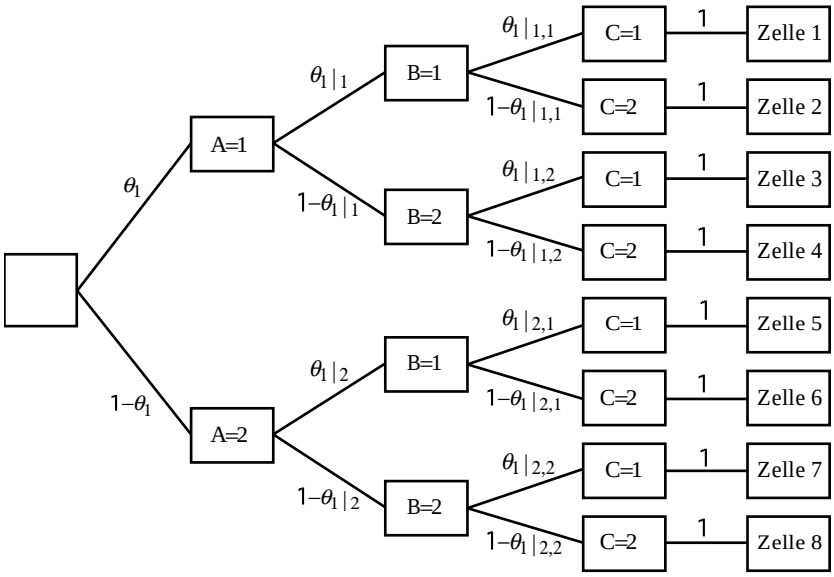

Abbildung 6.1:
Ein saturiertes binäres MVB-Modell zu einer $A\times B\times C$-Kontingenztafel mit binären Variablen, aus dem sich durch Gleichheitsrestriktionen und Parameterfixierungen Submodelle erzeugen lassen, die datenäquivalent zu verschiedenen hierarchischen log-linearen Modellen sind.

Abbildung 6.1 stellt exemplarisch ein binäres MVB-Modell für die 2×2×2-Kontingenztafel vor, aus dem sich datenäquivalente Modelle zu fast allen hierarchischen log-linearen Modellen zur gleichen Kontingenztafel erzeugen lassen. Analoges gilt für Tafeln mit mehr als zwei Kategorien pro Variable — dies würde ein nichtbinäres MVB-Modell erforderlich machen — oder für Tafeln mit mehr als drei Variablen, was mehr als drei Ebenen im Ereignisbaum erzwingen würde.

Das Grundprinzip eines geeigneten MVB-Rahmenmodells für die Analyse einer mehrdimensionalen Kontingenztafel besteht immer darin, für jede Variable der Kontingenztafel eine Ebene im zugehörigen Ereignisbaum vorzusehen. Auf der ersten Ebene werden die unbedingten Effekte der erste Variablen repräsentiert (als Abweichung zwischen θ_1 und 0.5), auf der zweiten Ebene die bedingten Effekte der zweiten Variablen innerhalb der Stufen der ersten Variablen (als Abweichung zwischen $\theta_{1|1}$ und 0.5 sowie $\theta_{2|1}$ und 0.5) usw. Durch Parametergleichsetzungen und/oder Parameterfixierungen kann man Submodelle erzeugen, die zu bestimmten log-linearen Modellen datenäquivalent sind. Tabelle 6.1 gibt eine Übersicht über die geeigneten Restriktionsmuster zu einer 2×2×2-Kontingenztafel.

Tabelle 6.1:
Hierarchische log-lineare Modelle für eine *A*×*B*×*C*-Kontingenztafel mit binären Variablen und Parameterrestriktion zum MVB-Modell aus Abbildung 6.1, die Submodelle liefern, welche zu den log-linearen Modellen datenäquivalent sind.

Nr.	hierarchisches log-lineares Modell	Restriktion im MVB-Modell
1)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,j}^{AB} + \lambda_{i,k}^{AC} + \lambda_{j,k}^{BC} + \lambda_{i,j,k}^{ABC}$	keine Parameterrestriktion
2)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,j}^{AB} + \lambda_{i,k}^{AC} + \lambda_{j,k}^{BC}$	**nicht möglich**
3)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,j}^{AB} + \lambda_{i,k}^{AC}$	$\theta_{1\|1,1} = \theta_{1\|1,2}$ & $\theta_{1\|2,1} = \theta_{1\|2,2}$
4)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,j}^{AB} + \lambda_{j,k}^{BC}$	$\theta_{1\|1,1} = \theta_{1\|2,1}$ & $\theta_{1\|1,2} = \theta_{1\|2,2}$
5)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,k}^{AC} + \lambda_{j,k}^{BC}$	**anderer Baum erforderlich**
6)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,j}^{AB}$	$\theta_{1\|1,1} = \theta_{1\|1,2} = \theta_{1\|2,1} = \theta_{1\|2,2}$
7)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{i,k}^{AC}$	wie 3) & $\theta_{1\|1} = \theta_{1\|2}$
8)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{j,k}^{BC}$	wie 4) & $\theta_{1\|1} = \theta_{1\|2}$
9)	$\lambda_0 + \lambda_i^A + \lambda_j^B + \lambda_k^C$	wie 6) & $\theta_{1\|1} = \theta_{1\|2}$
10)	$\lambda_0 + \lambda_i^A + \lambda_j^B$	wie 9) & $\theta_{1\|1,1} = 0.5$
11)	$\lambda_0 + \lambda_i^A + \lambda_k^C$	wie 9) & $\theta_{1\|1} = 0.5$
12)	$\lambda_0 + \lambda_j^B + \lambda_k^C$	wie 9) & $\theta_1 = 0.5$
13)	$\lambda_0 + \lambda_i^A$	wie 9) & $\theta_{1\|1} = \theta_{1\|1,1} = 0.5$
14)	$\lambda_0 + \lambda_j^B$	wie 9) & $\theta_1 = \theta_{1\|1,1} = 0.5$
15)	$\lambda_0 + \lambda_k^C$	wie 9) & $\theta_1 = \theta_{1\|1} = 0.5$
16)	λ_0	wie 9) & $\theta_1 = \theta_{1\|1} = \theta_{1\|1,1} = 0.5$

Wie man sieht, gibt es lediglich zwei hierarchische log-lineare Modelle, bei denen sich ein datenäquivalentes MVB-Modell nicht durch einfache Gleichheitsrestriktionen oder Parameterfixierungen zum Modell aus Abbildung 6.1 erzeugen läßt. In einem dieser Fälle (Modell 5) liegt das lediglich an der Zuordnung der Variablen zu den Ebenen des Ereignisbaums: Eine Null-Wechselwirkung zwischen den Variablen der ersten beiden Ebenen (hier: *A* und *B*) bei gleichzeitig vorhandenen Wechselwirkungen dieser Variablen mit der Variablen auf der dritten

Ebene (hier: *C*), läßt sich nicht ohne weiteres im MVB-Modell realisieren. Man kann aber leicht Abhilfe schaffen, indem man die Ebenen vertauscht: Wenn ein neues Baummodell formuliert wird, in dem entweder *A* und *C* oder *B* und *C* auf den ersten beiden Ebenen angesiedelt werden, läßt sich analog zu Tabelle 6.1 ein Muster von Gleichheitsrestriktionen finden, derart, daß das restringierte MVB-Modell zum log-linearen Modell 5) analog ist.

Ein größeres Problem wirft das log-lineare Modell 2) ohne Dreifach-Wechselwirkung auf. Hierzu läßt sich in der Tat auch nach Umstellung der Variablen kein datenäquivalentes MVB-Modell finden, das lediglich auf Gleichheitsrestriktionen oder Parameterfixierungen beruht. Mit anderen Worten: Wer dieses log-lineare Modell testen will, kommt mit Verfahren bzw. Computerprogrammen, die für MVB-Modelle konzipiert wurden, tatsächlich nicht aus. Interessanterweise ist dieses hierarchische log-lineare Modell auch dasjenige, dessen Interpretation recht problematisch ist, da es nicht einem einfachen Unabhängigkeits- oder konditionalen Unabhängigkeitsmodell entspricht (vgl. Andersen, 1996, S. 308 ff). Unter praktischen Gesichtspunkten ist der Verlust, der mit einem Verzicht auf die Analyse dieses Modells verbunden ist, also nicht sehr groß.

Alle anderen Parameterrestriktionen, die in Tabelle 6.1 aufgeführt sind, ergeben sich in einfacher Weise aus der Interpretation der entsprechenden log-linearen Modelle als Unabhängigkeits- oder konditionale Unabhängigkeitsmodelle. Die Hypothese etwa, daß *B* und *C* innerhalb der Stufen von *A* unabhängig sind, läßt sich einerseits durch die Fixierung der entsprechenden λ-Parameter auf 0 (vgl. Modell 3), andererseits aber auch durch Gleichsetzung der bedingten Effekte von *C* innerhalb der Stufen von *A* formulieren. Die Beziehung zwischen Wechselwirkungshypothesen in log-linearen Modellen und Parametergleichheitshypothesen in MVB-Modellen ist völlig analog zu der Beziehung zwischen Interaktionshypothesen in mehrfaktoriellen ANOVA-Modellen und Hypothesen über einfache (oder einfache einfache) Haupteffekte eines Faktors innerhalb der Stufen anderer Faktoren (vgl. Bredenkamp, 1980, S. 24 ff). Genauso wie es bei der Analyse mehrfaktorieller Versuchspläne manchmal sinnvoll sein kann, auf Standard-ANOVA-Modelle mit Haupt- und Wechselwirkungseffekten zugunsten von Modellen mit einfachen oder einfachen einfachen Haupteffekten[68] zu verzichten, so kann es bei der Analyse von Kontingenztafeln manchmal sinnvoll sein, statt eines log-linearen Modells ein analog zu Abbildung 6.1 konstruiertes MVB-Modell zugrunde zu legen. Dies ist etwa dann der Fall, wenn ein Training oder eine Therapie evaluiert wurde und *C* den Erfolg vs. Mißerfolg der entsprechenden Intervention indi-

[68]Ein „einfacher Haupteffekt" einer unabhängigen Variablen in einem mehrfaktoriellen Versuchsplan ist ein bedingter Haupteffekt innerhalb einer Stufe eines anderen Faktors. Ein „einfacher einfacher Haupteffekt" ist entsprechend ein bedingter Haupteffekt innerhalb einer Stufenkombination zweier anderer Faktoren (vgl. z.B. Bredenkamp, 1980).

ziert. Ist *A* eine Kontrollvariable (z.B. Geschlecht der Vp) und *B* eine Indikatorvariable für die Art der Intervention (z.B. Training vs. Placebo), dann ist der Nachweis von Wechselwirkungseffekten zwischen *C* und den anderen Variablen kaum informativ: Man weiß nur, *daß* der Interventionserfolg durch *A* und/oder *B* beeinflußt wird, aber nicht *wie*. Nützlicher ist es dann, nach den einfachen einfachen Effekten von *C* innerhalb der Stufen von *A* und *B* zu fragen. So könnte sich beispielsweise ergeben, daß *B* und *C* innerhalb der ersten Stufe von *A* unabhängig sind ($\theta_{1|1,1} = \theta_{1|1,2}$), während in der zweiten Stufe von *A* das Training der Placebo-Behandlung stark überlegen ist ($\theta_{1|2,1} > \theta_{1|2,2}$). Ein vergleichsweise einfach und direkt interpretierbares Ergebnismuster ist bei der Analyse mit log-linearen Standardmodellen nicht zu erwarten.

Neben der Möglichkeit, bedingte Effekte innerhalb der Stufen von Kontrollvariablen zu analysieren, sprechen auch sogenannte „strukturelle Nullzellen" (structural zeros) für eine Analyse von Kontingenztafeln mit MVB-Modellen anstelle von log-linearen Modellen. Eine strukturelle Nullzelle liegt vor, wenn eine bestimmte Kombination der Variablenausprägungen *prinzipiell* ausgeschlossen ist.[69] So hat man es etwa mit strukturelle Nullzellen zu tun, wenn neben dem Familienstand auch die Berufstätigkeit erhoben wird und katholische Priester als Berufsgruppe in der Kontingenztafel auftauchen: Die Zellen, die für nichtledige Priester vorgesehen sind, müssen zwangsläufig leer sein. Selbstverständlich dürfen strukturelle Nullzellen nicht in die Parameterschätzung und Modellprüfung eingehen, da dies artifiziell verzerrte Ergebnisse zur Folge hätte. Zwar lassen sich log-lineare Modelle — z.B. sogenannte Quasi-Unabhängigkeitsmodelle — auch für unvollständige Tafeln formulieren und (mit korrigierten Freiheitsgraden) unter bestimmten Bedingungen auch schätzen und testen (Bishop et al., 1975; Kapitel 5); jedoch fallen dann bestimmte Parameter oder ganze Parametergruppen zwangsläufig aus der Modellgleichung heraus, was die Interpretation der Ergebnisse zusätzlich erschwert (Andersen, 1990, Kapitel 7).

Viel unproblematischer ist die Analyse unvollständiger Kontingenztafeln auf der Basis von MVB-Modellen. Man kann hierbei etwa die bedingten Effekte des Familienstandes nur innerhalb der Berufsgruppen betrachten, die mit allen Ausprägungen der Variable „Familienstand" vereinbar sind. Umgekehrt kann man die Effekte aller Berufsgruppen (inklusive Priester) innerhalb der Gruppe lediger Vpn betrachten und für die restlichen Daten (Berufsgruppen ohne katholische Priester sowie alle nichtledigen Probanden) beispielsweise ein Unabhängigkeitsmodell zwischen Berufstätigkeit und Familienstand formulieren. Denkbar

[69] Hiervon abzugrenzen ist eine „zufällige Nullzelle" (random zero), die zwar eine Wahrscheinlichkeit größer als null hat, jedoch aufgrund eines zu kleinen Stichprobenumfangs in der Stichprobe nicht besetzt ist. Strukturelle Nullzellen sind a priori bekannt, zufällige Nullzellen dagegen erst nach der Datenerhebung.

sind auch ganz andere Ansätze, die alle nichtleeren Zellen der Kontingenztafel als Ausprägungen einer einzigen kategorialen Variablen auffassen und die multivariate Kontingenztafelstruktur komplett ignorieren.

Zusammenfassend kann festgehalten werden, daß die meisten inhaltlich interessierenden log-linearen Modelle alternativ auch durch datenäquivalente MVB-Modelle ersetzt werden können. Die Analyse von Kontingenztafeln mit MVB-Modellen hat darüber hinaus besondere Vorteile, wenn konditionale Effekte von Kernvariablen innerhalb der Stufen von Kontrollvariablen interessieren, oder wenn unvollständige Kontingenztafeln mit strukturellen Nullzellen vorliegen. Insgesamt sind MVB-Modelle verglichen mit log-linearen Modellen der flexiblere Analyserahmen für multivariate kategoriale Datensätze. Man verzichtet offensichtlich auf nicht viel, wenn man Kontingenztafeln generell auf der Basis geeignet konstruierter MVB-Modellen untersucht. Die Informationen, die hierarchische log-lineare Analysen liefern, können fast immer auch durch MVB-Analysen aufgedeckt werden (vgl. Tabelle 6.1). Umgekehrt eröffnen MVB-Modelle neue Möglichkeiten der Kontingenztafelanalyse, die Anwendern, welche sich auf hierarchische log-lineare Modelle beschränken, verschlossen bleiben.

6.2 Genetische Modelle

Theorien der geschlechtlichen Vererbung eignen sich hervorragend zur Formalisierung in Termini von MVB-Modellen. Wissenschaftshistorisch betrachtet dürfte die Genetik die ersten MVB-Modelle überhaupt hervorgebracht haben („statistical genetics"). So ist beispielsweise schon Bernsteins (1925) Blutgruppen-Vererbungsmodell ein MVB-Modell im Sinne der Definition (3.4-7), wenn auch natürlich die Bezeichnung „Verarbeitungsbaummodell" weder von Bernstein noch von anderen statistisch orientierten Genetikern benutzt wurde (Batchelder, 1999; Hu & Batchelder, 1994). Tatsächlich ist diese Bezeichnung im genetischen Kontext auch irreführend. Treffender wäre es, genetische MVB-Modelle als „multinomiale Abstammungsbaummodelle" zu bezeichnen: Modelliert wird, mit welchen Wahrscheinlichkeiten Individuen einer Population die verschiedenen Allele eines Gens (oder auch mehrerer Gene) mütterlicherseits ererben (1. Ebene des „Abstammungsbaumes") und mit welchen bedingten Wahrscheinlichkeiten Allele väterlicherseits ererbt werden, gegeben daß mütterlicherseits ein bestimmtes Allel vererbt wird (2. Ebene des „Abstammungsbaumes"). Prinzipiell können entsprechende Modelle natürlich auch auf Sequenzen von Generationen verallgemeinert werden.

Ein Beispiel ist das erwähnte Modell von Bernstein (1925), bei dem es um die Vererbung der A-, B- und 0-Allele des Blutgruppengens geht. Konkret zielt das Modell darauf ab, eine Schätzung der Allel-Häufigkeiten in einer Population zu

ermöglichen, wobei neben den üblichen Annahmen zur Beziehung zwischen den sechs Blutgruppengenotypen (AA, AB, A0, BB, B0, 00) und den vier beobachtbaren Phänotypen (A, B, AB, 0) die Zusatzannahmen zugrunde gelegt werden, daß (a) die mütterlicherseits und väterlicherseits ererbten Allele stochastisch unabhängig sind („random mating"), (b) die verschiedenen Allele in der weiblichen und der männlichen Population gleiche relative Häufigkeiten aufweisen (keine Plazierung des Gens auf den Geschlechtschromosomen) und (c) ein sogenanntes Hardy-Weinberg-Äquilibrium vorliegt, so daß die Allelhäufigkeiten über die Generationen hinweg stabil bleiben (zur Herleitung von Hardy-Weinberg-Äquilibria in Abhängigkeit von der Fitness der verschiedenen Phänotypen vgl. z.B. Sutton & Wagner, 1985).

6.2.1 Ein Modell zur Schätzung von Vaterschaftswahrscheinlichkeiten

Zur Veranschaulichung der MVB-Formalisierung von genetischen Fragestellungen soll ein recht einfaches Problem betrachtet werden, das unmittelbar an Bernsteins Modell anknüpft. Trotz ihrer Einfachheit ist die Lösung des Problems insbesondere für evolutionspsychologische Ansätze sehr nützlich und meines Wissens auch neu. Sie betrifft ein zentrales Problem vieler evolutionspsychologischer Theorien, nämlich die Abschätzung der Wahrscheinlichkeit, daß — bezogen auf eine bestimmte Population — der nominelle Vater eines Kindes, d.h. der von der Mutter benannte Kindesvater (im folgenden Präsumtivvater genannt), auch der biologische Vater (im folgenden: Vater) des Kindes ist.[70] Biologische Fakten (z.B. in Einzelfällen nachweisbare Phänotyp-Inkompatibilitäten zwischen Präsumtivvater und Präsumtivkind) zeigen klar, daß diese Vaterschaftswahrscheinlichkeit p im allgemeinen kleiner als 1 ist (z.B. Ashton, 1980). Zahlreiche evolutionsbiologische und evolutionspsychologische Theorien gehen jedoch von der weitergehenden Annahme aus, daß p nicht nur marginal, sondern bedeutsam kleiner als 1 ist. Folglich versprechen biologische oder psychologische Mechanismen, die direkt oder indirekt darauf abzielen, die Investition der Ressourcen eines Mannes in eigene Kinder zu maximieren und zugleich Ressourcen von nicht eigenen Kindern fernzuhalten, einen beträchtlichen Fitneß-Vorteil gegenüber alternativen Mechanismen, die keine Differenzierung zwischen eigenen und nicht eigenen Kindern vorsehen. Letzteres würde nämlich bedeuten, daß auf die eigenen Kinder um so

[70] Ich betrachte nur den weitaus häufigsten Ein-Mann-Fall („one man case"), d.h. den Fall, daß für ein Kind genau ein Präsumtivvater benannt wird. Der zu beschreibende Modellansatz läßt sich — wie deutlich werden wird — prinzipiell auch auf verschiedene „multiple men cases" erweitern, verlangt dann jedoch bestimmte Verallgemeinerungen, insbesondere auf mehr als eine Vaterschaftswahrscheinlichkeit. Da es hier nur auf die Illustrierung des Prinzips ankommt, sollen entsprechend komplexe Modelle ausgeklammert werden.

weniger Ressourcen entfallen, je kleiner *p* ist, was wiederum deren Fitneß beeinträchtigen würde.

Ein möglicher psychologischer Mechanismus, der diesem übergeordneten Ziel dient, könnte die maskuline Form der Eifersucht sein. Es ist vermutet worden, daß Eifersucht bei Männern anders als bei Frauen primär darauf abzielt, sexuelle Kontakte außerhalb der Ehe bzw. außerhalb der Partnerschaftsbeziehung zu verhindern, während „emotionale Untreue" aus der Sicht des Mannes von untergeordneter Bedeutung ist. Eine dauerhafte emotionale Beziehung der Partnerin zu einem anderen Mann, die nicht mit sexuellen Kontakten einhergeht, wäre demnach für die betroffenen Männern vergleichsweise weniger bedrohlich als eine kurze, unverbindliche sexuelle Beziehung, aus der ein genetisch fremdes Kind hervorgehen könnte. Für Frauen sollte es dagegen genau umgekehrt sein: Für sie ist es unter den Bedingungen einer Jäger-Sammler-Kultur, die — wie man vermutet — die größte Phase der Evolutionsgeschichte der Menschheit geprägt hat (Cosmides & Tooby, 1987; Tooby & Cosmides, 1992), bedrohlicher, wenn der Mann als Konsequenz aus einer dauerhaften emotionalen Bindung an eine andere Frau möglicherweise nicht mehr als Ressourcengeber zur Verfügung steht. Eine unverbindliche sexuelle Affäre des Partners ist für die betroffene Ehefrau bzw. Partnerin vergleichsweise weniger konsequenzenreich. Obwohl diese evolutionspsychologische Eifersuchtstheorie nicht unumstritten ist (DeSteno & Salovey, 1996; Harris & Christenfeld, 1996), gibt es doch eine Reihe von überzeugenden empirischen Argumenten, die für die Theorie sprechen (Buss, Larsen & Westen, 1996; Buss, Larsen, Westen & Semmelroth, 1992; Buunk, Angleitner, Oubaid & Buss, 1996).

Eng verknüpft mit der Eifersuchtsproblematik ist Gewalt von Männern gegen Frauen und auch Kinder. Wie man aufgrund der geschilderten evolutionspsychologischen Überlegungen vermuten muß, ist Gewalt gegen Frauen oft mit vermuteten oder tatsächlichen sexuellen Beziehungen zu anderen Männern gekoppelt und hat das erklärte Ziel, derartige Beziehungen zu verhindern oder zu beenden (Daly & Wilson, 1988; Daly, Wilson & Weghorst, 1982; Wilson & Daly, 1996). Soweit sich Gewalt von Männern gegen Kinder richtet, sind Stiefkinder, Pflegekinder und andere genetisch nicht oder wenig verwandte Kinder im Vergleich zu eigenen Kindern weit überproportional betroffen (Daly & Wilson, 1987; 1988; 1996).

Das Thema Gewalt ist wiederum eng mit der Inzestproblematik verknüpft. Inzest scheint aus evolutionspsychologischer Sicht zunächst schwer erklärbar zu sein, da Inzest — sofern genetisch sehr nahe Verwandte betroffen sind — bekanntermaßen fitneßabträglich ist (Thornhill, 1990). Tatsächlich sind jedoch von Inzest weit überproportional Stiefkinder oder Adoptivkinder des Täters betroffen, weniger dagegen blutsverwandte Kinder. Sofern sexuelle Beziehungen zwischen genetischen Präsumtivverwandten vorliegen, sind solche, bei denen eine genetische Verwandtschaft sehr sicher ist, seltener betroffen als jene, bei denen die genetische

Verwandtschaft tendenziell unwahrscheinlicher ist. So sind etwa Beziehungen zwischen Verwandten mütterlicherseits seltener als zwischen Verwandten väterlicherseits (Thornhill, 1990; 1991; Thornhill & Thornhill, 1987; Welham, 1990).

Schließlich muß auch noch auf biologische und psychologische Mechanismen der intrauterinen Spermienkonkurrenz („sperm competition") verwiesen werden, die aus evolutionstheoretischer Perspektive nur dann einen Sinn ergeben, wenn man sie als adaptive Mechanismen betrachtet, die der Maximierung der Vaterschaftswahrscheinlichkeit *p* dienen (Baker & Bellis, 1995).

Für die Bewertung des adaptiven Wertes der o.g. Mechanismen in einer bestimmten Population ist die Kenntnis der Vaterschaftswahrscheinlichkeit *p* unumgänglich. Offenkundig kann man sich aber bei der Abschätzung von *p* nicht darauf beschränken, die relative Häufigkeit von Phänotyp-Inkompatibilitäten zwischen Präsumtivvater und Kind auszuzählen. Beim ABO-Blutgruppensystem mit den vier Phänotypen A, B, AB und 0 gibt es beispielsweise a priori $4^3 = 64$ verschiedene Phänotyp-Kombinationen der Blutgruppen von Präsumtivmutter, Präsumtivvater und Kind. Ist die Präsumtivmutter die biologische Mutter des Kindes — wovon im folgenden immer ausgegangen werden soll — so scheiden insgesamt acht der 64 Blutgruppenkombinationen aus, nämlich all jene, die in Tabelle 6.2 einfach umrandet sind. Darüber hinaus liegen Phänotyp-Inkompatibilitäten zwischen Präsumtivvater und Präsumtivkind vor, wenn Blutgruppenkombinationen vorkommen, die in Tabelle 6.2 doppelt umrandet sind. Insgesamt sind das 16 zusätzliche Blutgruppenkombinationen, so daß letztlich nur $64 - 8 - 16 = 40$ mögliche Blutgruppenkombinationen von Mutter, Vater und Kind übrig bleiben (vgl. z.B. Essen-Möller, 1938, S. 32f.).

Tabelle 6.2 faßt die Häufigkeiten der verschiedenen Blutgruppenkombinationen in einer Stichprobe von Mickey, Gjertson und Terasaki (1986) zusammen. Es handelt sich um insgesamt 1393 sukzessive „Ein-Mann-Fälle" (vgl. Fußnote 70) mit kaukasischer Abstammung, die in der Paternity Evaluation Section des UCLA Tissue Typing Laboratory untersucht wurden. Betont sei, daß dies keine repräsentative Stichprobe aus der Gesamtbevölkerung der betreffenden Region ist. Vielmehr handelt es sich um Fälle, bei denen eine Begutachtung der Vaterschaft angeordnet oder gewünscht wurde, weil der Präsumtivvater bezweifelt, der wirkliche Vater des betreffenden Kindes zu sein. Dies läßt vermuten, daß *p* hier deutlich kleiner als in der Gesamtbevölkerung ist.

Wie Tabelle 6.2 zeigt, führt keiner der 1393 Fälle zu Blutgruppeninkonsistenzen zwischen Mutter und Kind; allerdings liegen insgesamt 39 Blutgruppeninkonsistenzen zwischen Mutter, Kind und Präsumtivvater vor. Damit ist klar, daß die Mindestquote von Nichtvätern an den untersuchten Präsumtivvätern $39/1393 = .028$ beträgt. Hieraus läßt sich aber lediglich ableiten, daß $1 - .028 = .972$ eine Obergrenze für eine vernünftige Schätzung von *p* ist. Vermutlich ist das

wahre *p* deutlich kleiner, da Präsumtivväter, die nicht die wirklichen Väter des betreffenden Kindes sind, nicht in jedem Falle inkompatible Blutgruppenkombinationen generieren müssen.

Tabelle 6.2:
Häufigkeiten der verschiedenen AB0-Blutgruppenkombinationen von Mutter, Präsumtivvater und Kind in der Stichprobe von Mickey, Gjertson und Terasaki (1986, S. 126/127). Blutgruppeninkompatibilitäten zwischen Mutter und Kind sind einfach umrandet und kommen in der Stichprobe nicht vor. Inkompatibilitäten zwischen den Blutgruppen zwischen Präsumtivvater, Mutter und Kind sind doppelt umrandet.

Blutgruppen von		Blutgruppe des Präsumtivvaters			
Mutter	Kind	0	A	B	AB
0	0	239	127	38	3
	A	11	124	3	16
	B	4	3	22	7
	AB	0	0	0	0
A	0	93	56	8	0
	A	173	203	20	16
	B	2	2	18	5
	AB	4	0	15	2
B	0	33	11	2	1
	A	2	6	1	1
	B	25	11	12	1
	AB	1	14	0	0
AB	0	0	0	0	0
	A	12	15	2	0
	B	8	8	1	0
	AB	2	6	1	3

Wie kann man zu einer besseren Schätzung von *p* gelangen? Ashton (1980) ist im Rahmen einer Familienstudie auf Hawaii so vorgegangen, daß er die Frau-Mann-Kind-Tripel zunächst in „konventionelle Familien" und „Familien mit nichtbiologischem Vater" aufteilte. Innerhalb der letzten Gruppe wurde nun ausgezählt, wie häufig inkompatible Blutgruppenkombinationen vorlagen. Diese relative Häufigkeit inkompatibler Blutgruppenkombinationen wurde dann benutzt, um die Anzahl inkompatibler Blutgruppen bei den konventionellen Familien zu einem approximativen Schätzer für *p* zu adjustieren.

Die Methode von Ashton basiert auf der problematischen Annahme, daß der Anteil inkompatibler Blutgruppenkombinationen bei Mutter-Nichtvater-Kind-Tripeln unabhängig davon ist, ob es sich um eine konventionelle Familie (mit unbekannterweise nichtbiologischem Vater) oder eine Familie mit bekanntermaßen nichtbiologischem Vater handelt. Stimmt diese Annahme, dann ist der Ashton-Schätzer brauchbar. Zu vermuten ist aber, daß die Annahme zumindest nicht exakt stimmt, da in einigen Fällen Phänotyp-Inkompatibilitäten erst dazu führen dürften, daß eine Familie als „Familie mit nichtbiologischem Vater" klassifiziert wird. Dies würde bedeuten, daß der Anteil inkompatibler Blutgruppenkombinationen in dieser Population größer ist als in der Population konventioneller Familien mit unbekannterweise nichtbiologischem Vater, was eine Überschätzung von *p* für die letztgenannte Population zur Folge hätte.

Eine Alternative besteht darin, nicht nur einen, sondern eine Vielzahl genetischer Marker simultan zu untersuchen. Dadurch steigt die Wahrscheinlichkeit, daß ein nichtbiologischer Vater in mindestens einem der Marker eine Phänotyp-Inkompatiblität generiert (vgl. Li & Chakravarti, 1985, S. 816). Mickey et al. (1986) sind diesen Weg gegangen, indem sie zusätzlich zum AB0-System auch das HLA-, das Rh- und das MNS-System in ihrer Stichprobe untersuchten. Durch das AB0-System allein werden 39 Präsumtivväter als Väter ausgeschlossen (s.o.), durch die drei anderen genetischen Marker kommen weitere 308 Ausschlüsse hinzu, so daß insgesamt mindestens 347 Präsumtivväter der Gesamtstichprobe (ein Anteil von 347 / 1393 = .249) nichtbiologische Väter sind. Damit erhalten wir zwar eine realistischere Schätzung von *p* in der hier vorliegenden Gesamtpopulation, aber wir haben immer noch das Problem, daß 1 – .249 = .751 nur eine Obergrenze für eine vernünftige Schätzung von *p* ist, da mit einiger Sicherheit immer noch einige nichtbiologische Väter unentdeckt geblieben sind, wenn auch vermutlich nur wenige. Durch „genetic fingerprinting" ließe sich noch mehr Sicherheit erzielen (Baker & Bellis, 1995, S. 200), doch damit wären zusätzlicher zeitlicher Aufwand und weitere Kosten verbunden.

Im folgenden soll gezeigt werden, daß man auf einfachem Wege mit unproblematischen Zusatzannahmen zu einer asymptotisch erwartungstreuen, konsistenten und effizienten Schätzung von *p* auf der Grundlage eines einzigen genetischen Markers (z.B. der AB0-Blutgruppe) kommen kann. Dies mag überraschen, da einzelne genetische Marker — speziell die AB0-Blutgruppe — bekanntlich recht unsichere Indikatoren für Vaterschaftsentscheidungen in Einzelfällen sind. Doch darum geht es hier nicht: *p* soll lediglich für eine bestimmte Population von Individuen vernünftig geschätzt werden und das verlangt — wie deutlich werden wird — nicht zwangsläufig ein nahezu fehlerfreies Entscheidungsverfahren für Einzelfälle. Die Methode berücksichtigt sogar solche scheinbar hoffnungslosen Fälle korrekt, bei denen der Präsumtivvater nicht der Vater des Kindes ist, den-

noch aber zufällig dieselbe Blutgruppe wie der biologische Vater aufweist. Damit widerlegt sie die Aussage von Baker und Bellis (1995), daß es hierfür keine Kontrollmöglichkeit geben kann: „(...) there is obviously no control for the latter problem" (a.a.O., S. 200).

Um die vorzustellende Methode anzuwenden, braucht man nur eine möglichst große Stichprobe der Blutgruppen von Mutter-Präsumtivvater-Kind-Tripeln, wie sie beispielsweise in Tabelle 6.2 dargestellt ist. Legt man nun die bereits geschilderten Annahmen von Bernsteins (1925) Blutgruppenmodell zugrunde und ergänzt diese um die unproblematische Annahme, daß die Vaterschaftswahrscheinlichkeit p unabhängig von der Genotyp-Kombination von Mutter und Präsumtivvater ist, so ergibt sich unmittelbar das in Abb. 6.2 illustrierte MVB-Modell.

In Abbildung 6.2 ist nur ein Teil des gesamten Ereignisbaummodells dargestellt, nämlich der Teil, der sich auf den AA-Genotyp der Mutter bezieht. Das Gesamtmodell erhält man, wenn man analoge Teilmodelle für die anderen Allelkombinationen der Mutter ergänzt. Der Grundgedanke ist ganz einfach: Es werden zunächst alle denkbaren Genotypkombinationen von Mutter und Präsumtivvater *mit Berücksichtigung der Allelreihenfolge* betrachtet, so daß beispielsweise die Reihenfolgen A0 und 0A der Mutter oder des Präsumtivvaters separat aufzuführen sind. Das erste Allel ist immer das an das Kind vererbte (im Falle des Präsumtivvaters: gegebenenfalls vererbte) Allel, das zweite das nicht vererbte Allel. Insgesamt ergibt das $3 \cdot 3 = 9$ mögliche Allelkombinationen bei der Mutter und genauso viele beim Präsumtivvater, so daß durch Kreuzklassifikation $9 \cdot 9 = 81$ Äste des Modells resultieren. Im Einklang mit Bernsteins (1925) Modell wird davon ausgegangen, daß die vier Allele von Mutter und Präsumtivvater komplett stochastisch unabhängig sind und gleiche Wahrscheinlichkeiten a (für das A-Allel), b (für das B-Allel) und $1-a-b$ (für das 0-Allel) aufweisen. Für ein gegebenes Mutter-Präsumtivvater-Allelquadrupel ist die Wahrscheinlichkeit immer p, daß der Präsumtivvater der Vater des Kindes ist. In diesem Fall (Präsumtivvater = Vater) ergibt sich der Genotyp des Kindes durch Kombination der jeweils ersten Allele von Mutter und Vater. Ist der Präsumtivvater mit Wahrscheinlichkeit $1-p$ nicht der Vater, so ergibt sich der Genotyp des Kindes durch das erste Allel der Mutter und ein „Zufallsallel" des unbekannten Vaters, nämlich mit Wahrscheinlichkeit a ein A-Allel, mit Wahrscheinlichkeit b ein B-Allel und mit Wahrscheinlichkeit $1-a-b$ ein 0-Allel. Jeder der 81 Äste des Modells wird somit in $1 + 3 = 4$ Unteräste aufgesplittet, so daß das Gesamtmodell $81 \cdot 4 = 324$ Äste umfaßt. Die Phänotyp-Kombinationen von Mutter, Präsumtivvater und Kind, in die diese Äste einmünden, ergeben sich unmittelbar aus den entsprechenden Genotypen. Ganz rechts in Abbildung 6.2 ist die Nummer der zugehörigen Beobachtungskategorie angegeben. Man erhält sie aus Tabelle 6.2, indem man die 64 Zellen der Tabelle zeilenweise von oben links bis unten rechts durchnumeriert.

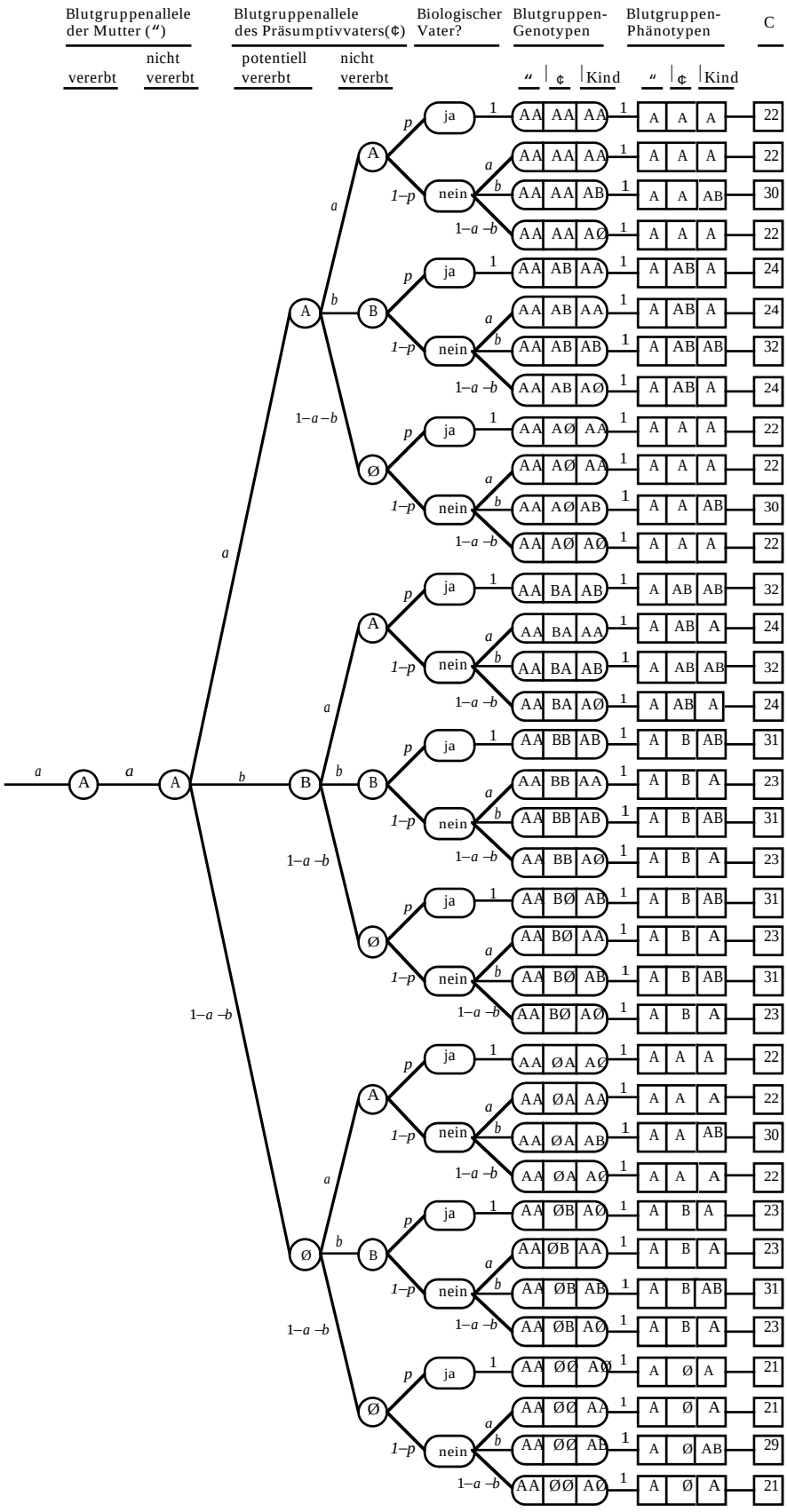

Abbildung 6.2:
Ein MVB-Teilmodell zur Modellierung der AB0-Blutgruppen-Phänotypen von Mutter, Kind und Präsumtivvater. Die A-, B- und 0-Allelwahrscheinlichkeiten in der zugrundeliegenden Population werden mit *a*, *b* und 1–*a*–*b* notiert. *p* ist die Wahrscheinlichkeit, daß der Präsumtivvater der tatsächliche Vater des Kindes ist. Weitere Annahmen des Modells werden im Text erläutert.

6.2.2 Prüfung des Modells zu Vaterschaftswahrscheinlichkeiten

Das MVB-Modell gemäß Abbildung 6.2 wurde mittels AppleTree (Rothkegel, 1998) an die in Tabelle 6.2 reproduzierten Daten von Mickey et al. (1986) angepaßt. Da es sich um ein nichtbinäres MVB-Modell handelt, mußte es zunächst in ein datenäquivalentes binäres Modell überführt werden, um mit AppleTree analysierbar zu sein (vgl. Abschnitt 3.5). Dies kann erreicht werden, indem im ersten Schritt Allel A gegen die beiden anderen Allele kontrastiert wird; erst auf der nächsten Ebene wird jeweils B gegen 0 kontrastiert. In Abschnitt 4 des Anhangs ist die EQN-Datei des binär reparametrisierten MVB-Modells sowie die MDT-Datei mit den Daten gemäß Tabelle 6.2 wiedergegeben (abzüglich der acht Kategorien, die durch die angenommene Mutter-Kind-Beziehung ausgeschlossen werden).

Die Log-Likelihood-Modellanpassungsstatistik fällt mit $G^2 = 53.19$ bei $df = 64 - 8 - 3 - 1 = 52$ (Kategorien insgesamt minus Ausschlüsse durch Mutter-Kind-Beziehung minus freie Parameter minus 1) nahezu auf den Erwartungswert der $\chi^2_{(52)}$-Verteilung bei Modellgültigkeit ($E(\chi^2_{(df)}) = df$). Die hohe Anzahl von Freiheitsgraden des Modelltests läßt vermuten, daß auch stark liberalisierte Versionen des dargestellten Modells noch identifizierbar und testbar gewesen wären. Doch offensichtlich besteht kein Anlaß, das Modell abzuschwächen. Zwar ist die Stärke des Modelltests bei $N = 1393$, $df = 52$ und $\alpha = .05$ nur für mindestens moderate Modellverletzungen ($w = .2$, approx. $1-\beta = 0.99$), nicht aber für kleine Modellverletzungen ($w = .1$, approx. $1-\beta = 0.36$) ausreichend (Buchner, Faul & Erdfelder, 1992; Erdfelder et al., 1996, vgl. Abschnitt 4.3.3); doch fallen im gegebenen Fall sowohl G^2 als auch Anpassungsindizes wie AIC (hier: –50.81) und BIC (hier: –323.25) so eindeutig aus, daß kein Zweifel an der empirischen Adäquatheit des Modells bestehen kann (vgl. Gleichungen (4.3.6-2) und (4.3.6-3)).

Die ML-Parameterschätzung für p liegt mit .745 (90%-Konfidenzintervall: [.697, .793]) — wie zu erwarten — knapp unter dem von Mickey et al. angegebenen Schätzwert von .751. Die erwartete Anzahl von Vätern in der Gesamtstichprobe ist also mit $.745 \cdot 1393 = 1037.79$ um etwa acht kleiner als die von den Autoren nach dem Ausschlußverfahren ermittelte Anzahl. Die ML-Parameterschätzungen für die Allelhäufigkeiten betragen $\hat{a} = .273$, $\hat{b} = .078$ und $1 - \hat{a} - \hat{b} = .650$ und entsprechen damit fast exakt den von Mickey et al. (1986, S. 125, Tabelle 1) aus den Blutgruppen-Randverteilungen geschätzten Werten (.273, .070 und .657).

Durch Anwendung des vorgestellten Modells auf repräsentative Blutgruppenstichproben aus bestimmten geographischen Regionen oder Kulturen kann man mit minimalen Aufwand unverfälschte Schätzungen von p erhalten. Bisherige Befunde, die allerdings auf problematischen Annahmen basieren, lassen vermuten, daß p kulturspezifisch zwischen minimal .70 und maximal .986 variieren

kann (vgl. Baker & Bellis, 1995, S. 200; Welham, 1990, S. 101f). Zu bedenken ist, daß $1-p$ nicht unmittelbar auf die Verbreitung von sexueller Untreue in einer Kultur schließen läßt (Baker & Bellis, 1995, S. 200). Dennoch wird deutlich, daß kulturspezifisch unterschiedlicher Anpassungsdruck in bezug auf die Maximierung von p besteht. Hat man genaue Schätzungen für p, so lassen sich unter Berücksichtigung kultureller Besonderheiten (z.B. der Rolle des Mannes in der Familie) Hypothesen über die Verbreitung der o.g. adaptiven Mechanismen und weiterer Reproduktionsstrategien von Männern und Frauen ableiten.

Abschließend sei erwähnt, daß das vorgestellte Modell nicht nur zur Schätzung von p, sondern auch zur Schätzung der *bedingten* Vaterschaftswahrscheinlichkeit, gegeben ein bestimmtes Blutgruppentripel von Mutter, Präsumtivvater und Kind, eingesetzt werden kann. Diese bedingte Wahrscheinlichkeit entspricht den „recruitment probabilities", die im Zusammenhang mit dem LCA-Modell zum Experiment von Levine (1966) in Abschnitt 2.6 bereits vorgestellt wurden. Man betrachtet hierzu alle Äste des Modells, die in das betreffende Blutgruppentripel einmünden *und* dadurch gekennzeichnet sind, daß der Präsumtivvater auch der Vater des Kindes ist. Die Summe der Wahrscheinlichkeitsschätzungen dieser Äste wird dann dividiert durch die Summe aller Wahrscheinlichkeitsschätzungen zum gegebenen Blutgruppentripel, auch denen, bei denen der Präsumtivvater nicht der biologische Vater ist.

Als illustrierendes Beispiel sei die zweite Blutgruppenkombination (0, A, 0) betrachtet (d.h. die Mutter hat Blutgruppe 0, der Präsumtivvater hat A und das Kind hat ebenfalls 0). Wie die EQN-Datei im vierten Abschnitt des Anhangs zeigt, führen insgesamt vier Äste in diese Kategorie, davon nur einer über einen biologischen Vater (Genotyp der Mutter: 00, Genotyp des Vaters: 0A, jeweils 0 wird vererbt). Aus den Wahrscheinlichkeiten dieser vier Äste ergibt sich die bedingte Vaterschaftswahrscheinlichkeit bzw. „recruitment probability" wie folgt:

$$
\begin{aligned}
Pr(\text{Vater} \mid (0,\text{A},0)) &= \frac{Pr(\text{Vater} \cap (0,\text{A},0))}{Pr((0,\text{A},0))} \quad = \\
&= \frac{(1-a-b)^3 \cdot a \cdot p}{(1-a-b)^3 (a \cdot p + (a+(1-a-b)+(1-a-b)) \cdot (a \cdot (1-p))} = \qquad (6.2.2\text{-}1) \\
&= \frac{a \cdot p}{a \cdot p + (2 \cdot (1-b)-a) \cdot (a \cdot (1-p))} \; .
\end{aligned}
$$

Durch Einsetzen der aus den Mickey-Daten ermittelten Parameterschätzwerte erhält man die Schätzung .65, die etwas unter dem Schätzwert für die unbedingte Vaterschaftswahrscheinlichkeit p liegt. Durch analoge Ableitungen kann man relativ mühelos bedingte Vaterschaftswahrscheinlichkeiten für alle anderen Blutgruppenkombinationen der Tabelle 6.2 erhalten. Deutlich wird dabei, daß — von inkompatiblen Blutgruppenkombination mit bedingter Vaterschaftswahrschein-

lichkeit 0 abgesehen — die ermittelten Schätzwerte im allgemeinen im mittleren Bereich liegen und damit nicht sehr informativ sind. Dies ist eine Konsequenz der Tatsache, daß hier lediglich ein einziger genetischer Marker betrachtet wurde. Will man informativere Vaterschaftswahrscheinlichkeiten, so muß der geschilderte Modellansatz auf mehrere simultan zu betrachtende genetische Marker verallgemeinert werden. Auch wenn das System der Modellgleichungen dabei erheblich komplizierter wird, stößt man bei der Verallgemeinerung nicht auf grundsätzlich neue Probleme.

Am Rande sei erwähnt, daß die in der Humangenetik verbreitete Methode der Schätzung individueller Vaterschaftswahrscheinlichkeiten von Essen-Möller (1938) ein Spezialfall der hier dargestellten Methode ist. Essen-Möllers Methode basiert auf dem Bayesschen Theorem und geht von einer Apriori-Vaterschaftswahrscheinlichkeit von 0.5 aus (Aickin, 1984, S. 906). Setzt man nun beispielsweise in die o.g. Formel (6.2-1) für *p* den Wert 0.5 ein, so ergibt sich numerisch dasselbe Resultat wie es nach Essen-Möllers Methode resultieren würde (vgl. Essen-Möller, 1938, S. 32f). Dies gilt analog für alle anderen denkbaren Phänotypkonstellationen. Die hier favorisierte Methode der „recruitment probabilities" hat jedoch den Vorteil, auf die mehr oder minder willkürliche Festlegung von Apriori-Vaterschaftswahrscheinlichkeiten, die Essen-Möllers Methode zwingend voraussetzt, verzichten zu können. Man schätzt diese Apriori-Wahrscheinlichkeit vielmehr aus der Grundgesamtheit, welcher der konkrete Einzelfall entstammt.[71] Damit ist die bedingte Vaterschaftswahrscheinlichkeit eindeutig interpretierbar: Von allen Fällen der relevanten Grundgesamtheit (z.B. Begutachtungsfälle in einem humangenetischen Labor), welche die im konkreten Einzelfall vorliegende Phänotypkonstellation aufweisen, gibt die betreffende Wahrscheinlichkeit den Anteil biologischer Väter an den Präsumtivvätern an (ähnlich: Mickey et al., 1986). Aussagen über andere Grundgesamtheiten sind nicht intendiert und ohne entsprechende Stichprobendaten auch nicht möglich. Durch diese klare und unzweideutige Interpretierbarkeit der geschätzten statistischen Größe sind eine Reihe von Einwänden nicht anwendbar, die in den letzten 15 Jahren gegen Essen-Möllers Methode erhoben wurden (Aickin, 1984; Li & Chakravarti, 1985). Auch der Einwand, daß der Schätzung der bedingten Vaterschaftswahrscheinlichkeit ungeprüfte Annahmen zugrunde liegen (Aickin, 1984), greift nicht mehr, denn das verwendete statistische Modell ist testbar und — zumindest im Falle der Daten von Mickey et al. (1986) —

[71] Die von Li und Chakravarti (1985, S. 815f) vorgeschlagene Methode leistet ebenfalls eine Schätzung der Apriori-Vaterschaftswahrscheinlichkeit, setzt aber voraus, daß die Exklusionswahrscheinlichkeiten von Nonvätern bei den verwendeten genetischen Markern bekannt sind. Dies ist problematisch, denn Exklusionswahrscheinlichkeiten sind keine eindeutig bestimmbaren markerspezifischen Konstanten; vielmehr hängen sie von der Allelverteilung in der relevanten Grundgesamtheit ab. Die hier favorisierten „recruitment probabilities" berücksichtigen die Allelverteilung in der Population automatisch.

auch empirisch adäquat. Selbst Konfidenzintervalle für die Schätzwerte lassen sich problemlos konstruieren, wenn dies erwünscht sein sollte. Damit bietet sich die vorgestellte Methode auch für Einzelfallbegutachtungen an.

6.3 Latent-Class-Modelle

Ebenso wie log-lineare Modelle (vgl. Abschnitt 6.1) gehören auch Latent-Class-Modelle inzwischen zu den Standardanalyseinstrumenten für mehrdimensionale Kontingenztafeln (Formann, 1984; Goodman, 1974; Langeheine, 1988; Langeheine & Rost, 1996; Lazarsfeld & Henry, 1968). In gewisser Weise bilden sie das kategoriale Pendant zur konfirmatorischen Faktorenanalyse für multivariat normalverteilte Daten. Bei der Latent-Class-Analyse (LCA) wird angenommen, daß die kategorialen Variablen Y_1, Y_2, ..., Y_m, welche die m-dimensionale Kontingenztafel aufspannen, innerhalb bestimmter (unbekannter) Subpopulationen — den sogenannten „latenten Klassen" — stochastisch unabhängig sind (Annahme der lokalen stochastischen Unabhängigkeit). Dadurch, daß die Kontingenztafel auf Daten verschiedener latenter Klassen in unbekannten Anteilen beruht, können sich in der beobachteten Kontingenztafel dennoch massive Abhängigkeiten zwischen den Variablen Y_1, Y_2, ..., Y_m, zeigen. Analog zu einem faktorenanalytischen Modell, das Korrelationen zwischen stetig verteilten Zufallsvariablen durch Einführung von Faktoren erklärt, erklärt also ein LCA-Modell Abhängigkeiten zwischen beobachteten kategorialen Variablen durch Einführung latenter Klassen, innerhalb derer die Abhängigkeiten verschwinden.

Wenn X die latente Klassenvariable mit den Ausprägungen $x = 1, 2, ..., c$ bezeichnet (insgesamt c latente Klassen), gilt aufgrund der Annahme der lokalen stochastischen Unabhängigkeit innerhalb einer beliebigen Klasse x für die „lokalen" Zellwahrscheinlichkeiten die Produktregel

$$p(Y_1 = y_1, \ldots, Y_m = y_m | X = x) = \prod_{j=1}^{m} p(Y_j = y_j | X = x) \qquad (6.3\text{-}1)$$

Durch Gewichtung dieser klassenspezifischen Zellwahrscheinlichkeiten mit den latenten Klassenwahrscheinlichkeiten $p(X = x)$ und Aufsummierung über die c latenten Klassen erhält man die LCA-Modellgleichung für die Wahrscheinlichkeit einer beliebigen Zelle der beobachteten Kontingenztafel:

$$p(Y_1 = y_1, \ldots, Y_m = y_m) = \sum_{x=1}^{c} p(X = x) \cdot \prod_{j=1}^{m} p(Y_j = y_j | X = x) \qquad (6.3\text{-}2)$$

Wie man durch Vergleich mit der Modellgleichung (3.4-7) sofort erkennt, bilden LCA-Modelle eine Teilklasse der allgemeinen MVB-Modelle. Anders als

log-lineare Modelle sind LCA-Modelle also nicht nur datenäquivalent, sondern sogar äquivalent zu bestimmten MVB-Modellen. Wir haben es mit insgesamt $G = c \cdot m + 1$ Parametergruppen $\theta_{g,l}$ zu tun, innerhalb derer sich die Parameterwerte, die Elemente des Intervalls [0, 1] sind, zu 1 aufsummieren müssen: Für jede der c latenten Klassen gibt es m bedingte Kategoriewahrscheinlichkeiten $p(Y_j = y_j | X = x)$; zusätzlich sind die latenten Klassenwahrscheinlichkeiten $p(X = x)$ zu berücksichtigen. Jeder Ast des zugehörigen Baummodells entspricht einem Produkt von $m+1$ Parametern, nämlich einer Klassenwahrscheinlichkeit und m bedingten Kategoriewahrscheinlichkeiten (vgl. Gleichung 6.3-2). Jeder Parameter des Modells kommt — sieht man von restringierten LCA-Modellen zunächst einmal ab — entweder einmal oder niemals pro Ast vor, so daß die a- und b-Koeffizienten der Modellgleichung (3.4-7) allesamt entweder den Wert 1 oder den Wert 0 annehmen. Die Konstante $c_{k,j,i}$ ist immer 1. Die Anzahl $I_{k,j}$ der Äste, über die pro Beobachtungskategorie $C_{k,j}$ (hier: pro Zelle der Kontingenztafel) summiert wird, ist immer c, nämlich gerade die Anzahl latenter Klassen.

Sind alle Variablen der beobachteten Kontingenztafel dichotom — liegt also eine 2^m-Kontingenztafel vor — *und* wird ein Modell mit nur $c = 2$ latenten Klassen betrachtet, so sind alle Parameter des Modells binär, und das entsprechende Modell geht in einen Spezialfall des binären MVB-Modells (3.5-1) über.

Eine weitere Möglichkeit, LCA-Modelle einzuordnen, bietet die in Abschnitt 5.4.2.1 eingeführte Klasse der Latent-Class-MVB-Modelle. LCA-Modelle bilden gewissermaßen den einfachsten Spezialfall dieser Modellklasse; man erhält sie aus einem Latent-Class-MVB-Modell, indem man als MVB-Modell innerhalb der latenten Klassen das log-lineare Modell der totalen stochastischen Unabhängigkeit mit klassenspezifischen Parametern wählt (vgl. Abschnitt 6.1).

6.3.1 Ein Modell zur Konzeptidentifikationsprozedur von Levine (1966)

Zur Illustration der Spezialfallbeziehung zwischen MVB-Modellen und LCA soll das LCA-Modell zum Experiment von Levine (1966), welches in Abschnitt 2.5 bereits vorgestellt wurde, als MVB-Modell aufgefaßt und ausgewertet werden. Die Modellbeschreibungsdatei (EQN-Datei), die zur Analyse des beschriebenen LCA-Modells mit Hilfe der Programme MBT oder AppleTree erforderlich ist, ist in Abschnitt 5 des Anhangs als Datei „LCA8KL.EQN” aufgeführt. Da beide Programme binäre MVB-Modelle voraussetzen, das LCA-Modell zur Levine-Prozedur aber wegen der acht latenten Klassenwahrscheinlichkeiten $p(h_j)$, die eine sieben freie Parameter umfassende Parametergruppe bilden, kein binäres Modell ist, müssen die Klassenwahrscheinlichkeit zunächst einmal binär reparametrisiert werden. Dies wurde dadurch erreicht, daß jede Klassenwahrscheinlichkeit wie folgt als Produkt drei binärer Hilfsparameter geschrieben wurde:

$$
\begin{aligned}
p(h_1) &= h_1 \cdot h_2 \cdot h_3 \\
p(h_2) &= h_1 \cdot h_2 \cdot (1 - h_3) \\
p(h_3) &= h_1 \cdot (1 - h_2) \cdot h_4 \\
p(h_4) &= h_1 \cdot (1 - h_2) \cdot (1 - h_4) \\
p(h_5) &= h_1 \cdot h_7 \cdot h_5 \\
p(h_6) &= h_1 \cdot h_7 \cdot (1 - h_5) \\
p(h_3) &= h_1 \cdot (1 - h_7) \cdot h_6 \\
p(h_4) &= h_1 \cdot (1 - h_7) \cdot (1 - h_6)
\end{aligned}
\qquad (6.3.1\text{-}1)
$$

Die Werte bzw. Schätzwerte für die sieben Hilfsparameter h_j, j = 1, ..., 7, sind dabei natürlich nur mittelbar von Interesse. Primär interessieren die Klassenwahrscheinlichkeiten bzw. deren Schätzwerte, die gemäß Gleichungssystem (6.3.1-1) aus (Schätz-)Werten der Hilfsparameter rückgerechnet werden können.

Das Gleichungssystem (6.3.1-1) entspricht 2·2·2 = 8 „Basisästen" des zugehörigen MVB-Modells. Dieses Basisäste müssen dann noch in jeweils in 2^4 = 16 Unteräste für alle mögliche Antwortmuster auf die vier dichotomen Blank-Trial-Aufgaben untergliedert werden. Die klassenspezifischen Parameterprodukte, die diesen 16 Antwortmustern entsprechen, ergeben sich unmittelbar aus den in Tabelle 2.2 zusammengestellten klassenspezifischen Wahlwahrscheinlichkeiten des jeweils rechten Items. Insgesamt hat das Baummodell zum 8-Klassen-LCA-Modell somit 8·16 = 128 Äste, deren Wahrscheinlichkeiten der in Abschnitt 5 des Anhangs angegeben Datei „LCA8KL.EQN" entnommen werden können.

6.3.2 Prüfung des Modells zur Konzeptidentifikationsprozedur

Bei der Evaluierung des vorgestellten Konzeptidentifikationsmodells konnte auf Daten einer Gruppenuntersuchung zurückgegriffen werden, die im Rahmen einer Denkpsychologie-Übung an der Universität Trier im Sommersemester 1998 durchgeführt wurde.[72] Im Kern handelte es sich um eine konzeptuelle Replikation der Studie von Levine (1966), die dem Ziel diente, den Studierenden das Levine-Paradigma zu veranschaulichen und zu erläutern.

Insgesamt 70 Studierende zweier Parallelkurse nahmen an der Gruppenuntersuchung teil. Wie in der Originalstudie von Levine (1966) wurden den Studierenden zunächst die acht möglichen Konzepte erläutert. Aus technischen Gründen wurden nicht dieselben Konzepte wie in Abschnitt 2.5 dargestellt gewählt; der gesamte Aufbau und Ablauf des Experiments bleibt aber hiervon unberührt. Es wurde den Vpn erklärt, daß der Versuchsleiter sich eines dieser Konzepte ausgewählt habe und es die Aufgabe der Vp sei, dieses Konzept zu identifizieren. Dies habe dadurch zu geschehen, das die Vp von zwei Zeichen immer dasjenige

[72]Ich danke den Lehrveranstaltungsleitern, Prof. Dr. K.F. Wender und Dr. M. Steffens, für die Durchführung der Untersuchung und die Überlassung der Rohdaten.

(entweder das linke oder das rechte) anzukreuzen habe, das ihrer Meinung nach unter das Konzept des Versuchsleiters fällt.

Insgesamt 16 Zeichenpaare wurden den Vpn präsentiert. Nach Beantwortung des ersten, sechsten, elften und sechzehnten Paares wurden jeweils Rückmeldung gegeben, ob die Antwort „rechts" oder „links" richtig war. Auf die Antworten zu den verbleiben Zeichenpaaren 2 bis 5, 7 bis 10 und 12 bis 15 wurde jeweils keine Rückmeldung gegeben. Es handelte sich jeweils um die gleiche Sequenz von vier blank trials, die völlig analog zu der Sequenz in den Tabellen 2.1 bzw. 2.2 (Kopfzeile) aufgebaut war. Wenn das LCA-Modell zum Experiment von Levine (1966) zutrifft, sollten die Vpn innerhalb von vier blank trials immer die gleiche Konzepthypothese testen und — bis auf stochastisch unabhängige Flüchtigkeitsfehler mit Wahrscheinlichkeit a_j und b_j (vgl. Tabelle 2.2) — die konzeptkonformen Wahlen treffen.

Der Versuchsleiter hatte sich Konzept 6 ausgedacht. Durch die Rückmeldung nach dem ersten Zeichenpaar werden implizit vier der acht Konzepthypothesen (1, 3, 4 und 7) ausgeschlossen, durch das Feedback nach dem sechsten Zeichenpaar weitere zwei (5 und 8) und durch das Feedback nach dem elften Paar wird zusätzlich noch Hypothese 2 ausgeschlossen. Perfektes Gedächtnis und fehlerfreie Anwendung der Hypothesenfalsifikationsstrategie vorausgesetzt, kann man also noch dem elften Durchgang die korrekte Hypothese 8 im Prinzip kennen und in der letzten Sequenz von blank trials nur noch hypothesenkonforme Antworten geben.

Tabelle 6.3:
Ergebnisse der Anpassung des 8-Klassen-LCA-Modells zum Experiment von Levine (1966) an Daten einer Replikationsstudie.

Schätzer bzw.	Blank Trials des Experiments		
Statistik	2 bis 5	7 bis 10	12 bis 15
Modelltest $G^2_{(7)}$	3.31	9.21	4.89
AIC	-10.69	-4.79	-9.11
BIC	-26.43	-34.53	-24.86
Fehlerrate $\hat{x}$	.067 [.039, .096]	.043 [.022, .063]	.064 [.037, .092]
$\hat{p}(h_1)$	.031	.094	.029
$\hat{p}(h_2)$	.426	.577	.064
$\hat{p}(h_3)$	.028	.031	.033
$\hat{p}(h_4)$	.010	.013	.026
$\hat{p}(h_5)$	.080	.032	.077
$\hat{p}(h_6)$	.146	.143	.632
$\hat{p}(h_7)$	.028	.094	.050
$\hat{p}(h_8)$	.250	.016	.090

Das LCA-Modell mit der von Levine (1966) selbst vorgeschlagenen Annahme homogener Fehlerwahrscheinlichkeiten ($\forall\, j\colon a_j = b_j = x$) erwies sich für alle drei untersuchten Sequenzen von blank trials als adäquat. Wie Zeile 1 der Tabelle 6.3 zeigt, ist die Log-Likelihood-Anpassungsstatistik G^2 bei $df = 16 - 1 - 7 - 1 = 7$ (Zellen der Kontingenztafel – 1 – 7 freie Klassenwahrscheinlichkeiten – 1 freie Fehlerwahrscheinlichkeit x) für alle drei Datensätze kleiner oder kaum größer als der χ^2-Erwartungswert bei Modellgültigkeit. Zwar ist die Teststärke des Modelltests selbst bei mittleren Modellabweichungen ($w = .3$) bei $\alpha = .05$, $N = 70$ und $df = 7$ noch ungenügend (approx. $1-\beta = .40$); aber ähnlich wie schon beim Blutgruppenmodell im letzten Abschnitt ist die Anpassungsgüte in Termini von G^2, AIC und BIC deskriptiv so überzeugend, daß man von der Angemessenheit des Modells für den gegebenen Datensatz ausgehen kann.

Der ML-Schätzer für die Fehlerwahrscheinlichkeit x liegt mit durchschnittlich .06 deutlich über dem von Levine (1996) angegeben Wert .02. Dies mag auf die Gruppenprüfungssituation zurückzuführen sein. Vielleicht noch überraschender ist, daß die geschätzten Hypothesenwahrscheinlichkeiten für Konzepte, die durch die Rückmeldungen des Vl bereits als ausgeschlossen gelten können, dennoch recht hoch sind. So wird im letzten Antwortmuster der Vpn (catch trials 12 bis 15) nur von knapp 2/3 der Vpn die korrekte Konzepthypothese zugrunde gelegt. Dies kann z.B. darauf beruhen, daß die Vpn nicht alle Hypothesen simultan evaluiert haben, sondern immer nur eine aus dem Hypothesenraum zufällig gezogene Konzepthypothese. Es kann auch bedeuten, daß fast alle Vpn das gesuchte Konzept zwar schon kennen, dennoch aber ein anderes Konzept — sozusagen als Falsifikationsabsicherung — sicherheitshalber noch ein weiteres Mal überprüfen.

In jedem Falle dürfte durch das Beispiel demonstriert worden sein, daß MVB-Modelle auch in den Kontexten, in denen sie nicht unmittelbar der Formalisierung sequentieller kognitiver Prozeßmodelle dienen, sinnvolle Aufgaben in der kognitiven Psychologie erfüllen können. Insbesondere beinhalten sie kognitionspsychologisch begründete LCA-Modelle, die aus der „Probabilisierung" deterministischer Modelle durch zusätzliche Meßfehlerannahmen hervorgehen, als Spezialfälle.

6.4 Binomialmischverteilungsmodelle

Finite Binomialmischverteilungen können als Submodelle von Latent-Class-Modellen aufgefaßt werden (Rost & Erdfelder, 1996). Sie ergeben sich aus den beiden Annahmen, daß (a) ein bestimmtes Latent-Class-Modell gilt (vgl. Abschnitt 6.3) und (b) die Verteilung jeder beobachteten kategorialen Variablen Y_j innerhalb der c latenten Klassen eine Binomialverteilung mit klassenspezifischen Parametern ist.

Sinnvoll ist Annahme (b) z.B. dann, wenn jede Variable Y_j Ausprägungen 0, 1, ..., n_j annimmt, die als „Anzahl der Erfolge (oder Mißerfolge) in n_j unabhängigen Versuchen (oder Aufgaben)" interpretiert werden kann. Rost (1985) hat plausibel gemacht, daß man eine solche Interpretation auch (n_j+1)-stufigen Ratingskalen unterlegen kann, so daß das Anwendungsspektrum von Binomialmischverteilungen größer ist als es auf den ersten Blick scheinen mag.

Liegt nur eine beobachtete Variable Y mit Ausprägungen y = 0, 1, ..., n vor, so resultiert aus den obigen Annahmen das Modell einer univariaten finiten Binomialmischung mit der Wahrscheinlichkeitsfunktion

$$\begin{aligned} p(Y=y) &= \sum_{x=1}^{c} p(X=x)\cdot p(Y=y|X=x) = \\ &= \sum_{x=1}^{c} p(X=x)\cdot \binom{n}{y}\cdot \pi_x{}^{y}\cdot (1-\pi_x)^{(n-y)} \ . \end{aligned} \tag{6.4-1}$$

Hierbei ist — wie schon im letzten Abschnitt — $p(X = x)$ die Wahrscheinlichkeit (bzw. die relative Größe) der x-ten latenten Klasse und π_x die Erfolgswahrscheinlichkeit (bzw. Mißerfolgswahrscheinlichkeit) innerhalb der x-ten latenten Klasse (vgl. z.B. Rost & Erdfelder, 1996). Erdfelder (1989; 1993a) hat ein auf dem EM-Algorithmus basierendes Computerprogramm vorgestellt, das Maximum-Likelihood-Analysen dieser Modellklasse ermöglicht.[73] Eingesetzt wurde es bislang z.B. zur Prüfung von Stufentheorien der Entwicklungen, nach denen sich die Lösungswahrscheinlichkeiten für bestimmte Problemlöseaufgaben im Verlaufe der Ontogenese nicht kontinuierlich, sondern in diskreten Schüben verbessern (für weitere Anwendungen siehe z.B. Erdfelder, 1989). Folgt die Anzahl gelöster Probleme für jedes Kind einer Binomialverteilung, so sollte sich die Verteilung der gelösten Aufgaben in einer altersheterogenen Gruppe folglich als finite Binomialmischverteilung darstellen, wobei die Anzahl der latenten Klassen gerade der Anzahl der in der Stichprobe vertreten Entwicklungsstadien entspricht. Für Piagetsche Klasseninklusionsprobleme fiel die Anpassungsgüte des Modells (6.4-1) gut aus, für Umkehrlernprobleme nach Kendler und Kendler (1959) dagegen nicht (Erdfelder, 1990a; 1993a). Die Frage, ob kognitive Entwicklung in diskreten Stufen oder kontinuierlich verläuft, kann demnach nicht für alle kognitiven Fähigkeiten einheitlich beantwortet werden.

Liegen mehrere lokal binomialverteilte Variablen Y_1, ..., Y_m vor und sind diese — wie es das LCA-Modell verlangt — lokal stochastisch unabhängig, so re-

[73] Das Programm BINOMIX erlaubt darüber hinaus ML-Analysen von Binomialmischungen mit einer stetig verteilten mischenden Variablen, nämlich sogenannte Beta-Binomial-Mischungen. Da Beta-Binomial-Mischungen nicht als Spezialfälle von MVB-Modellen darstellbar sind, sollen sie hier ausgeklammert werden.

sultiert aus den obigen Annahmen das von Rost (1985) vorgeschlagene „LCA-Modell für Rating-Daten", nämlich eine multivariate finite Binomialmischung:

$$p(Y_1 = y_1, \ldots, Y_m = y_m) = \sum_{x=1}^{c} p(X = x) \cdot \prod_{j=1}^{m} p(Y_j = y_j | X = x) = \\ = \sum_{x=1}^{c} p(X = x) \cdot \prod_{j=1}^{m} \binom{n_j}{y_j} \cdot \pi_{x,j}{}^{y_j} \cdot (1 - \pi_{x,j})^{(n_j - y_j)} \quad . \tag{6.4-2}$$

Man kann dieses Modell beispielsweise zur Analyse von ratingbasierten Fragebögen einsetzen, wobei die klassenspezifischen „Erfolgswahrscheinlichkeiten" $\pi_{x,j}$ die Tendenz reflektieren, auf dem Item Y_j in Klasse x hohe Ratings zu geben.

Aus den beiden Tatsachen, daß (a) finite Binomialmischungen Submodelle von LCA-Modellen sind und (b) LCA-Modelle zur Familie der MVB-Modelle gehören, folgt nicht unmittelbar, daß finite Binomialmischungen ebenfalls MVB-Modelle sind. Zu belegen ist noch, daß die Restriktion, die die Binomialverteilungsannahme LCA-Modellen auferlegt, im Rahmen von MVB-Modellen darstellbar ist. Dieser Nachweis kann jedoch schnell nachgeliefert werden. Ein Blick auf die Gleichungen (6.4-1) und (6.4-2) zeigt, daß sie der Form (3.4-7) genügen, die für allgemeine MVB-Modelle charakteristisch ist. Die multiplikative Konstante $c_{k,j,i}$ entspricht in diesem Fall dem Binomialkoeffizienten (Gleichung (6.4-1)) bzw. dem Produkt der m Binomialkoeffizienten (Gleichung (6.4-2)). Es liegen $G = m+1$ Parametergruppen vor, von denen maximal eine — nämlich $p(X = x)$ — nichtbinär ist. Werden nur zwei oder gar nur eine latente Klasse spezifiziert ($c = 2$ oder $c = 1$), so reduzieren sich (6.4-1) und (6.4-2) auf Spezialfälle der binären MVB-Modellgleichung (3.5-1).

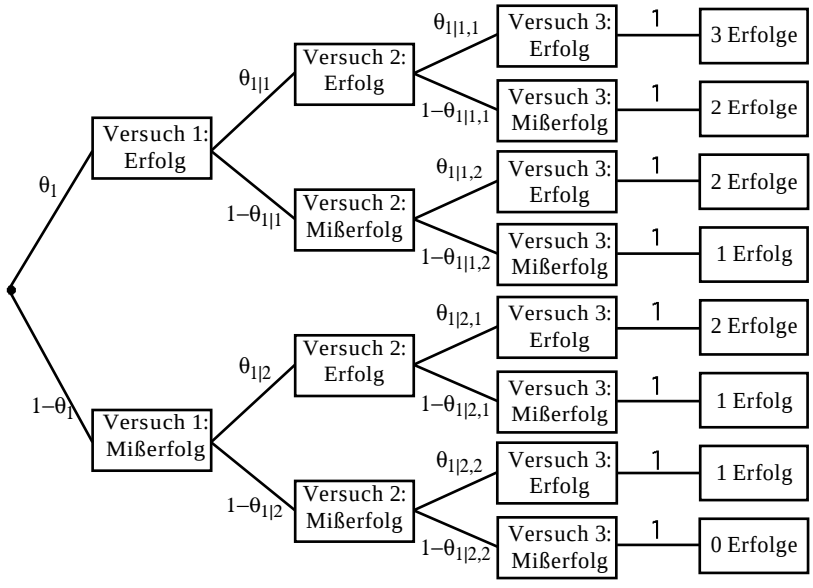

Abbildung 6.3:
Ereignisbaumdiagramm zu einem binären MVB-Modell, aus dem sich durch Gleichsetzung aller Parameter die (3, π)-Binomialverteilung als Spezialfall gewinnen läßt.

Abbildung 6.3 illustriert die Tatsache, daß jede Binomialverteilung als restringiertes binäres MVB-Modell aufgefaßt werden kann, an einem einfachen Beispiel. Es handelt sich um die Ereignisbaumrepräsentation aller möglichen Sequenzen von Erfolgen und Mißerfolgen in $n = 3$ Versuchen. Die empirische Kategorie, in die ein Ast des Modells einmündet, ist jeweils die Anzahl der Erfolge in den drei Versuchen. Im Modell wurde eine Parametrisierung gewählt, die keinerlei Vorannahmen macht: Weder müssen die drei Versuche untereinander stochastisch unabhängig sein noch müssen die Erfolgswahrscheinlichkeiten in den drei Versuchen gleich ausfallen. Es handelt sich folglich um ein Modell mit sieben binären Parametern, das bei nur drei unabhängigen Kategoriewahrscheinlichkeiten stark überparametrisiert ist. Aus der strengstmöglichen Gleichsetzungsrestriktion, nämlich $\theta_1 = \theta_{1|1} = \theta_{1|2} = \theta_{1|1,1} = \theta_{1|1,2} = \theta_{1|2,1} = \theta_{1|2,2} = \pi$, resultiert die ($n = 3, \pi$)-Binomialverteilung als Spezialfall dieses Modells. Völlig analog lassen sich Binomialverteilungen mit anderem n als restringierte binäre MVB-Modelle konstruieren. Man muß lediglich für jeden zusätzlichen Versuch eine weitere Ebene im Ereignisbaumdiagramm vorsehen und die Kategorienanzahl auf $(n+1)$ erweitern.

Die Einordnung von Binomialverteilungsmodellen in die Familie der binären MVB-Modelle bietet einige Vorteile. Abgesehen von den Möglichkeiten der Parameterschätzung und Modellgeltungsprüfung, die sich daraus ergeben, eröffnet diese Einordnung den Blick für die Vielfalt von Modellen, die in Betracht zu ziehen sind, wenn sich das einfache Binomialverteilungsmodell als zu restriktiv erweist. Eine Liberalisierungsmöglichkeit resultiert aus der Annahme, daß die Ergebnisse der n Versuche zwar stochastisch unabhängig sind, die Erfolgswahrscheinlichkeiten in den verschiedenen Versuchen sich jedoch unterscheiden können. Handelt es sich bei den „Versuchen" beispielsweise um Aufgaben mit unterschiedlichen Schwierigkeitsgraden, so ist dieses liberalisierte Binomialmodell recht plausibel. Es ist zwar global identifizierbar, aber saturiert und deshalb isoliert leider nicht testbar (Bamber & van Santen, 1985; 2000). Testbarkeit kann jedoch durch Erweiterung auf ein Mehr-Gruppen-Design mit Gleichheitsrestriktionen zwischen den Gruppen hergestellt werden (vgl. Abschnitt 4.6).

Eine zweite Liberalisierungsmöglichkeit ergibt sich aus der Annahme, daß ein einfaches Binomialmodell zwar innerhalb bestimmter Subpopulationen gilt, die Erfolgswahrscheinlichkeit jedoch zwischen den Subpopulationen variieren kann. Dies entspricht dem finiten Binomialmischverteilungsmodell gemäß Gleichung (6.4-1), das zugleich ein Latent-Class-MVB-Modell im Sinne von Abschnitt 5.4.2.1 ist. Erweitert man derartige Modelle auf verbundene Modelle für mehrere Gruppen, so ergibt sich zusätzlich die Möglichkeit, die Klassengrößenparameter $p(X = x)$ oder Erfolgswahrscheinlichkeiten $\pi_{x,j}$ zwischen den Gruppen Gleichheits-

restriktionen zu unterziehen. Im günstigsten Fall kann so auch eine Kombination der beiden erwähnten Liberalisierungsmöglichkeiten erwogen werden, vorausgesetzt, es lassen sich genügend begründete Gleichsetzungsrestriktionen definieren, so daß die Parameteranzahl nicht größer als $n \cdot K$ wird.

6.4.1 Ein Modell zum Konjunktionsfehler

Worin liegt der Nutzen dieser Modellklasse für die Kognitionspsychologie? Ich möchte dies anhand eines Modells zum sogenannten Konjunktionsfehler illustrieren, der bekanntlich zu den robustesten Effekten im Bereich der Psychologie des probabilistischen Urteilens zählt. Unter dem Konjunktionsfehler versteht man das Phänomen, daß Vpn unter bestimmten Bedingungen die Konjunktion zweier Ereignisse A und B verglichen mit A bzw. B als wahrscheinlicher beurteilen. Für die Wahrscheinlichkeitsurteile gilt also $p(A \cap B) > p(A)$ und $p(A \cap B) > p(B)$, was offensichtlich im Widerspruch zu den Kolmogoroffschen Axiomen für Wahrscheinlichkeitsmaße (als normierte Mengenmaße) steht (vgl. z.B. Krengel, 1991, S. 2–6). Zahlreiche Aufgaben führen bei den meisten untersuchten Probanden zu diesem Konjunktionsfehler. Zuallererst ist hier die bekannte Linda-Aufgabe aus der klassischen Arbeit von Tversky und Kahnemann (1983) zu nennen: Linda wird den Vpn als 31 Jahre alte, alleinstehende, redegewandte und sehr intelligente Akademikerin mit erfolgreichem Abschluß im Fach Philosophie vorgestellt, die als Studentin in Fragen der sozialen Gerechtigkeit politisch stark engagiert war und aktiv an Anti-Kernkraftdemonstrationen teilgenommen hat. Die Vpn werden gebeten, eine Reihe möglicher gegenwärtiger Tätigkeiten von Linda nach ihrer Wahrscheinlichkeit zu ordnen, darunter die Alternativen „Linda ist in der Frauenbewegung aktiv" (im folgenden:A), „Linda ist Bankangestellte" (im folgenden:B) sowie „Linda ist Bankangestellte und in der Frauenbewegung aktiv" (im folgenden:$A \cap B$). Die große Mehrheit der Probanden (mehr als 80%) hält bei dieser Aufgabe die Alternative $A \cap B$ für wahrscheinlicher als B, begeht also den Konjunktionsfehler.

Bereits Tversky und Kahnemann (1983) haben eine Reihe von Erklärungen bzw. Artefakthypothesen zum Konjunktionsfehler überprüft und wieder verworfen, darunter die Hypothese, daß die Aussagen über Lindas gegenwärtige Tätigkeiten sprachlich fehlinterpretiert werden, so daß die logisch intendierte Bedeutung mit der rezipierten Bedeutung nicht übereinstimmt. Trotz Verwendung eindeutigerer sprachlicher Formulierungen resultierte sowohl bei Tversky und Kahneman (1983) als auch in zahlreichen Replikationsstudien immer wieder ein hoher Konjunktionsfehleranteil, und zwar nicht nur bei der Linda-Aufgabe, sondern auch bei inhaltlich veränderten Aufgaben mit gleicher Struktur.

Lediglich eine Modifikation der Aufgabe führte zu einer erheblichen Reduktion des Konjunktionsfehleranteils: Bittet man die Vpn nicht um *Rangbildungen*

subjektiver Wahrscheinlichkeiten, sondern um *Schätzungen absoluter Häufigkeiten*, so fällt die Konjunktionsfehlerrate auf 25% (Tversky & Kahneman, 1983) und zum Teil sogar deutlich darunter (Fiedler, 1988; Hertwig, 1994). Fiedler (1988) hat diesen „Häufigkeitsformateffekt" als erster gründlich und systematisch untersucht, indem er von verschiedenen Konjunktionsfehleraufgaben parallele Versionen im Wahrscheinlichkeitsformat (Rangierung von subjektiven Wahrscheinlichkeiten) und Häufigkeitsformat (Schätzung absoluter Häufigkeiten) erstellt hat. Das Häufigkeitsformat zur Linda-Aufgabe sieht beispielsweise so aus, daß man die Probanden bittet, sich nicht eine Frau, sondern sehr viele (z.B. 2000) Frauen vorzustellen, welche die gleichen Eigenschaften wie Linda haben. Die Vpn müssen dann die absoluten Häufigkeiten für die o.g. Kategorien A, B und $A \cap B$ schätzen. Über die verschiedenen Aufgaben hinweg beobachtete Fiedler (1988) im Durchschnitt wieder 73% Konjunktionsfehler im Wahrscheinlichkeitsformat, jedoch nur 23% Konjunktionsfehler im Häufigkeitsformat. Hertwig (1994) hat gezeigt, daß der Konjunktionsfehleranteil im Häufigkeitsformat sogar unter 10% liegen kann.

Wie ist der Häufigkeitsformateffekt zu erklären? Eine Möglichkeit ist, daß unser kognitives System auf die Repräsentation und Verarbeitung von Häufigkeitsinformationen vorbereitet ist, nicht jedoch auf die Verarbeitung von Informationen zu Einzelfallwahrscheinlichkeiten (Gigerenzer, 1991). Diese Erklärung paßt gut zu zahlreichen anderen Befunden und läßt sich auch evolutionspsychologisch plausibel machen (Cosmides & Tooby, 1996).

Dennoch stößt diese Hypothese an Grenzen. Zunächst einmal bliebe zu erklären, warum bei einigen Aufgaben immerhin noch bis zu 30% Konjunktionsfehler auch im Häufigkeitsformat zu beobachten sind, wenn doch unser kognitives System auf Häufigkeitsverarbeitung spezialisiert ist. Unerklärt bleibt zweitens auch der sehr hohe Konjunktionsfehleranteil im Wahrscheinlichkeitsformat. Wenn das kognitive System Einzelfallwahrscheinlichkeiten schlicht nicht verarbeiten kann, sollten dabei Fehlerraten auf dem Rateniveau resultieren. Die tatsächlichen Fehlerraten sind jedoch viel höher, was die Vermutung nahelegt, daß Wahrscheinlichkeitsinformationen sehr wohl systematisch verarbeitet werden, wenn auch mit wahrscheinlichkeitstheoretisch falschen Ergebnissen. Drittens fällt auf, daß sich Häufigkeitsformat und Wahrscheinlichkeitsformat nicht nur in der Frage nach Häufigkeiten versus Wahrscheinlichkeiten unterscheiden (im folgenden: Frageformat), sondern zusätzlich auch in der Anwortmodalität „Rangbildung" versus „Absolutschätzung" (im folgenden: Antwortformat). Wegen dieser Konfundierung ist zunächst einmal unklar, ob der Vorteil des Häufigkeitsformats wirklich zwingend an Häufigkeitsbeurteilungsaufgaben gekoppelt ist.

Fiedler (1988) hat die Konfundierung von Frage- und Antwortformat diskutiert, hielt deren Effekt aber aufgrund einer Kontrollstudie, in der im Wahrscheinlichkeitsformat Absolutschätzungen verlangt wurden, für nicht sehr bedeutsam.

Jüngst haben jedoch Hertwig und Chase (in press) signifikante Effekte des Antwortformats bei Fragen nach Wahrscheinlichkeiten gefunden: Bei Rangbildungen resultierten deutlich mehr Fehler als bei Absolutschätzungen von Wahrscheinlichkeiten. Das wirft die Frage auf, ob ein ähnlicher Antwortformateffekt auch bei Fragen nach Häufigkeiten zu beobachten ist, so daß möglicherweise das Antwortformat „Rangbildung" versus „Absolutschätzung" insgesamt mehr Varianz in den Fehlerraten aufklärt als das Frageformat. Erdfelder, Bröder und Brandt (1998) sind dieser Frage systematisch nachgegangen, indem sie Frageformat und Antwortformat orthogonal zueinander in einem 2×2-Design variierten. Das in mehreren Experimenten replizierte Ergebnis entsprach den Erwartungen: Die Hauptvarianzquelle der Fehlerraten ist das Antwortformat, d.h. Rangbildungen führen zu mehr Konjunktionsfehlern als Absolutschätzungen. Darüber hinaus gibt es einen zweiten Haupteffekt des Frageformats, wobei der Vorteil der Häufigkeitsfrage im Vergleich zur Wahrscheinlichkeitsfrage bei Absolutschätzungen tendenziell (jedoch nicht signifikant) größer ausfiel als bei Rangbildungsaufgaben. In einigen Experimenten war bei Rangbildungsaufgaben praktisch kein Unterschied mehr zwischen Häufigkeits- und Wahrscheinlichkeitsfragen auszumachen.

Diese Befunde werfen die neue Frage nach den Gründen für den Vorteil der Absolutschätzung im Vergleich zur Rangbildung auf. Im folgenden wird ein differentialpsychologisches Modell der Informationsverarbeitung bei Häufigkeits- und Wahrscheinlichkeitsbeurteilungen vorgestellt, das sowohl den Effekt des Antwortformats als auch den Effekt des Frageformats erklären kann.

Das Modell basiert auf der Annahme, daß nur eine Teilmenge der Gesamtpopulation über den abstrakten mathematischen Wahrscheinlichkeitsbegriff im Kolmogoroffschen Sinne verfügt und ihn *interpretationsunabhängig* korrekt anwenden kann (im folgenden: Gruppe W). „Interpretationsunabhängig" bedeutet hierbei, daß die korrekte Verwendung des Wahrscheinlichkeitsbegriffs nicht an eine bestimmte Interpretation der Wahrscheinlichkeit als subjektive Wahrscheinlichkeit, logische Wahrscheinlichkeit oder asymptotische relative Häufigkeit gekoppelt ist. Ob eine Person zu dieser Gruppe gehört oder nicht sollte im wesentlichen vom Vorwissen bzw. der Ausbildung abhängen. *Wenn* eine Person der Gruppe W eine Aufgabe als Mengenmaßbeurteilungsaufgabe erkennt, dann urteilt sie — abgesehen von Flüchtigkeitsfehlern — im Einklang mit den Kolmogoroffschen Axiomen und macht folglich nur selten Konjunktionsfehler. Hierbei sind Frage- und Antwortformat unerheblich, d.h. die Fehlerrate liegt nicht nur im unteren Bereich, sondern ist auch immer etwa gleich groß. Erkennt eine Person der Gruppe W eine Aufgabe dagegen nicht als Mengenmaßaufgabe, dann wird sie möglicherweise mit höherer Wahrscheinlichkeit Konjunktionsfehler begehen. Möglich ist dies z.B. dann, wenn (a) eine Person nicht explizit auf den mathematischen Charakter einer Aufgabe hingewiesen wird und (b) durch den Kontext eine

nichtmathematische Interpretation des Wahrscheinlichkeitsbegriffs nahegelegt wird. Tatsächlich ist der Wahrscheinlichkeitsbegriff in der Alltagssprache mehrdeutig (Hertwig, 1994; Hertwig & Chase, in press). Er wird beispielsweise häufig auch im Sinne von „plausibel" verwendet. Eine Wahrscheinlichkeitsrangbildungsaufgabe kann daher zulässigerweise auch als Plausibilitätsrangbildungsaufgabe verstanden werden, sofern nicht explizit darauf hingewiesen wird, daß es sich um eine mathematische Aufgabe handelt.

Neben der Gruppe W postuliert das Modell eine Gruppe I. Hierzu zählen Personen, die über den mathematischen Mengenmaßbegriff nicht verfügt und deshalb nur intuitive Urteile abgeben können, unabhängig davon, ob nun auf den mathematischen Charakter der Aufgabe hingewiesen wird oder nicht. Die Konjunktionsfehlerrate hängt bei diesen Personen wesentlich von der inhaltlichen Einbettung und von Kontextfaktoren ab. Sie werden unter sonst gleichen Bedingungen beispielsweise häufiger Konjunktionsfehler bei Rangbildungen und seltener Konjunktionsfehler bei Absolutschätzungen machen, weil letztere extensionale Interpretionen der Aufgabe begünstigen, was Konjunktionsfehlern entgegenwirkt (Tversky & Kahneman, 1983).

Zusätzlich zu den beiden Gruppen W und I nimmt das Modell eine weitere Subpopulation H an, die sich bei Häufigkeitsbeurteilungen genauso verhält wie Gruppe W, jedoch bei Wahrscheinlichkeitsbeurteilungsaufgaben genauso wie Gruppe I. Gruppe H verfügt also über den Kolmogoroffschen Wahrscheinlichkeitsbegriff, sofern Wahrscheinlichkeiten frequentistisch interpretiert bzw. Häufigkeitsbeurteilungen verlangt werden. Mit anderen Interpretationen des mathematischen Wahrscheinlichkeitsbegriffs kann Gruppe H nicht umgehen und urteilt demzufolge (wie Gruppe I) intuitiv.

Wendet man dieses „Drei-Gruppen-Modell des Konjunktionsfehlers" auf Datensätze an, die aus Rangbildungs- bzw. Absolutschätzungsurteilen auf mehrere Konjunktionsfehleraufgaben resultieren, so entspricht es — wenn es um einige eher technische Zusatzannahmen ergänzt wird — einem Binomialmischverteilungsmodell mit zwei latenten Klassen. Als abhängige Variable ist hierbei die Anzahl der Konjunktionsfehler in den n bearbeiteten Aufgaben zu betrachten. Nimmt man unabhängige Antworten an, so muß die Konjunktionsfehleranzahl für Personen der Gruppe W und bei Häufigkeitsbeurteilungen zusätzlich für Personen der Gruppe H binomialverteilt sein mit einer niedrigen Fehlerwahrscheinlichkeit π_1. Die Annahme der Homogenität dieser Wahrscheinlichkeit über die Aufgaben hinweg ist unproblematisch, denn die inhaltliche Einkleidung sollte bei dieser Gruppe im Regelfall keine Rolle spielen.

Problematischer ist die Binomialverteilungsannahme für die Gruppe I und bei Wahrscheinlichkeitsbeurteilungen zusätzlich für Gruppe H. Hier *kann* der Inhalt der Aufgabe eine Rolle spielen. Wenn man allerdings hinreichend parallele

Aufgaben konstruiert, sollte die Binomialverteilungsannahme zumindest als Approximation brauchbar sein. Zu erwarten ist in dieser Gruppe bei Aufgaben des Linda-Typs eine im allgemeinen recht hohe Konjunktionsfehlerwahrscheinlichkeit π_2, die bei Absolutschätzungen (wegen der dadurch nahegelegten extensionalen Interpretation) vielleicht etwas niedriger als bei Rangbildungen ausfällt.

Da Gruppe H sich bei Wahrscheinlichkeitsbeurteilung wie Gruppe I, bei Häufigkeitsbeurteilungen dagegen wie Gruppe W verhält, gibt es für beide Frageformate jeweils nur zwei latente Klassen („fehlerarme" und „fehlerträchtige Vpn"). Allerdings sollte die Wahrscheinlichkeit der latenten Klasse 1 (fehlerarm) bei Wahrscheinlichkeitsbeurteilungsaufgaben kleiner sein als bei Häufigkeitsbeurteilungsaufgaben. Das Frageformat sollte sich also auf die Klassengrößenparameter auswirken, nicht jedoch auf die Konjunktionsfehlerraten, denn die hängen nur von der Gruppenzugehörigkeit ab. Umgekehrt ist zu erwarten, daß das Antwortformat keinen Einfluß auf die Klassengrößenparameter hat, möglicherweise aber auf Konjunktionsfehlerrate innerhalb der Population der „fehlerträchtigen" Vpn.

6.4.2 Prüfung des Modells zum Konjunktionsfehler

Das im letzten Abschnitt geschilderte Modell kann anhand von drei Konjunktionsfehlerexperimenten überprüft werden, die im Rahmen des vom Autor geleiteten Projekts „Multinomiale Modelle kognitiver Täuschungen" durchgeführt wurden.[74] Alle drei Experimente verwendeten das 2×2-Design von Erdfelder et al. (1998), d. h. es wurden jeweils vier komplett randomisierte Versuchsgruppen gebildet, die aus der Kreuzklassifikation des Frageformats (Wahrscheinlichkeitsfrage vs. Häufigkeitsfrage) mit dem Antwortformat (Rangbildung vs. Absolutschätzung) resultierten. In jeder Bedingung mußte jede Vp dieselben fünf Konjunktionsfehleraufgaben bearbeiten, die bereits Erdfelder et al. (1998) verwendet hatten. Neben der oben geschilderten Linda-Aufgabe waren dies vier weitere Aufgaben aus der Sammlung von Fiedler (1988).

Alle drei Experimente dienten dem Zweck, die Effekte verschiedener Maßnahmen zur Reduktion des Konjunktionsfehlers zu explorieren. In Experiment 1 wurden die Vpn (N = 120 Studierende) explizit auf den mathematischen Charakter der Aufgaben hingewiesen und angehalten, Fehler zu vermeiden. Zusätzlich wurde die Teilmengenrelation zwischen den drei Ereignissen A, B und $A \cap B$ dadurch salient gemacht, daß keine weiteren Distraktoralternativen dargeboten wurden. Experiment 2 (N = 80 Studierende) gab keinen expliziten Hinweis auf den mathematischen Charakter der Aufgabe und verwendete stattdessen das von Tversky

[74]Ich danke Martin Brandt für konzeptuelle Mitarbeit und die Programmierung der Experimente und Johannes Hetzenegger, Anna Kirnich, Anette Rüth, Regine Scheib, Momme von Sydow sowie Volker Wiebe für die Datenerhebung.

(1994) vorgeschlagene Verfahren des „unpacking" der kritischen Grundmenge *A* (z.B. Linda ist Bankangestellte): Bietet man vor *A* die beiden konstituierenden Teilmengen $A \cap B$ (Linda ist Bankangestellte *und* in der Frauenbewegung aktiv) sowie $A \cap$ *nicht-B* (Linda ist Bankangestellte *und nicht* in der Frauenbewegung aktiv) dar, die zusammengenommen *A* ausschöpfen, so sollte nach Tversky (1994) die Teilmengenrelation offensichtlich werden und der Konjunktionsfehler verschwinden. Experiment 3 (N = 80 Studierende) schließlich kombinierte die Maßnahmen der Experimente 1 und 2: Es wurde auf den mathematischen Charakter der Aufgabe hingewiesen und zusätzlich das „Unpacking"-Format von Tversky (1994) verwendet. Eine zusätzliche Besonderheit des dritten Experiments bestand darin, daß die Grundmengen für alle Häufigkeitsbeurteilungsaufgaben auf 100 Exemplare festgesetzt und die Wahrscheinlichkeiten in Prozent geschätzt wurden, so daß das numerische Intervall der möglichen Antworten zwischen den beiden Frageformaten nicht mehr variierte.

Für jedes Experiment und jede der vier Versuchsbedingungen wurde die Anzahl der Vpn ausgezählt, die 0, 1, ..., 5 Konjunktionsfehler in den fünf Aufgaben gemacht hatten (vgl. Tabelle 6.4). Zusätzlich wurden die mittlere Konjunktionsfehlerzahl pro Aufgabe berechnet und mit BINOMIX (Erdfelder, 1993a) die G^2-Anpassungsstatistiken für das 1-Klassen- und das 2-Klassen-Binomialmischverteilungsmodell für jede Versuchsbedingung getrennt bestimmt. Varianzanalysen über die mittleren Fehlerzahlen wurden ebenfalls durchgeführt, sind jedoch in diesem Zusammenhang nicht von Interesse und bleiben deshalb im Tabelle 6.4 ausgespart.

Tabelle 6.4 zeigt, daß die Maßnahmen zur Reduktion des Konjunktionsfehlers zwar insofern wirksam waren, als die Konjunktionsfehlerraten verglichen mit anderen Untersuchungen tendenziell niedriger ausfallen, vor allem bei Wahrscheinlichkeitsrangbildungen. Dennoch kann von einer Elimination des Konjunktionsfehlers — gerade in den Experimenten mit „unpacking" der relevanten Grundmenge — nicht die Rede sein. Die entsprechende Vorhersage von Tverskys (1994) „support theory" muß also als widerlegt gelten. Das schon von Erdfelder et al. (1998) berichtete Ergebnismuster — ein im wesentlichen additiver Effekt des Frage- und des Antwortformats auf die mittlere Konjunktionsfehlerrate — spiegelt sich auch in diesen drei Experimenten wider, wenn auch auf insgesamt niedrigerem Niveau.

Für die Modellprüfung relevanter sind die Verteilungen der Fehlerhäufigkeiten in den vier Versuchsbedingungen. Schon rein deskriptiv sind andeutungsweise bimodale Fehlerverteilungen zu erkennen, was auf die Adäquatheit einer Zwei-Klassen-Binomialmischung hindeutet, sie aber nicht belegt (siehe dazu Erdfelder, 1990a). Die G^2-Anpassungsstatistiken sprechen dagegen eine recht klare Sprache: Das Modell der Zwei-Klassen-Mischung paßt immer, das Ein-Klassen-

Modell (identisch mit dem Modell einer einfachen Binomialverteilung) dagegen nur in zwei von zwölf Fällen. Zugrunde gelegt wurde bei allen Tests ein relativ hohes Typ-1-Fehlerrisiko von $\alpha = .10$, um das β-Fehlerrisiko nicht zu exzessiv ansteigen zu lassen.

Tabelle 6.4:
Ergebnisse von drei Experimenten zur Reduktion des Konjunktionsfehlers, in denen das Frageformat (Rangbildung vs. Schätzung) und das Antwortformat (Häufigkeit vs. Wahrscheinlichkeit) in 2×2-Designs manipuliert wurde. In der ersten Zeile jeder Zelle sind jeweils die Häufigkeiten von 0, 1, ..., 5 Konjunktionsfehlern in den 5 Aufgaben wiedergegeben, darunter die durchschnittlichen Fehlerraten pro Aufgabe und die G^2-Modellanpassungsstatistiken für das 1- und das 2-Klassen-Binomialmischverteilungsmodell.

VG	Stat.	Experiment 1	Experiment 2	Experiment 3
Rang-	Häuf.	14 4 5 4 2 1	3 4 4 4 3 2	4 3 2 5 3 3
bildung:	∅	26%	46%	49%
Häufig-	1-Kl.-G^2	$G^2_{(4)} = 22.03$, $p < .10$	$G^2_{(4)} = 8.02$, $p < .10$	$G^2_{(4)} = 16.94$, $p < .10$
keit	2-Kl.-G^2	$G^2_{(2)} = 0.58$, n.s.	$G^2_{(2)} = 0.27$, n.s.	$G^2_{(2)} = 1.20$, n.s.
Rang-	Häuf.	5 5 5 4 6 5	1 6 3 2 6 2	1 1 2 4 7 5
bildung:	∅	51%	52%	70%
Wahr.-	1-Kl.-G^2	$G^2_{(4)} = 23.96$, $p < .10$	$G^2_{(4)} = 12.13$, $p < .10$	$G^2_{(4)} = 6.16$, n.s.
keit	2-Kl.-G^2	$G^2_{(2)} = 0.50$, n.s.	$G^2_{(2)} = 3.26$, n.s.	$G^2_{(2)} = 0.12$, n.s.
Schät-	Häuf.	13 8 5 1 2 1	13 3 1 1 2 0	12 3 1 3 0 1
zung:	∅	23%	16%	19%
Häufig-	1-Kl.-G^2	$G^2_{(4)} = 15.50$, $p < .10$	$G^2_{(4)} = 29.57$, $p < .10$	$G^2_{(4)} = 30.68$, $p < .10$
keit	2-Kl.-G^2	$G^2_{(2)} = 1.27$, n.s.	$G^2_{(2)} = 0.26$, n.s.	$G^2_{(2)} = 2.51$, n.s.
Schät-	Häuf.	10 8 5 3 3 1	4 2 6 4 2 2	0 6 7 5 1 1
zung:	∅	29%	44%	44%
Wahr.-	1-Kl.-G^2	$G^2_{(4)} = 13.81$, $p < .10$	$G^2_{(4)} = 10.43$, $p < .10$	$G^2_{(4)} = 3.81$, n.s.
keit	2-Kl.-G^2	$G^2_{(2)} = 0.21$, n.s.	$G^2_{(2)} = 2.73$, n.s.	$G^2_{(2)} = 0.53$, n.s.

Damit ist eine notwendige, jedoch allein noch nicht hinreichende Bedingung für die Bewährung des Drei-Klassen-Modells zum Konjunktionsfehler erfüllt. Zusätzlich zu prüfen ist, ob die Klassenwahrscheinlichkeiten und die Fehlerwahrscheinlichkeiten innerhalb der latenten Klassen dem oben geschilderten Restriktionsmuster genügen. Dazu ist zunächst pro Experiment die Anpassungsgüte des unrestringierten 2-Klassen-Binomialmodells über alle vier Versuchsbedingungen hinweg zu bestimmen. Wie in Abschnitt 4.3.4 gezeigt wurde, ergeben sich diese Omnibus-Anpassungsstatistiken einfach durch Summierung der G^2-Statistiken für die einzelnen Versuchsbedingungen (vgl. dazu Tabelle 6). Für Experiment 1 erhält man so $G^2_{(8)} = 2.56$, für Experiment 2 $G^2_{(8)} = 6.52$ und für Experiment 3 $G^2_{(8)} = 4.36$. Fügt man dem Modell nun die Restriktionen hinzu, daß (a) die la-

tenten Klassenwahrscheinlichkeiten nicht durch das Antwortformat und (b) die Fehlerwahrscheinlichkeiten innerhalb der beiden Klassen nicht durch das Frageformat beeinflußt werden, so resultieren für das so restringierte Modell $G^2_{(14)}$-Statistiken von 12.59, 10.00 und 15.06. Die zugehörigen $\Delta G^2_{(6)}$-Differenzenstatistiken von 10.03, 3.48 und 10.70 sind bei 6 Freiheitsgraden zwar nicht optimal, selbst bei Zugrundelegung eines Signifikanzniveaus von $\alpha = .10$ aber noch nullhypothesenkonform. Einschränkend muß allerdings angemerkt werden, daß der Modelltest wegen der geringen Gesamtstichprobenumfänge von $N = 120$ bzw. $N = 80$ auch bei $\alpha = .10$ nur eine mäßige Teststärke aufweist (approx. $1-\beta = .79$ bzw. .61), selbst wenn man von mittleren Effektstärken sensu Cohen (1988) ausgeht ($w = .3$). Die Modelltests sind also allenfalls für „große" Modellverletzungen im Sinne von Cohen (1988) hinreichend sensitiv.

Tabelle 6.5:
Ergebnisse der ML-Parameterschätzung zu den drei Experimenten aus Tabelle 6.4 auf der Grundlage des 2-Klassen-Binomialmischverteilungsmodells.

Parameter	Experiment 1	Experiment 2	Experiment 3
$p(X = 1)$ (Häuf.)	.76	.63	.50
$p(X = 1)$ (Wahr.)	.50	.37	.11
π_1-Rangbildung	.15	.25	.16
π_1-Schätzung	.11	.06	.04
π_2-Rangbildung	.73	.76	.75
π_2-Schätzung	.56	.62	.50

Die ML-Parameterschätzung für das restringierte Modell ist aus zwei Gründen nicht ganz unproblematisch. Erstens ist das Modell — wie alle finiten Mischverteilungsmodelle ohne Ordnungsrestriktionen — nur lokal, nicht aber global identifizierbar, weil man durch Vertauschung der Klassenindizes identische Vorhersagen für die Kategoriewahrscheinlichkeiten erhält. In der Ergebnistabelle 6.5 wird grundsätzlich davon ausgegangen, daß die latente Klasse 1 die geringeren, Klasse 2 dagegen die höheren Fehlerraten produziert. In Abhängigkeit von der Startwertkombination kann der EM-Algorithmus für die gleichen Daten aber auch genau das umgekehrte Ergebnismuster liefern, was nicht irritieren sollte. Die Modellanpassungsstatistik G^2 ist von diesem Problem natürlich nicht betroffen.

Zweitens weist das Modell bei den vorliegenden Daten offensichtlich mehrere lokale Likelihood-Maxima auf, denen auch unterschiedliche Modellanpassungsstatistiken entsprechen. Dies ist insofern bemerkenswert, als das Problem lokaler Maxima bei MVB-Modellen ansonsten recht selten auftritt (vgl. Abschnitt 4.2.1). Binomialmischverteilungsmodelle haben vermutlich deshalb häufiger lokale Likelihood-Maxima, weil ihren Log-Likelihood-Funktionen Polynome höhe-

ren Grades zugrunde liegen als das normalerweise bei MVB-Modellen der Fall ist. Da 0 bis 5 Konjunktionsfehler vorkommen können, ist die Log-Likelihood-Funktion im gegebenen Fall ein logarithmiertes Polynom fünften Grades. Bei Polynomen fünften Grades sind mehrere Maxima im Parameterraum $[0, 1]^S$ durchaus leicht vorstellbar. Praktisch läßt sich das Problem nur dadurch in den Griff bekommen, daß man den EM-Algorithmus mehrfach von unterschiedlichen Zufallsstartwerten aus startet. Die Wahrscheinlichkeit, von einer diese Startwertekombinationen aus das globale Maximum zu erwischen, ist sehr hoch (vgl. Abschnitt 4.2.1). Die Ergebnisse, die in der Tabelle 6.5 berichtet werden, entsprechen der kleinsten G^2-Statistik (und damit dem höchsten Maximum der Likelihood-Funktion) bei 10 Replikationen des EM-Schätzverfahrens mittels AppleTree (Rothkegel, 1998).

Wie Tabelle 6.5 zeigt, passen die Parameterschätzungen insgesamt recht gut zu den Prognosen des Drei-Gruppen-Modells. Zunächst einmal ist der Anteil der latenten Klasse 1 — der Klasse mit geringen Fehlerwahrscheinlichkeiten — bei Aufgaben im Häufigkeitsformat grundsätzlich höher als bei Aufgaben im Wahrscheinlichkeitsformat. Setzt man die Größe der latenten Klasse 1 für Häufigkeits- und Wahrscheinlichkeitsaufgaben gleich, so resultieren durchweg signifikante χ^2-Differenzen-Statistiken von $\Delta G^2_{(1)} = 6.03$ (Experiment 1), $\Delta G^2_{(1)} = 5.32$ (Experiment 2) und $\Delta G^2_{(1)} = 12.53$ (Experiment 3). Nach dem Drei-Klassen-Modell des Konjunktionsfehlers muß dies so sein, denn die latente Klasse 1 besteht bei Wahrscheinlichkeitsaufgaben nur aus der Gruppe W, bei Häufigkeitsaufgaben dagegen aus den Gruppen W und H. Die Differenz der Größe von Klasse 1 bei Häufigkeits- und bei Wahrscheinlichkeitsaufgaben entspricht also gerade der Größe von Gruppe H. Interessanterweise scheint diese Differenz in allen drei Experimenten relativ ähnlich auszufallen (.26, .26 und .39).

Erwartungsgemäß fallen die Konjunktionsfehlerraten in Klasse 1 für die Rangbildungs- und die Schätzaufgaben (Zeile 4 und 5 der Tabelle) nicht nur niedrig, sondern auch recht ähnlich aus. Dies gilt zumindest für Experiment 1 ($\Delta G^2_{(1)} = 0.25$, n.s.) und — trotz signifikanter Differenzenstatistik — auch für Experiment 3 ($\Delta G^2_{(1)} = 3.32$, $p < .10$). In Experiment 2 fällt die Differenz dagegen wesentlich größer aus ($\Delta G^2_{(1)} = 8.01$, $p<.10$), was insofern leicht zu erklären ist, als die Vpn in diesem Experiment nicht explizit auf den mathematischen Charakter der Aufgaben hingewiesen wurden. In Verbindung mit einer Rangbildungsinstruktion ist es für die Vpn deshalb naheliegend, die Aufgaben eher im Sinne einer „Plausibilitätseinschätzung" der verschiedenen Alternativen zu verstehen, so daß der Kolmogoroffsche Mengenmaßbegriff keine Anwendung mehr finden muß. Dies kann zu mehr „Konjunktionsfehlern" — die in diesem Kontext ja gar keine Fehler mehr sind — führen. Eine Absolutschätzungsinstruktion macht dagegen schon per se klar, daß es sich um eine mathematische Aufgabe handelt, so daß sich

der fehlende Hinweis auf den mathematischen Charakter der Aufgaben nicht schädlich auswirkt.

In der latenten Klasse 2 gibt es zwar eine Tendenz zu mehr Konjunktionsfehlern bei Rangbildungsinstruktionen, die jedoch nur in Experiment 2 ($\Delta G^2_{(1)}$ = 3.59) und Experiment 3 signifikant ausfällt ($\Delta G^2_{(1)}$ = 13.02), nicht dagegen in Experiment 1 ($\Delta G^2_{(1)}$ = 2.49). Dieses Ergebnis ist insgesamt vereinbar mit der Annahme, daß die Personen, die ihren Urteilen nicht den Kolmogoroffschen Mengenmaßbegriff zugrunde legen, generell hohe Konjunktionsfehlerraten haben, die in Abhängigkeit von Kontextcharakteristika der Aufgaben mehr oder minder stark variieren können.

Alle geschilderten Ergebnisse und die daraus abgeleiteten Interpretationen stehen unter dem Vorbehalt, daß die durchgeführten Tests wegen der recht kleinen Stichprobenumfänge nicht sehr teststark sind. Außerdem steht noch eine systematische Validierung dieses Modells aus. So sollte es etwa möglich sein, die Größe der Gruppe W durch gezielte Selektion von Vpn — beispielsweise die Untersuchung von Mathematikern — oder durch Trainingsmaßnahmen positiv zu beeinflussen. Die gleichen Maßnahmen sollten unter konstanten Randbedingungen *keinen* Effekt auf die Konjunktionsfehlerraten innerhalb der latenten Klassen haben — eine sicherlich nicht triviale Vorhersage. Es bleibt abzuwarten, ob das Modell derart strengen Prüfversuchen standhält.

6.5 Markoff-Ketten-Modelle und latente Markoff-Ketten-Modelle

Markoff-Ketten-Modelle (MK-Modelle) bildeten den formalen Kern zahlreicher mathematischer Lerntheorien, die zwischen 1950 und 1975 recht einflußreich waren (vgl. Atkinson et al., 1965; Tack, 1976; Wickens, 1982). MK-Modelle basieren auf finiten Markoff-Ketten-Prozessen. Ein solcher Prozeß kann wie folgt definiert werden (vgl. z.B. Ross, 1997):

> Zu jedem diskreten Zeitpunkt $t = 0, ..., T$ sei genau einer von J diskreten Zuständen („states") aus dem Zustandsraum $\mathbf{Z} = \{1, 2, ..., J\}$ gegeben. Die (beobachtbare) kategoriale ZV $X(t)$ gebe den Zustand zum Zeitpunkt t an. Alle ZVn $X(t)$ seien auf dem gleichen Wahrscheinlichkeitsraum definiert, so daß $\{X(t), t = 0, ..., T\}$ ein stochastischer Prozeß in der diskreten Zeit ist. Gilt für alle Zeitpunkte $t \geq 0$ und für beliebige Zustände $j(t) \in \mathbf{Z}$

$$\begin{aligned} &Pr\Big(X(t+1) = j(t+1) \,\Big|\, X(t) = j(t), X(t-1) = j(t-1), \ \ldots, X(0) = j(0)\Big) \\ &\quad = Pr\Big(X(t+1) = j(t+1) \,\Big|\, X(t) = j(t)\Big) \ \text{— kurz: } \pi_{j(t),j(t+1)} \ \text{—,} \end{aligned} \tag{6.5-1}$$

so heißt $\{X(t),\ t = 0, ..., T\}$ „finiter Markoff-Ketten-Prozeß"[75] oder kurz „Markoff-Kette".

Ein Markoff-Ketten-Modell (MK-Modell) ist ein parametrisiertes multinomiales Modell zur $J^{(T+1)}$-Kontingenztafel der möglichen Zustandskombinationen über die T+1 Zeitpunkte hinweg, demzufolge die Zellwahrscheinlichkeiten einem Markoff-Ketten-Prozeß folgen. Im Grunde genommen wird also lediglich behauptet, daß die Wahrscheinlichkeit eines Zustandes zum Zeitpunkt t+1 (z.B. des Lernzustandes eines Items vor dem (t+1)-ten Gedächtnistest) vom vorherigen Zustand zum Zeitpunkt t (dem Lernzustand vor dem t-ten Gedächtnistest) abhängen kann, nicht aber von den Zuständen zu früheren Zeitpunkten t–1, t–2, etc. Die Vergangenheit kann also nur über den jeweils zuletzt erreichten Zustand Einfluß auf den nächsten Zustand nehmen, nicht über die Vorgeschichte dieses Zustands. Dies ist die sogenannte „memoryless property" von Markoff-Modellen, oder kurz die Markoff-Eigenschaft (Wickens, 1982, S. 15).

Wie man leicht sehen kann, sind MK-Modelle ebenfalls spezielle MVB-Modelle bzw. — bei binärem Zustandsraum — binäre MVB-Modelle. Aus der Definition einer Markoff-Kette folgt nämlich unmittelbar, daß für alle Zellwahrscheinlichkeiten der betrachteten Kontingenztafel

$$Pr\big(X(T) = j(T), X(T-1) = j(T-1),\ \ldots, X(1) = j(1), X(0) = j(0)\big) = \pi_{j(0)} \cdot \pi_{j(0),j(1)} \cdot \pi_{j(1),j(2)} \cdot \ \cdots \cdot \pi_{j(T-1),j(T)} \qquad (6.5\text{-}2)$$

gilt. In Gleichung (6.5-2) ist $\pi_{j(0)}$ die „initial-state probability" des Zustandes $j(0) \in \mathbf{Z}$ zum Zeitpunkt $t = 0$; die anderen Parameter sind Übergangswahrscheinlichkeiten $\pi_{j(t),j(t+1)}$ für $t = 0, 1, ..., T-1$ gemäß Gleichung (6.5-1). Da einige oder sogar alle Übergangswahrscheinlichkeiten gleich sein können, ist (6.5-2) maximal ein Polynom (T+1)-ten Grades in den Modellparametern. Es handelt sich offenkundig um einen Spezialfall der allgemeinen MVB-Modellgleichung (3.4-7), wobei $G = 1 + T \cdot J$ Parametergruppen vorliegen, nämlich die „initial-state probabilities" sowie zusätzlich die Übergangswahrscheinlichkeiten für T Zeitpunkte und jeweils J Ausgangszustände des vorherigen Zeitpunkts. Handelt es sich um ein sogenanntes *stationäres* MK-Modell, bei dem die Übergangswahrscheinlichkeiten $\pi_{j(t),j(t+1)}$ vom Zustand $j(t)$ in den Zustand $j(t+1)$ für $t = 0, ..., T-1$ immer gleich sind — also nur vom jeweiligen Ausgangs- und Zielzustand, nicht aber vom Zeitpunkt abhängen — so reduziert sich die Anzahl der Parametergruppen auf $G = 1 + J$. Die Anzahl freier Parameter ist in jeder der Parametergruppen maximal $L_g = J-1$. Falls also ein

[75] Finite Markoff-Ketten-Prozesse bilden eine Teilklasse von Markoff-Prozessen. Letztere setzen nicht notwendigerweise einen Zustandsraum **Z** mit endlich vielen Zuständen voraus; es kann sich auch um unendlich viele Zustände handeln. Ferner kann die Zeit t auch kontinuierlich statt diskret sein (vgl. Wickens, 1982, S. 15).

binärer Zustandsraum mit $J = 2$ Zuständen vorliegt oder — bei nichtbinärem Zustandsraum — die Übergangswahrscheinlichkeiten in Termini binärer Parameter reparametrisiert werden, wird aus dem MK-Modell ein binäres MVB-Modell.

Die Ereignisbaumrepräsentation von MK-Modellen bereitet insofern keinerlei Probleme, als die Konstruktionsvorschrift sehr einfach ist: Man sieht einfach — beginnend bei $t = 0$ — für jeden Zeitpunkt eine zusätzliche Ebene im Baum vor; jede Astgabelung entspricht den J Folgezuständen, die auf einen gegebenen Ausgangszustand folgen können. Praktisch stößt man hierbei jedoch schon bei mäßig großem T oder J auf Probleme, weil die Baumdiagramme schnell unübersichtlich werden. Dies bedeutet auch, daß eine Analyse von MK-Modellen mit Software, die speziell auf MVB-Modelle zugeschnitten ist, zwar prinzipiell immer möglich ist; die Spezifikation des zu analysierenden Modells über Baumdiagramme bzw. die korrespondierenden Astwahrscheinlichkeiten kann jedoch durchaus zeitraubend sein. Trotzdem hat die MVB-Analyse von MK-Modellen Vorteile. Diese werden besonders dann deutlich, wenn sich ein Standard-MK-Modell, das durch die Stationaritätsannahme gekennzeichnet ist, als empirisch nicht haltbar erweist. Betrachtet man MK-Modelle als MVB-Modelle, so werden schnell Verallgemeinerungsmöglichkeiten erkennbar, die dabei helfen können, die Quelle der schlechten Datenanpassung zu beseitigen.

Eine naheliegende Verallgemeinerungsmöglichkeit von MK-Modellen bilden sogenannte latente MK-Modelle (Langeheine & van de Pol, 1990; Wiggins, 1973). In einem latenten MK-Modell ist die ZV $X(t)$, für die ein MK-Modell gilt, latent, d.h. die J Zustände des MK-Prozesses sind nicht direkt beobachtbar. Beobachtbar ist dagegen eine ZV $Y(t)$ mit I möglichen Ausprägungen (z.B. Verhaltensweisen oder Vp-Antworten aus einem Antwortraum $\mathbf{R} = \{1, 2, ..., I\}$), die über ein Meßmodell mit der latenten Zustandsvariablen $X(t)$ verknüpft ist. Bezeichnet $p_{i(t)\,|\,j(t)}$ die bedingte Wahrscheinlichkeit, daß $Y(t)$ die Ausprägung $i(t) \in \{1, ..., I\}$ annimmt, gegeben daß die latente Zustandsvariable $X(t)$ die Ausprägung $j(t) \in \{1, ..., J\}$ annimmt, dann ist die Wahrscheinlichkeit eines bestimmten Antwortmusters über die $(T+1)$ Zeitpunkte nach dem latenten MK-Modell gegeben durch:

$$Pr\big(Y(T) = i(T), Y(T-1) = i(T-1),\ \ldots, Y(1) = i(1), Y(0) = i(0)\big) =$$
$$\sum_{j(0)}^{J}\sum_{j(1)}^{J}\cdots\sum_{j(T)}^{J} \pi_{j(0)} \cdot p_{i(0)|j(0)} \cdot \pi_{j(0),j(1)} \cdot p_{i(1)|j(1)} \cdot \ \cdots\ \cdot \pi_{j(T-1),j(T)} \cdot p_{i(T)|j(T)} \qquad (6.5\text{-}3)$$

Wie man an der Modellgleichung (6.5-3) unschwer erkennt, gehören latente MK-Modelle ebenso wie gewöhnliche MK-Modelle zur Familie der MVB-Modelle und können folglich ebenfalls mit MVB-Software analysiert werden. Das Meßmodell eines latenten MK-Modells — die Wahrscheinlichkeiten $p_{i(t)\,|\,j(t)}$ — kann alternativ auch in das Strukturmodell — die Matrix der Übergangswahrscheinlich-

keiten — integriert werden, indem man einen neuen Zustandsraum $\mathbf{Z}^* = \mathbf{Z} \times \mathbf{R}$ kreiert, der alle möglichen Kombinationen von Elementen des latenten Zustandsraumes **Z** und des Antwortraumes **R** umfaßt. Man splittet also gewissermaßen jeden latenten Zustand aus **Z** in alle möglichen empirischen Manifestation des betreffenden Zustandes auf. Die „initial state probabilities" und die Übergangswahrscheinlichkeiten zu $\mathbf{Z}^*$ ergeben sich einfach als Produkte der *p*- und der π-Parameter (Wickens, 1982, S. 83-88). Jedes Element von $\mathbf{Z}^*$ kann dann deterministisch einer bestimmten Antwort aus **R** zugeordnet werden. Dies macht das latente MK-Modell konzeptuell einfacher, ohne daß ein Verlust ein Allgemeinheit damit verbunden wäre. Man kann wegen der Möglichkeit einer entsprechenden Reformulierung des Zustandsraumes auch sagen, daß latente MK-Modelle immer gewöhnlichen MK-Modellen äquivalent sind, bei denen verschiedene Zustände in das gleiche beobachtbare Ereignis einmünden können.

Latente MK-Modelle spielen wegen des Dekomponierungsproblems (vgl. Abschnitt 2.6) in der kognitiven Psychologie eine wichtigere Rolle als MK-Modelle, bei denen jeder Zustand eineindeutig einem bestimmten beobachtbaren Ereignis zugeordnet ist. Beispielsweise verbergen sich hinter der erfolgreichen Reproduktion eines Items und dem Scheitern der Reproduktion im allgemeinen mehr als nur zwei verschiedene kognitive Zustände. Ein Beispiel für die kognitionspsychologische Modellierung von Reproduktionsdaten über ein latentes MK-Modell mit mehreren latenten Zuständen behandeln die folgenden beiden Abschnitte.

6.5.1 Ein Dual-Code-Markoff-Modell verbalen Lernens

Mehr-Speicher-Konzeptionen des Gedächtnisses eignen sich hervorragend zur Formalisierung in Termini von latenten MK-Modellen, da die (erfolgreiche oder nicht erfolgreiche) Speicherung in einem speziellen Gedächtnissystem und der (erfolgreiche oder nicht erfolgreiche) Abruf aus einem bestimmten Gedächtnissystem zwanglos als Elemente des Zustandsraumes eines latenten MK-Modells aufgefaßt werden können. Für die Anpassung und Prüfung eines solchen Modells sind typischerweise wiederholte Beobachtungen der Probanden bei mehreren Gedächtnistests zu denselben Items erforderlich. Beispielsweise hat Erdfelder (1993b) ein latentes MK-Modell zur wiederholten freien oder geförderten Reproduktion von abstrakten und konkreten Items vorgestellt, welches die Kernannahmen der Dual-Code-Theorie des verbalen Lernens von Paivio (1971; 1986) formalisiert. Dieses Modell soll hier kurz vorgestellt werden. Der nächste Abschnitt behandelt die Prüfung verschiedener Modellvarianten anhand neuropsychologischer Daten.

Einer der am besten gesicherten Befunde der episodischen Langzeitgedächtnisforschung ist der Bildhaftigkeitseffekt: Bilder werden besser erinnert als Wörter, und konkrete Wörter werden besser erinnert als abstrakte Wörter. Dieser Effekt

und zahlreiche verwandte Befunde lassen sich gut im Rahmen von Paivios dualer Kodierungstheorie interpretieren (zusammenfassend: Wippich & Bredenkamp, 1979). Paivio (1971; 1986) geht davon aus, daß es zwei Speichersysteme im episodischen Langzeitgedächtnis gibt: Ein verbales System, das auf die Analyse und Speicherung von verbal-semantischen Informationen in Form von „Logogenen" beschränkt ist, und ein imaginales System, das Informationen in analog-bildhafter Form über „Imagene" abspeichert. Abstrakte Wörter können lediglich im verbalen System repräsentiert werden, konkrete Wörter und Bilder dagegen sowohl im verbalen als auch im imaginalen System. Konkret-bildhafte Information wird also im Gegensatz zu abstrakter Information doppelt kodiert und hat schon allein deshalb einen Vorteil in Gedächtnistests: Wenn der Abruf aus dem verbalen System scheitert, bleibt immer noch die Option, die Information aus dem imaginalen System abzurufen („code redundancy principle"). Hinzu kommt, daß Paivio (1971) generell höhere Abrufwahrscheinlichkeit für Imagene im Vergleich zu Logogenen annimmt. Dies verschafft konkret-bildhafter Information einen weiteren Gedächtnisvorteil gegenüber abstrakter Information.

Über die neurobiologischen Grundlagen der beide Gedächtnissysteme hat Paivio ebenfalls Spekulationen angestellt. Paivio (1986) vermutete, daß das verbale System in der linken Hirnhemisphäre (speziell im Temporalbereich) zu lokalisieren ist, das imaginale System dagegen eher in der rechten Hemisphäre. Die bislang vorliegenden Befunde sprechen allerdings eher gegen diese Lateralisierungshypothese (Ehrlichman & Barrett, 1983; Richardson, 1990). Es scheint keinen eindeutigen Zusammenhang zwischen Hirnhemisphären und der Speicherung verbaler versus bildlicher Informationen zu geben.

Die Grundannahmen der Dual-Code-Theorie lassen sich recht zwanglos im Rahmen eines latenten MK-Modells formalisieren. Erdfelder (1993b) hat für wiederholte Lern-Test-Abfolgen konkreter und abstrakter Wörter folgende Annahmen zugrunde gelegt:

(1) Vor jedem Reproduktionstest ist ein Item in einem der folgenden episodischen Gedächtniszustände:
   – weder im imaginalen noch im verbalen Gedächtnissystem gespeichert;
   – nur im verbalen Gedächtnissystem gespeichert;
   – nur im imaginalen Gedächtnissystem gespeichert;
   – in beiden Gedächtnissystemen gespeichert.

(2) Vor der ersten Lernphase sind alle Items in keinem System gespeichert.

(3) Wenn ein Item *nicht* im imaginalen Gedächtnissystem gespeichert ist, ist es nach dem nächsten Lerndurchgang mit Wahrscheinlichkeit α_c imaginal gespeichert, sofern es sich um ein konkretes Wort handelt, und mit Wahrscheinlichkeit α_a imaginal gespeichert, sofern es sich um ein abstraktes Wort handelt. Es sollte natürlich gelten: $\alpha_c > \alpha_a$.

(4) Wenn ein Item *nicht* im verbalen Gedächtnissystem gespeichert ist, ist es im nächsten Durchgang mit Wahrscheinlichkeit β verbal gespeichert, wobei β nicht davon abhängt, ob das Item abstrakt oder konkret ist.

(5) Die Speicherung im imaginalen und im verbalen System erfolgt stochastisch unabhängig voneinander, d.h. die Wahrscheinlichkeit, daß ein weder verbal noch imaginal gespeichertes Item nach dem nächsten Lerndurchgang sowohl verbal als auch imaginal gespeichert ist, beträgt $\alpha_c \cdot \beta$ bei konkreten Wörtern und $\alpha_a \cdot \beta$ bei abstrakten Wörtern.

(6) Ist ein Item irgendwann einmal in einem der beiden episodischen Gedächtnissysteme gespeichert, so bleibt es bis zum Ende des Lernexperiments in diesem Gedächtnissystem, wird also nicht wieder „vergessen". Dies bedeutet jedoch nicht, daß das Item immer auch abrufbar sein muß (vgl. Annahme 7).

(7) Ist ein Item verbal gespeichert, so kann es in dem betreffenden Gedächtnistests mit Wahrscheinlichkeit r_v abgerufen werden; ist es imaginal gespeichert, so kann es mit Wahrscheinlichkeit r_i abgerufen werden. Es wird erwartet, daß $r_v < r_i$. Der Abruf aus dem verbalen und dem imaginalen System erfolgt stochastisch unabhängig voneinander, d.h. die Wahrscheinlichkeit, daß ein verbal und imaginal gespeichertes Item auch aus beiden Systemen abgerufen werden kann, ist $r_v \cdot r_i$. Selbstverständlich kann ein Item aus einem System nie abgerufen werden, wenn es in diesem System nicht gespeichert ist.

(8) Ein Item wird in einem Durchgang dann und nur dann erfolgreich reproduziert, wenn es aus mindestens einem der beiden Gedächtnissysteme abgerufen werden kann.

Aus diesen Annahmen läßt sich problemlos die in Tabelle 6.5.1 wiedergegebene Matrix von Übergangswahrscheinlichkeiten für konkrete Wörter ableiten. Die analoge Matrix für abstrakte Wörter erhält man, indem man den Parameter α_c durch α_a ersetzt.

Anders als in Tabelle 1 bei Erdfelder (1993b) werden in Tabelle 6.5.1 nicht nur Übergänge zwischen den möglichen Speicherungszuständen, sondern Übergänge zwischen *Kombinationen* von Speicherungs- und Abrufzuständen angegeben. Es wird also der erweiterte Zustandsraum $\mathbf{Z}^* = \mathbf{Z} \times \mathbf{R}$ betrachtet, was den Vorteil hat, daß die Übergangsmatrix alle Modellparameter umfaßt. Die Prüfung des in Tabelle 6.5.1 dargestellten Modells anhand eines neuropsychologischen Datensatzes wird im folgenden Abschnitt behandelt.

Tabelle 6.6:
Matrix der Übergangswahrscheinlichkeiten zum Dual-Code-Markoff-Modell für die wiederholte freie oder geförderte Reproduktion konkreter Wörter (nach Erdfelder, 1993b). „I" und „V" stehen für Speicherung im imaginalen bzw. verbalen System, die mit Querbalken versehen Buchstaben für fehlende Speicherung in dem betreffenden System. Das Symbol „+" steht für erfolgreichen Abruf, das Symbol „–" für nicht erfolgreichen Abruf. Die entsprechende Matrix für abstrakte Wörter ergibt sich, wenn man den Parameter α_c durch α_a ersetzt.

Durchgang t		Bedingte Wahrscheinlichkeit im Durchgang t+1						
Speicher	Abruf	$\bar{I}\,\&\,\bar{V}$	$I\,\&\,\bar{V}$	$I\,\&\,\bar{V}$	$\bar{I}\,\&\,V$	$\bar{I}\,\&\,V$	$I\,\&\,V$	$I\,\&\,V$
		–	–	+	–	+	–	+
$\bar{I}\,\&\,\bar{V}$	–	$(1-\alpha_c)\cdot(1-\beta)$	$\alpha_c\cdot(1-\beta)\cdot(1-r_i)$	$\alpha_c\cdot(1-\beta)\cdot r_i$	$(1-\alpha_c)\cdot\beta\cdot(1-r_v)$	$(1-\alpha_c)\cdot\beta\cdot r_v$	$\alpha_c\cdot\beta\cdot(1-r_i)\cdot(1-r_v)$	$\alpha_c\cdot\beta\cdot(1-(1-r_i)\cdot(1-r_v))$
$I\,\&\,\bar{V}$	–	0	$(1-\beta)\cdot(1-r_i)$	$(1-\beta)\cdot r_i$	0	0	$\beta\cdot(1-r_i)\cdot(1-r_v)$	$\beta\cdot(1-(1-r_i)\cdot(1-r_v))$
$I\,\&\,\bar{V}$	+	0	$(1-\beta)\cdot(1-r_i)$	$(1-\beta)\cdot r_i$	0	0	$\beta\cdot(1-r_i)\cdot(1-r_v)$	$\beta\cdot(1-(1-r_i)\cdot(1-r_v))$
$\bar{I}\,\&\,V$	–	0	0	0	$(1-\alpha_c)\cdot(1-r_v)$	$(1-\alpha_c)\cdot r_v$	$\alpha_c\cdot(1-r_i)\cdot(1-r_v)$	$\alpha_c\cdot(1-(1-r_i)\cdot(1-r_v))$
$\bar{I}\,\&\,V$	+	0	0	0	$(1-\alpha_c)\cdot(1-r_v)$	$(1-\alpha_c)\cdot r_v$	$\alpha_c\cdot(1-r_i)\cdot(1-r_v)$	$\alpha_c\cdot(1-(1-r_i)\cdot(1-r_v))$
$I\,\&\,V$	–	0	0	0	0	0	$(1-r_i)\cdot(1-r_v)$	$(1-(1-r_i)\cdot(1-r_v))$
$I\,\&\,V$	+	0	0	0	0	0	$(1-r_i)\cdot(1-r_v)$	$(1-(1-r_i)\cdot(1-r_v))$

6.5.2 Prüfung des Dual-Code-Markoff-Modells

Klein (1989) hat unter anderem Paar-Assoziations-Lernen (PAL) bei 9 Epilepsiepatienten mit linkstemporalem Fokus, 11 Epilepsiepatienten mit rechtstemporalem Fokus und 16 gesunden Kontrollpersonen untersucht. Insgesamt 21 abstrakte Wortpaare und 21 konkrete Wortpaare — in zufälliger Reihenfolge angeordnet — mußten gelernt werden; anschließend wurde für jedes Wortpaar der Liste das erste Wort dargeboten und die Vp mußte das zweite Wort dazu reproduzieren (Lern-

Prüf-Methode des PAL). Die Lern- und die Prüfphase wurde insgesamt fünf Mal wiederholt. Dieses Experiment erfüllt die strukturellen Voraussetzungen, die an Daten zu richten sind, die zur Prüfung des Dual-Code-MK-Modells aus dem letzten Abschnitt herangezogen werden.

Tabelle 6.7:
Daten der Untersuchung zum Paar-Assoziations-Lernen von Klein (1989). In den Zellen stehen die Häufigkeiten der verschiedenen Erfolg-Mißerfolg-Sequenzen bei abstrakten und konkreten Wortpaaren für drei verschiedene Probandengruppen. „+" steht für erfolgreiche Reproduktion in dem betreffenden Durchgang, „-" für nicht erfolgreiche Reproduktion.

Durchgang			Kontrollgruppe		Linksfok. Epil.		Rechtsfok. Epil.	
1	2	3	abstr.	konkr.	abstr.	konkr.	abstr.	konkr.
+	+	+	15	112	1	23	1	26
+	+	–	0	2	0	1	0	2
+	–	+	3	7	1	1	2	2
+	–	–	2	2	1	0	0	0
–	+	+	35	96	2	38	2	61
–	+	–	8	4	1	6	1	7
–	–	+	47	71	5	37	8	52
–	–	–	226	42	178	83	217	81

Erdfelder (1993b) hat die ersten vier Lern-Prüf-Durchgänge des Experiments mit einem speziell auf das Dual-Code-MK-Modell zugeschnittenen Programm analysiert und in wenigstens zwei von drei Fällen unakzeptabel hohe Anpassungsindizes erhalten. Die Pearson-Anpassungsstatistiken lauteten $X^2_{(25)} = 54.50$ (Kontrollpersonen), $X^2_{(25)} = 32.11$ (linkstemporale Epileptiker) und $X^2_{(25)} = 65.47$ (rechtstemporale Epileptiker). Diese Analyse wurde nunmehr für die ersten drei Lern-Prüfdurchgänge mittels AppleTree (Rothkegel, 1998; 1999) wiederholt, um das Problem der vielen Nullzellen, die in der kompletten 2^4-Kontingenztafeln zu den ersten vier Lern-Prüf-Durchgängen enthalten waren, zu umgehen. Wie Tabelle 6.7 zeigt, verschwindet das Nullzellenproblem weitgehend, wenn man nur die Erfolg-Mißerfolg-Sequenzen der ersten vier Durchgänge betrachtet.

Leider ändert der Übergang von vier zu drei Lern-Prüf-Durchgängen aber nichts an den mäßigen Anpassungsindizes des Modells. Zwar sind die Likelihood-Quotienten-Anpassungsstatistiken nun kleiner, was jedoch wegen der Reduktion der Freiheitsgrade auf $df = 2\cdot(8–1)\text{-}5 = 9$ keine wirkliche Verbesserung der Anpassungsgüte bedeutet. Für die Kontrollpersonen ist $G^2_{(9)} = 31.41$, für die linkstemporalen Epileptiker immerhin $G^2_{(9)} = 12.16$, für die rechtstemporalen Epileptiker jedoch wieder $G^2_{(9)} = 29.46$. Zwei der drei Anpassungsstatistiken sind auf dem Ni-

veau $\alpha = .05$ — ein vernünftiges Signifikanzniveau bei den gegebenen Stichprobenumfängen — signifikant.

Worauf ist die mäßige Anpassungsgüte des Modells zurückzuführen? Erdfelder (1993b) hat vermutet, daß die Annahme gleicher β-Parameter für abstrakte und konkrete Wortpaare für Kleins (1989) Daten möglicherweise verletzt ist, da die verwendeten konkreten Wörter nach den Wortnormentabellen von Baschek, Bredenkamp, Oehrle und Wippich (1977) nicht nur bildhafter, sondern auch bedeutungshaltiger waren als die abstrakten Wörter. Dies kann bedeuten, daß konkrete Wörter nicht nur leichter imaginal, sondern zusätzlich auch leichter verbal enkodierbar waren als die abstrakten Wörter. Sollte dies tatsächlich der Grund für die schlechte Modellanpassung sein, wäre dies für das Modell sehr problematisch, denn nach Einführung zweier verschiedener β-Parameter für abstrakte und konkrete Wörter wäre das Modell nur noch lokal, jedoch nicht mehr global identifizierbar. Es wäre dann keine spezifische Aussage über Speicher- und Abrufwahrscheinlichkeiten für das imaginale versus verbale Gedächtnissystem mehr möglich.

Ein weiterer möglicher Grund für die hohen G^2-Statistiken könnte die Stationaritätsannahme für die Übergangswahrscheinlichkeiten zwischen den Gedächtniszuständen sein. Tatsächlich ist diese Annahme speziell für den ersten Lerndurchgang nicht sehr plausibel: Warum sollte die Speicherung eines Items ohne jede vorherige Lernerfahrung genauso wahrscheinlich sein wie dessen erfolgreiche Speicherung nach wiederholten Lernphasen? Um eine Antwort auf diese Frage zu finden, wurde das Dual-Code-MK-Modell auf ein latentes MK-Modell verallgemeinert, das zwar für die Lern-Prüf-Durchgänge 2 und 3 stationäre Übergangswahrscheinlichkeiten annimmt, für den ersten Lerndurchgang jedoch beliebige (mutmaßlich kleinere) Speicherungswahrscheinlichkeiten α_c, α_a und β zuläßt. Auch dieses 8-Parameter-Modell ist für zwei 2^3-Kontingenztafeln zu abstrakten und konkreten Items noch global identifizierbar. Das diesem latenten MK-Modell entsprechende MVB-Modell ist trotz des nicht binären Zustandsraumes ein binäres Modell, da alle Parameter der Übergangsmatrix binär sind. Die entsprechende Modelleingabedatei (EQN-Datei) für die AppleTree- und GPT-Computerprogramme ist allerdings sehr umfangreich und findet sich in Abschnitt 7 des Anhangs. Eine Analyse mit dem auf 640 KB RAM beschränkten MBT-Programm ist wegen der Größe des Modells nicht möglich.

Wie die Ergebnistabelle 6.8 zeigt, erweist sich die Liberalisierung des Dual-Code-MK-Modells zu einem nichtstationären MK-Modell als sehr wirksam. Die G^2-Modellanpassungsindizes in der letzten Zeile der Tabelle, die wegen der drei zusätzlichen Parameter nunmehr auf $df = 6$ basieren, belegen, daß das Modell für wenigstens zwei der drei Probandengruppen sehr gut paßt. Legt man ein α-Niveau von .10 zugrunde, was für mittlere Abweichungen im Cohenschen (1988) Sinne (*w*

= .3) bei den gegebenen Stichprobenumfängen und Freiheitsgraden eine Teststärke von mindestens (1–β) = .85 garantiert, so muß zwar für die Kontrollgruppe das Modell noch zurückgewiesen werden; dennoch liegt auch hierbei eine deutliche Verbesserung gegenüber dem stationären Modell vor.

Tabelle 6.8:
ML-Parameterschätzungen, 90%-Konfidenzintervalle (in Klammern) und G^2-Statistiken bei Anwendung des liberalisierten nichtstationären Dual-Code-MK-Modells auf die Daten von Klein (1989), die hier in Tabelle 6.7 wiedergegeben sind.

Parameter	Durchgänge	Kontrollgr.	LT-Epilep.	RT-Epilep.	$\Delta G^2_{(2)}$
a_c (Sp. im.)	1	.36 (.31-.41)	.14 (.09-.20)	.13 (.08-.19)	30.06, p<.10
a_c (Sp. im.)	2 und 3	.50 (.45-.56)	.28 (.23-.34)	.36 (.30-.42)	16.99, p<.10
a_a(Sp. im.)	1	.05 (.02-.07)	.01 (.00-.03)	.00 (.00-.03)	5.54, p<.10
a_a(Sp. im.)	2 und 3	.11 (.08-.15)	.01 (.00-.04)	.00 (.00-.05)	6.56, p<.10
β (Sp. verb.)	1	.13 (.00-.27)	.06 (.01-.09)	.05 (.01-.09)	2.06, n.s.
β (Sp. verb.)	2 und 3	.12 (.00-.25)	.02 (.00-.08)	.03 (.00-.09)	1.39, n.s.
r_i (Abr. im.)	alle	.96 (.93-.98)	.90 (.84-.97)	.90 (.84-.95)	4.09, n.s.
r_v (Abr. v.)	alle	.23 (.00-.52)	.39 (.00-.80)	.57 (.18-.96)	1.44, n.s.
$G^2_{(6)}$		11.68, p<.10	2.33, n.s.	2.33, n.s.	

Auch die ML-Parameterschätzungen ergeben ein Ergebnismuster, das auf dem Hintergrund der Dual-Code-Theorie und der bislang vorliegenden Befunde zur Lateralisierungshypothese durchaus zu erwarten ist: Die Schätzwerte der Speicherungswahrscheinlichkeiten im imaginalen System liegen für konkrete Wörter durchweg höher als für abstrakte Wörter, die Schätzer für den ersten Durchgang unter (oder kaum über) denen für die folgenden Durchgänge. Erwartungsgemäß fällt auch die geschätzte Abrufwahrscheinlichkeit aus dem imaginalen System, r_i, größer aus als die des verbalen Systems, r_v.

In der Spalte ganz rechts sind die Differenzen-$G^2_{(2)}$-Statistiken zum Test der Nullhypothese angegeben, daß sich die entsprechenden Parameter der Zeile nicht zwischen den drei untersuchten Vpn-Gruppen unterscheiden. Offensichtlich gibt es praktisch keinerlei Unterschiede zwischen den Gruppen in bezug auf die Abrufwahrscheinlichkeiten und die Speicherungswahrscheinlichkeiten im verbalen System. Deutliche Vorteile der Kontrollgruppe sind dagegen in puncto imaginale Kodierung zu erkennen, während auch diesbezüglich die Unterschiede zwischen den beiden klinischen Gruppen eher gering sind.

So gut wie die Ergebnisse interpretierbar sind, so klar stehen sie im Widerspruch zur Lateralisierungshypothese von Paivio (1986), wonach rechtshemisphärische zerebrale Schädigungen mit einem markanten Imagery-Defizit einhergehen müßten. Zwar gibt es ein Defizit rechtstemporaler Epileptiker im Vergleich zur ge-

sunden Kontrollgruppe, jedoch ist es nicht stärker, sondern tendenziell sogar eher schwächer ausgeprägt als das einer vergleichbaren Epileptikergruppe mit linkshemisphärischem Fokus. Zerebrale Schäden im Zusammenhang mit Epilepsie scheinen also unabhängig vom Fokus mit Imagery-Problemen einherzugehen; einen Hinweis auf eine besondere Bedeutung des rechtshemisphärischen Bereichs gibt es nicht. Diese Ergebnisse stehen im Einklang mit Resultaten anderer Autoren (Ehrlichman & Barrett, 1983; Richardson, 1990).

Insgesamt dürfte das Beispiel die Nützlichkeit der MVB-Modellierung auch im Zusammenhang mit der Modellierung dynamischer Lern- und Gedächtnisprozesse belegen. Allerdings macht das Beispiel zugleich deutlich, daß strenge Stationaritätsannahmen, wie sie gewöhnlichen oder latenten MK-Modellen typischerweise anhaften, problematisch sein können. Es wurde gezeigt, daß man die Stationaritätsannahme im Rahmen eines MVB-Ansatzes fallen lassen und damit einen erheblichen Gewinn an Anpassungsgüte erzielen kann.

6.6 Gemischte und gemischte latente Markoff-Ketten-Modelle

Die Idee eines Markoff-Ketten-Modells (Abschnitt 6.5) kann mit der Idee eines Latent-Class-Modells (Abschnitt 6.2) in naheliegender Weise kombiniert werden: Man nimmt an, daß einer $J^{(T+1)}$-Kontingenztafel der möglichen Zustandskombinationen über $(T+1)$ Zeitpunkte hinweg nicht eine homogene Population zugrunde liegt, sondern vielmehr mehrere Subpopulationen („latente Klassen“), innerhalb derer verschiedene MK-Modelle mit unterschiedlichen „initial state probabilities” und Übergangswahrscheinlichkeitsmatrizen gelten. Dies ist die Idee eines gemischten Markoff-Ketten-Modells (Langeheine & van de Pol, 1990; van de Pol & Langeheine, 1990).

Das sogenannte Mover-Stayer-Modell (Blumen, Kogan & McCarthy, 1955; Davies & Crouchly, 1986) ist der wohl bekannteste Spezialfall dieser Modellklasse. Man nimmt hierbei eine latente Klasse für die „stayer” an, die ihren Ausgangszustand über den gesamten Prozeß hinweg beibehalten, mithin also durch eine Matrix von Übergangswahrscheinlichkeiten gekennzeichnet sind, die eine Einheitsmatrix mit Einsen in der Hauptdiagonalen und sonst nur Nullen ist. Daneben gibt es noch eine zweite latente Klasse von „movern”, die beliebige Zustandsübergangswahrscheinlichkeiten aufweisen können. Obwohl bislang meines Wissens ausschließlich in der Soziologie eingesetzt, kann das Mover-Stayer-Modell auch in der kognitiven Psychologie — speziell in der angewandten kognitiven Psychologie oder Neuropsychologie — von großem Nutzen sein. Man kann damit etwa die Größe einer Klasse von „nichtlernenden Personen” abschätzen, die von einem Lernprozeß abweichen, der durch eine bestimmte Matrix von Übergangswahrscheinlichkeit gekennzeichnet ist.

Für die kognitive Psychologie wichtiger als gemischte MK-Modelle dürften gemischte latente MK-Modelle sein (vgl. Langeheine & van de Pol, 1990; van de Pol & Langeheine, 1990). Genau wie latente MK-Modelle erlauben sie latente Zustände $X(t)$, die nur probabilistisch mit beobachtbaren Variablen $Y(t)$ verknüpft sind. Analog zu gemischten MK-Modellen basieren sie auf der Idee, daß in jeder von mehreren latenten Klassen ein bestimmten latentes MK-Modell gilt. Die Modellgleichung

$$\begin{aligned} &Pr\big(Y(T)=i(T), Y(T-1)=i(T-1),\ \ldots, Y(1)=i(1), Y(0)=i(0)\big) = \\ &\sum_{s=1}^{c}\sum_{j(0)}^{J}\sum_{j(1)}^{J}\cdots\sum_{j(T)}^{J}\alpha_s\cdot\pi_{j(0),s}\cdot p_{i(0)|j(0),s}\cdot\pi_{j(0),j(1),s}\cdot p_{i(1)|j(1),s}\cdot\ \cdots \\ &\qquad\cdots\ \cdot\ \pi_{j(T-1),j(T),s}\cdot p_{i(T)|j(T),s} \end{aligned} \tag{6.6-1}$$

ergibt sich unmittelbar aus der Modellgleichung (6.5-3) des latenten MK-Modells, wenn man (1) für alle Parameter einen Index s für die latente Klassenzugehörigkeit vorsieht, (2) Parameter α_s, $s = 1, ..., c$, für die Wahrscheinlichkeit der s-ten latente Klasse ergänzt und (3) zusätzlich über alle latenten Klassen summiert.

Van de Pol, Langeheine und de Jong (1991) haben ein sehr leistungsfähiges Computerprogramm PANMARK entwickelt, das sehr viele Modellvarianten von gemischten latenten MK-Modellen auf der Basis des EM-Algorithmus analysieren kann. Allerdings erlaubt es keine Reparametrisierungen der Übergangswahrscheinlichkeiten zwischen den latenten Zuständen in Termini binärer Parameter und ist deshalb z.B. für eine Analyse des Dual-Code-Markoff-Modells aus Abschnitt 6.5.1 — obwohl ein Modell mit nur einer latenten Klasse — nicht geeignet. Umgekehrt sind jedoch alle gemischten latenten MK-Modelle gemäß Modellgleichung (6.6-1) zugleich auch MVB-Modelle und lassen sich folglich mit MVB-Software wie z.B. AppleTree und GPT analysieren. Allerdings wird der Gewinn an Flexibilität bei der Definition verschiedener Modellvarianten, den man durch MVB-Programme zweifellos erzielt, durch großen Aufwand bei der Konstruktion der Modelleingabedateien erkauft, da sich gemischte latente MK-Modelle gewöhnlich durch extrem viele Äste des zugehörigen Baummodells auszeichnen. Sofern ein Modell also in den PANMARK-Modellrahmen paßt, ist man daher gut beraten, dieses Programm bei der Datenanalyse auch zu verwenden.

6.7 Verallgemeinerte Signalentdeckungsmodelle

Ein Argument, das immer wieder gegen kognitionspsychologisch motivierte MVB-Modelle vorgebracht wird, ist die Annahme mehrerer diskreter kognitiver Zustände („Finite-State-Annahme"), die mit dieser Modellklasse scheinbar unausweichlich verbunden zu sein scheint (z.B. Kinchla, 1994; Yonelinas & Jacoby, 1996a). Für kognitionspsychologische Theorien, die stetig verteilte kontinuierliche

Größen annehmen (z.B. Empfindungsstärken, Vertrautheitsgrade etc.) dürften MVB-Modelle demnach ein prinzipiell ungeeigneter Formalisierungsrahmen sein.

Obwohl oftmals stochastische Modelle konzipierbar sind, die kontinuierliche theoretische Größen besser als MVB-Modelle repräsentieren können (Ashby, 1992; Luce, 1986; Townsend & Ashby, 1983), trifft die oben genannte Kritik in ihrer Allgemeinheit nicht zu. Denkbar ist nicht nämlich, daß Modelle mit stetig verteilten theoretischen Größen bestimmten MVB-Modellen äquivalent sind und vice versa. Unter diesen Umständen kann die kontinuierliche Natur der angenommenen theoretischen Größen offenbar kein Apriori-Argument mehr gegen die MVB-Modellierung sein.

Beispiele hierfür sind die schon mehrfach erwähnten Ein- und das Zwei-Hochschwellen-Modelle der Rekognition, die oftmals als Prototypen für „Finite-State-Modelle" gelten. Wie unter anderem Macmillan und Creelman (1991, Kapitel 4) gezeigt haben, sind die Modellgleichungen beider Modelle aber auch mit verallgemeinerten Signalentdeckungsmodellen konsistent, die kontinuierlich verteilte Vertrautheitswerte anstelle diskreter kognitiver Zustände (Wiedererkennen, Raten) annehmen. Wie in der Standardsignalentdeckungstheorie (vgl. Abschnitt 3.3) wird hierbei unterstellt, daß alte und neue Items im Rekognitionstest durch zwei Vertrautheitsverteilungen gekennzeichnet sind, welche die Dichtefunktionen $f_a(x)$ und $f_n(x)$ aufweisen. Fällt ein konkreter Vertrautheitswert größer als ein Kriteriumswert c aus, so wird „alt" geantwortet, sonst „neu". Anders als in der Standardsignalentdeckungstheorie sind die beiden Dichten aber keine Normalverteilungen, die um den Betrag d' gegeneinander verschoben sind. Die Modellgleichungen des 1HT-Modells erhält man, wenn man stattdessen annimmt, daß ein Schwellenwert s_a existiert, oberhalb dessen die Dichtefunktion $f_n(x)$ immer null ist, während unterhalb von s_a beide Dichten proportional zueinander sind:

$$\begin{aligned} &f_n(x) = 0, \text{ falls } x \geq s_a \\ &f_n(x) = k \cdot f_a(x) \text{ mit } k > 1, \text{ falls } x < s_a \end{aligned} \qquad (6.7\text{-}1)$$

Der Kriteriumswert c kann irgendwo unterhalb von s_a liegen. Bezogen auf diese Interpretation sind die Parameter des 1HT-Modells einfach Flächen unterhalb bestimmter Bereiche der Vertrautheitsverteilungen, nicht mehr Wahrscheinlichkeiten diskreter kognitiver Zustände.

Im Gegensatz zum 1HT-Modell ist das 2HT-Modell konsistent mit einem Signalentdeckungsmodell, das zwei Schwellenwerte s_n und s_a vorsieht: Unterhalb von s_n ist die Dichte für alte Items null, oberhalb von s_a ist die Dichte für neue Items null und dazwischen sind die beiden Dichten gleich:

$$
\begin{aligned}
&f_n(x) = 0, \text{ falls } x \geq s_a \\
&f_n(x) = f_a(x), \text{ falls } s_n < x < s_a \\
&f_a(x) = 0, \text{ falls } x \leq s_n
\end{aligned}
\qquad (6.7\text{-}2)
$$

Bei diesem Modell kann das Antwortkriterium irgendwo zwischen s_n und s_a liegen und die Modellparameter sind wiederum Integrale zu bestimmten Abschnitten der Dichtefunktionen (vgl. die Abbildung in Macmillan & Creelman, 1991, Kapitel 4).

Als Fazit bleibt festzuhalten, daß sowohl die empirische Adäquatheit als auch die empirische Inadäquatheit eines MVB-Modells keinen direkten Schluß auf die Angemessenheit oder Unangemessenheit der Finite-State-Annahme ermöglicht. Ein MVB-Modell kann adäquat sein, obwohl ein Modell mit stetig verteilten theoretischen Größen zutrifft, das diesem Modell äquivalent ist. Umgekehrt kann ein MVB-Modell auch scheitern, ohne daß die Finite-State-Annahme falsch sein muß: Es kann andere Finite-State-Modelle geben, die für die betreffenden Kategoriewahrscheinlichkeiten adäquat sind. Die bloße Tatsache jedenfalls, daß zur Formalisierung ein MVB-Modell herangezogen wird, bedeutet nicht automatisch, daß die Möglichkeit stetig verteilter theoretischer Größen negiert wird. Wenn man also beispielsweise die Standardsignalentdeckungstheorie den Schwellenmodellen vorzieht, dann kann das Argument dafür nicht lauten, daß letztere mit stetig verteilten Vertrautheitswerten prinzipiell unvereinbar sind. Es kann allenfalls lauten, daß die Verteilungsannahmen, die ein Schwellenmodell erzwingt, wenn man es als Signalentdeckungsmodell für kontinuierliche Vertrautheitswerte interpretieren möchte, im Vergleich zur Normalverteilungsannahme inadäquat oder unplausibel erscheinen. Ob derartige Vorbehalte gegenüber Schwellenmodellen wirklich begründet sind, läßt sich nur auf der Basis von Modellanpassungstests zu gegebenen empirischen ROC-Daten entscheiden, nicht durch Apriori-Argumente.

6.8 Modelle für Antworthäufigkeiten und Antwortzeiten

Vielen kognitionspsychologischen Anwendungen von MVB-Modellen liegt die Idee zugrunde, daß die modellierten psychologischen Prozesse unterschiedlich viel Zeit beanspruchen, weil (a) unterschiedliche kognitive Zustände mit unterschiedlichen Antwortzeitverteilungen verknüpft sind oder (b) unterschiedlich viele Zustände sequentiell durchlaufen werden müssen, ehe eine Antwort gegeben werden kann. Oftmals werden Ereignisbaumdiagramme so konstruiert, daß jeder Weg, der von der Wurzel des Baumes in eine bestimmte Antwortkategorie führt, auch als serieller, sich über die Zeit erstreckender Prozeß interpretiert werden kann, bei dem mehrere Verarbeitungsstufen nacheinander durchlaufen werden müssen. Diese Interpretation bleibt bei der Anpassung des Modells an die Daten allerdings unberücksichtigt, weil nur die Antworthäufigkeiten, nicht aber die Antwortzeiten

in die Analyse eingehen. Alle MVB-Modelle, die identische Kategoriewahrscheinlichkeiten implizieren, sind datenäquivalent, auch wenn mit den Modellen ganz unterschiedliche Prozeßannahmen verbunden sind, die anhand von Antwortzeitverteilungen prinzipiell diskriminierbar wären. Wann immer man nicht nur Antwortwahrscheinlichkeiten, sondern kognitive Prozesse im eigentlichen Sinne des Wortes modellieren will, liegt es deshalb eigentlich nahe, zusätzlich zu den Wahrscheinlichkeiten der verschiedenen Beobachtungskategorien auch deren Antwort- bzw. Reaktionszeitverteilungen explizit im Modell zu berücksichtigen. Ein solcher Ansatz ermöglicht eine strengere Prüfung der zugrundeliegenden Prozeßannahmen als sie mit einem auf Antwortwahrscheinlichkeiten beschränkten Modell erreicht werden kann.

Allerdings gibt es keinen kanonischen Weg der Erweiterung von MVB-Modellen auf Antwortzeiten. Ein Ereignisbaumdiagramm kann — wie im folgenden noch deutlicher werden wird — mit unterschiedlichen Prozeßannahmen verknüpft werden, die auf unterschiedliche Antwortzeitverteilungen hinauslaufen. Einige dieser Prozeßannahmen führen zu Modellerweiterungen, die selbst wieder als MVB-Modelle darstellbar sind, während andere aus der MVB-Modellfamilie hinausführen und nur mit veränderten statistischen Techniken analysierbar sind. Beispiele für beide Formen der Modellerweiterung werden in den folgenden beiden Abschnitten behandelt: Modelle mit zustandsabhängigen Antwortzeiten (Abschnitt 6.8.1) und Modelle mit seriell-additiven Antwortzeiten (Abschnitt 6.8.2).

6.8.1 Modelle mit zustandsabhängigen Antwortzeiten

Eine Möglichkeit der Erweiterung von MVB-Modellen auf Antwortzeiten ergibt sich aus der Annahme, daß die Antwortzeitverteilungen der Probanden zustandsabhängig sind. Der „Zustand” einer Vp entspricht dem Knoten im Ereignisbaumdiagramm, von dem aus eine bestimmte Antwort generiert wurde. Es handelt sich hierbei also immer um einen Knoten, der den Antwortkategorien direkt vorgeordnet ist. Sofern diesem Knoten selbst wieder andere Knoten im Ereignisdiagramm vorausgehen, bleiben letztere bei der Modellierung der Antwortzeiten unberücksichtigt. Es interessieren nur die Zustände, die direkt in bestimmte Antworten einmünden können. Wenn es m derartige Zustände Z_i, $i = 1, ..., m$, im Modell gibt, ist jedem dieser Zustände eine bestimmte Antwortzeitverteilung mit Verteilungsfunktion $F_i(t) := Pr(T \leq t)$ zugeordnet. Hierbei bezeichnet T die Zufallsvariable (ZV) „Antwortzeit” und t eine konkrete Realisation dieser ZV. Die Verteilung hängt wohlgemerkt nur vom Zustand Z_i, nicht aber von der Antwort ab, die letztlich gegeben wurde. Wenn diese Annahme in einem konkreten Einzelfall zu restriktiv erscheint, muß man unterschiedliche Zustände für unterschiedliche

Antwortkategorien vorsehen. Maximal kann *m* gleich der Anzahl der Äste im zugehörigen Ereignisbaumdiagramm werden. Im allgemeinen wird *m* jedoch kleiner als die Astanzahl sein, weil es einige Zustände gibt, die in mehrere Antwortkategorien einmünden können.

Konkrete Verteilungsannahmen bzw. Annahmen über Mittelwerte oder (homogene) Varianzen sind im allgemeinen nicht erforderlich. Die entsprechenden Momente sollen lediglich existieren, können aber völlig beliebig ausfallen. Damit die $F_i(t)$ als Verteilungsfunktionen von Antwortzeiten interpretierbar sind, fordern wir lediglich, daß $F_i(t) = 0$ für $t \leq 0$ und $0 \leq F_i(t) \leq 1$ für $t \geq 0$ und alle Zustände Z_i gilt; ferner muß natürlich $F_i(t)$ für $t \geq 0$ monoton steigen. Die Verteilung der beobachteten Antwortzeiten in einer bestimmten Antwortkategorie ergibt sich dann einfach als finite Mischung der zustandsabhängigen Antwortzeitverteilungen, die in der betreffenden Antwortkategorie potentiell vorkommen können (einführend zum Konzept der finiten Mischverteilung: Rost & Erdfelder, 1996). Die Anteile der zustandsabhängigen Antwortzeitverteilungen in der Mischung entsprechen gerade den bedingten Wahrscheinlichkeiten der verschiedenen Zustände Z_i, gegeben daß eine Antwort in die betreffende Kategorie fällt.

Das auf alte Items beschränkte Hochschwellen-Modell der Rekognition (vgl. Abschnitt 3.5.2) sei als Beispiel herangezogen, um die Modellannahmen zu konkretisieren. Werden nur alte Items berücksichtigt, ist es als Modell für Antwortwahrscheinlichkeiten nicht identifizierbar (vgl. Abschnitt 4.1), da einer freien Kategoriewahrscheinlichkeit — der Trefferrate — zwei freie Modellparameter *r* (Rekognitionsrate) und *b* (Antworttendenz zugunsten von „Alt"-Antworten) gegenüberstehen. Wenn man das Modell jedoch auf ein Modell mit zustandsabhängigen Antwortzeitverteilungen erweitert, wird es mit wenigen Zusatzannahmen nicht nur identifizierbar, sondern sogar testbar, ohne daß Distraktoritems zusätzlich berücksichtigt werden müssen.

Wie können Antwortzeiten in das Modell integriert werden? Das Hochschwellen-Modell läßt sich zwanglos als ein Modell mit zwei Zuständen interpretieren, nämlich dem „Zustand der Rekognition" (Z_r), der immer zu einer „Alt"-Antwort führt, und dem „Zustand des Ratens" (Z_b), der mit Wahrscheinlichkeit *b* zu einer „Alt"-Antwort und mit Wahrscheinlichkeit 1–*b* zu einer „Neu"-Antwort führt. Angenommen wird, daß die Antwortzeit im Zustand Z_r der Verteilungsfunktion $F_r(t)$ und die Antwortzeit im Zustand Z_b der Verteilungsfunktion $F_b(t)$ folgt, unabhängig davon, ob „alt" oder „neu" geraten wird. Diese Modellvorstellung ist nicht neu; sie ist bereits in den 60er Jahren als „Fast-Guess-Modell" bekannt geworden (Ollman, 1966; Yellott, 1971). Die Bezeichnung resultiert aus der Annahme, daß der Ratezustand Z_b typischerweise mit kürzeren Antwortzeiten assoziiert ist als der Rekognitionszustand Z_r.

Oftmals ist das Fast-Guess-Modell zur Schätzung der Erwartungswerte der zustandsabhängigen Antwortzeitverteilungen eingesetzt worden. Ein Ergebnis, welches die Validität der Modellannahmen überzeugend stützt, ist der Befund, daß Manipulationen des Geschwindigkeits-Genauigkeits-Austauschs selektiv auf die Zustandswahrscheinlichkeit r bzw. $1–r$ wirken, während die Erwartungswerte der zustandsabhängigen Antwortzeitverteilungen unverändert bleiben (z.B. Yellott, 1971). Wenn man also die Vpn durch unterschiedliche Instruktionen oder variierende Response-Deadlines zu laxem versus gründlichem Vorgehen motiviert, dann ändert das — modellimmanent betrachtet — lediglich die Bereitschaft zu „fast guesses" (r-Parameter) und sonst nichts. Trotz der Konstanz der mittleren Antwortzeiten im Zustand Z_r und Z_b wird bei laxem Vorgehen die durchschnittliche „Alt"-Antwortzeit kürzer, weil ein größerer Anteil dieser „Alt"-Antworten „fast guesses" aus dem Zustand Z_b heraus sind.

Link (1982, S. 476ff) hat mehrere Möglichkeiten diskutiert, wie man die Parameter r und b des Fast-Guess-Modells schätzen kann, wenn dazu keine Apriori-Hypothesen (z.B. $b = .50$) existieren. Erfreulicherweise kann die Grundidee von Link (1982) ohne viel Aufwand so modifiziert werden, daß erneut ein binäres MVB-Modell resultiert, welches mit der auf diese Modellklasse zugeschnittenen Software relativ leicht analysierbar ist. Hierzu muß die Antwortzeitvariable zunächst diskretisiert werden. Das kann durch Festlegung positiv-reellwertiger Grenzwerte $c_1 < c_2 < ... < c_{(C-1)}$ erreicht werden kann, die das Antwortzeitkontinuum in C Antwortzeitkategorien einteilen. Nehmen wir beispielsweise an, daß mit drei Grenzwerten vier Antwortzeitkategorien definiert werden, nämlich sehr schnelle Antworten ($T \leq c_1$), schnelle Antworten ($c_1 < T \leq c_2$), langsame Antworten ($c_2 < T \leq c_3$) und sehr langsame Antworten ($T > c_3$). Für das zu konstruierende MVB-Modell bilden „Alt"-Antworten, welche in die genannten vier Kategorien fallen, die Beobachtungskategorien C_1 bis C_4, während „Nein"-Antworten die Beobachtungskategorien C_5 bis C_8 bilden. Zusätzlich zu den Modellparametern r und b müssen Parameter zur Charakterisierung der zustandsabhängigen Antwortzeitverteilungen festgelegt werden. Für die Antwortzeitverteilung des Rekognitionszustands Z_r sei die Wahrscheinlichkeit $Pr(T \leq c_2)$, daß die Antwortzeit in Kategorie C_1 oder C_2 fällt, mit t_{r2} abgekürzt. Die bedingte Wahrscheinlichkeit $Pr(T \leq c_1 \,|T \leq c_2)$ wird mit t_{r1} und die bedingte Wahrscheinlichkeit $Pr(T \leq c_3 \,|T > c_2)$ mit t_{r3} notiert. Die analogen Wahrscheinlichkeiten für die Antwortzeitverteilung der geratenen Antworten seien, t_{b1}, t_{b2} und t_{b3}. Die Modellgleichungen für sehr schnelle (C_1), schnelle (C_2), langsame (C_3) und sehr langsame (C_4) „Alt"-Antworten ergeben sich nunmehr wie folgt:

$$
\begin{aligned}
Pr(C_1) &= p_1 = r \cdot t_{r_2} \cdot t_{r_1} + (1-r) \cdot b \cdot t_{b_2} \cdot t_{b_1} \\
Pr(C_2) &= p_2 = r \cdot t_{r_2} \cdot (1-t_{r_1}) + (1-r) \cdot b \cdot t_{b_2} \cdot (1-t_{b_1}) \\
Pr(C_3) &= p_3 = r \cdot (1-t_{r_2}) \cdot t_{r_3} + (1-r) \cdot b \cdot (1-t_{b_2}) \cdot t_{b_3} \\
Pr(C_4) &= p4 = r \cdot (1-t_{r_2}) \cdot (1-t_{r_3}) + (1-r) \cdot b \cdot (1-t_{b_2}) \cdot (1-t_{b_3})
\end{aligned}
\qquad (6.8.1\text{-}1)
$$

Die Kategoriewahrscheinlichkeiten sind jeweils Mischungen derjenigen Anteile von rekognizierten bzw. geratenen Antworten, die in eine bestimmte Antwortzeitkategorie fallen. Die Mischungsgewichte betragen r (rekognizierte Antworten) und $(1–r) \cdot b$ (geratene „Alt"-Antworten). Die entsprechenden Modellgleichungen für sehr schnelle (C_5), schnelle (C_6), langsame (C_7) und sehr langsame (C_8) „Neu"-Antworten ergeben sich ganz analog:

$$
\begin{aligned}
Pr(C_5) &= p_5 = (1-r) \cdot (1-b) \cdot t_{b_2} \cdot t_{b_1} \\
Pr(C_6) &= p_6 = (1-r) \cdot (1-b) \cdot t_{b_2} \cdot (1-t_{b_1}) \\
Pr(C_7) &= p_7 = (1-r) \cdot (1-b) \cdot (1-t_{b_2}) \cdot t_{b_3} \\
Pr(C_8) &= p_8 = (1-r) \cdot (1-b) \cdot (1-t_{b_2}) \cdot (1-t_{b_3})
\end{aligned}
\qquad (6.8.1\text{-}2)
$$

Hierbei handelt es sich allerdings nicht um Mischungen, da „Nein"-Antworten nur aus dem Ratezustand heraus erwachsen können.

Die acht Modellgleichungen zusammengenommen definieren zwar ein MVB-Modell, welches jedoch ohne weitere Parameterrestriktionen nicht identifizierbar ist: Sieben funktional unabhängigen Kategoriewahrscheinlichkeiten stehen acht freie Parameter gegenüber. Glücklicherweise läßt sich dieses Problem dadurch lösen, daß man die Grenzwerte c_i, welche die Antwortzeitkategorien definieren, geschickt wählt (Link, 1982). Eine naheliegende Idee ist beispielsweise, die c_i-Werte so zu wählen, daß die Restriktion $t_{r1} = t_{r2} = t_{r3} = .5$ erfüllt ist. Diese Wahl ist vorteilhaft, weil die drei Grenzwerte dann einfach interpretierbar sind: Sie markieren das 25%-Perzentil (c_1), den Median (c_2) und das 75%-Perzentil (c_3) der Antwortzeitverteilung rekognizierter Antworten.

Wie kann man Grenzwerte mit den gewünschten Eigenschaften finden? Wenn man den Hilfsparameter $h := b / (1–b)$ einführt und die Modellgleichungen (6.8.1-2) in die Modellgleichungen (6.8.1-1) einsetzt, ergibt sich:

$$
\begin{aligned}
p_1 &= r \cdot t_{r_2} \cdot t_{r_1} + h \cdot p_5 \\
p_2 &= r \cdot t_{r_2} \cdot (1-t_{r_1}) + h \cdot p_6 \\
p_3 &= r \cdot (1-t_{r_2}) \cdot t_{r_3} + h \cdot p_7 \\
p_4 &= r \cdot (1-t_{r_2}) \cdot (1-t_{r_3}) + h \cdot p_8
\end{aligned}
\qquad (6.8.1\text{-}3)
$$

Trägt man die Trefferraten p_1 bis p_4 auf der y-Achse gegen die zugehörigen Verpasserraten p_5 bis p_8 auf der x-Achse ab, so müssen sie folglich auf einer Geraden mit Steigung h und Ordinatenschnittpunkt $r \cdot .50 \cdot .50 = r \cdot .25$ liegen, wenn (a) das Mo-

dell gilt und (b) die drei Grenzwerte richtig gewählt wurden, so daß $t_{r1} = t_{r2} = t_{r3} = .50$ erfüllt ist. Um die drei Grenzwerte geeignet wählen zu können, benötigt man eine vorläufige Schätzung des Hilfsparameters *h* bzw. des Antworttendenzparameters *b*. Eine Möglichkeit der Schätzung ist die Herstellung einer speziellen Versuchsbedingung, in welcher der Rekognitionsparameter *r* null wird. Man könnte beispielsweise an eine extrem kurzfristige („subliminale") Darbietung der Items in der Lernphase denken. Gemäß Gleichungssystem (6.8.1-3) müssen für eine solche Versuchsbedingung p_1 bis p_4 zu p_5 bis p_8 proportional sein, wobei die Proportionalitätskonstante gerade *h* entspricht. Durch Anpassung einer Regressionsgerade durch den Ursprung an die Trefferraten und Verpasserraten der verschiedenen Antwortzeitkategorien kann dann eine Schätzung für *h* abgeleitet werden. Eine andere, noch einfachere Möglichkeit ist gegeben, wenn man einen Grenzwert c_0 für die Antwortzeiten festlegen kann, der aus dem Zustand Z_r heraus niemals unterschritten wird, wohl aber aus dem Ratezustand Z_b. Den Bereich $T < c_0$ unterteilt man dann ebenfalls in drei oder vier Antwortzeitkategorien und zählt die entsprechenden Treffer- und Verpasserraten in diesen Zeitkategorien aus. Da *r* in diesem Bereich qua Voraussetzung null sein muß, kann man mit einer Regression durch den Ursprung hier ebenfalls zu einer Schätzung von *h* gelangen.

Hat man auf einem dieser Wege eine vorläufige Schätzung $\hat{h}^*$ für *h* abgeleitet, setzt man sie zusammen mit den Treffer- und Verpasserraten für die Ausgangsstichprobe in das Gleichungssystem (6.8.1-3) ein und berechnet die Differenzen

$$\begin{aligned} &\hat{p}_1 - \hat{h}^* \cdot \hat{p}_5 \\ &\hat{p}_2 - \hat{h}^* \cdot \hat{p}_6 \\ &\hat{p}_3 - \hat{h}^* \cdot \hat{p}_7 \\ &\hat{p}_4 - \hat{h}^* \cdot \hat{p}_8 \end{aligned} \qquad (6.8.1\text{-}4)$$

Die Grenzwerte c_1 bis c_3 werden nun so angepaßt, daß alle vier Differenzen gleich groß werden: Durch Veränderung von c_1 kann man die relative Größe ersten beiden Differenzen, durch Veränderung von c_3 die der zweiten beiden Differenzen beeinflussen; c_2 schließlich beeinflußt die Beziehung der ersten beiden zu den letzten beiden Differenzen.

Der nächste Schritt besteht darin, das MVB-Modell mit den acht Modellgleichungen (6.8.1-1) und (6.8.1-2) und den Parameterfixierungen $t_{r1} = t_{r2} = t_{r3} = .50$ an die Häufigkeiten anzupassen, die aus den optimierten Grenzwerten resultieren.[76] Auf diese Weise werden endgültige Schätzer für *r* und *b* gewonnen. Hierbei sollte der Schätzer für *b* allerdings in der Nähe des vorläufigen Schätzers liegen, der in

[76] In Abschnitt 8 des Anhangs ist eine geeignete EQN-Datei zur Eingabe des Modells in die Computerprogramme MBT, GPT oder AppleTree wiedergegeben.

Gleichung (6.8.1-4) Verwendung fand. Ist dies nicht der Fall, so ist dies ein Hinweis auf falsche Voraussetzungen bei der Bestimmung der optimierten Grenzwerte.

Das Problem der statistischen Modellprüfung wurde von Link (1982) nicht explizit diskutiert. Natürlich bereitet es keine Schwierigkeiten, eine beliebige Modellanpassungsstatistik für das restringierte Modell zu berechnen. Da nunmehr nur noch fünf Parameter frei sind (r, b, t_{b1}, t_{b2} und t_{b3}), hat der Anpassungstest bei sieben funktional unabhängigen Kategoriewahrscheinlichkeiten rein rechnerisch zwei Freiheitsgrade. Die inferentielle Beurteilung der Anpassungsstatistik anhand der $\chi^2_{(2)}$-Verteilung ist jedoch problematisch, wenn die optimierten c_i-Grenzwerte anhand der gleichen Stichprobe geschätzt wurden, für die anschließend auch der Modellanpassungstest durchgeführt wird. Der Test ist dann für die Nullhypothese $t_{r1} = t_{r2} = t_{r3}$ voreingenommen. Besser ist es, wenn die Grenzwertoptimierung in einer Pilotuntersuchung erfolgt und die Modellanpassungsprüfung anhand einer unabhängig gewonnenen Stichprobe vorgenommen wird. Wird so verfahren, so ist die Interpretation der Modellanpassungsstatistik unproblematisch.

Die Erweiterung der Methode auf mehr als vier Antwortzeitkategorien liegt auf der Hand. Die Ergebnisse sollten sich — wenn das Modell gilt und die Grenzwertoptimierung analog erfolgt — bezüglich r und b nur unwesentlich unterscheiden. Der Vorteil einer feineren Kategorienaufteilung besteht im wesentlichen darin, daß man die Antwortzeitverteilungen $F_r(t)$ und $F_b(t)$ hierbei genauer abschätzen kann. Wenn man an der genauen Form und Lage der Verteilungsfunktionen interessiert ist, ist es natürlich effizienter, $F_r(t)$ und $F_b(t)$ von vornherein auf bestimmte Verteilungsfamilien zu beschränken (vgl. dazu Link, 1982). Allerdings dürfte der substanzwissenschaftliche Hintergrund — von den im nächsten Abschnitt zu besprechenden Ausnahmen abgesehen — nur selten eine überzeugende Begründung für derartige Verteilungsannahmen hergeben.

Die Möglichkeiten der Modellvalidierung sind vielfältig. Yellott (1991) hat das Fast-Guess-Modell durch Manipulationen des Genauigkeitskriteriums der Vpn validiert, indem er zeigte, daß derartige Maßnahmen selektiv auf den Modellparameter r wirken. Weitere Validierungsmöglichkeiten bestehen darin, den Antworttendenzparameter b oder die Schwierigkeit des Diskriminationsproblems zu manipulieren, die ausschließlich auf r und $F_r(t)$ wirken sollte. Schließlich kann das Fast-Guess-Modell auch dazu verwendet werden, den Distraktorentdeckungsparameter des 2HT-Modells zu schätzen: Wenn die Antworttendenz b aus dem Fast-Guess-Modell vorhergesagt werden kann, kann die Rate falschen Alarms zur Schätzung der unrestringierten Distraktorentdeckungswahrscheinlichkeit benutzt werden. Eine Gleichsetzung der Distraktorentdeckungsrate mit der Rekognitionsrate r ist nicht mehr erforderlich.

Prinzipiell kann die anhand des Fast-Guess-Modells illustrierte Vorgehensweise auf jedes kognitionspsychologisch motivierte MVB-Modell übertragen wer-

den. Man muß lediglich zusätzlich zu den Antworten die Antwortzeiten erheben und geeignet gewählten Antwortzeitkategorien zuweisen. Für die so ergänzte empirische Datenbasis formuliert man ein erweitertes MVB-Modell, das nicht nur den Pfad in eine bestimmte Antwortkategorie, sondern zusätzlich den Pfad in eine Antwortzeitkategorie angibt. Identifizierbarkeitsprobleme lassen sich leicht vermeiden, wenn experimentelle Bedingungen realisierbar sind, in denen jeweils nur einer der hypothetisch angenommenen Zustände Z_i, $i = 1, ..., m$, isoliert vorkommt. Anhand solcher „zustandsreinen" Bedingungen lassen sich die Antwortzeitverteilungen für die verschiedenen psychologischen Zustände schätzen. Kennt man diese „basis distributions", so ist es kein Problem, für weitere experimentelle Bedingungen, in denen mehrere Zustände vorkommen können, die Wahrscheinlichkeiten dieser Zustände als Mischungsanteile der „basis distributions" zu bestimmen (Yantis, Meyer & Smith, 1991). Apriori-Annahmen über die Form und Lage der „basis distributions" sind hierzu nicht erforderlich.

Eine Anwendung dieser Methode haben Yantis et al. (1991) zusammenfassend berichtet. In ihrem ersten Experiment mußten Vpn auf einen Pfeil, der nach links wies, möglichst schnell mit dem Druck der linken Taste und auf einen Pfeil, der nach rechts wies, möglichst schnell mit dem Druck der rechten Taste reagieren (einfache Wahlreaktionszeitaufgabe). Die Reaktion erfolgte in Bedingung 1 unvorbereitet, in Bedingung 2 durch einen 700 Millisekunden zuvor dargebotenen Hinweisreiz perfekt vorbereitet und in Bedingung 3 durch einen nur kurz zuvor dargebotenen Hinweisreiz kurzfristig vorbereitet. Der Hinweisreiz war entweder ein Wort, das immer den Rechts-Pfeil ankündigte, oder ein Nonwort, das immer den Links-Pfeil ankündigte. Die Antwortzeitverteilung in Bedingung 3 lies sich als Zwei-Komponenten-Mischung der „basis distributions" aus den Bedingungen 1 und 2 darstellen. Antwortpräparierung scheint demnach ein Alles-oder-Nichts-Prozeß zu sein: Entweder ist die Antwort in einem konkreten Durchgang vorbereitet, was mit einer charakteristischen Verteilung kurzer Antwortzeiten assoziiert ist, oder sie ist nicht vorbereitet, was mit einer charakteristischen Verteilung langer Antwortzeiten verknüpft ist. Antworten mit dazwischenliegenden Antwortzeitverteilungen gibt es nicht. Mehr oder minder gute partielle Vorbereitung einer Antwort drückt sich nur darin aus, daß sich die Mischungsanteile perfekt vs. nicht vorbereiteter Antworten ändern.

Anders sah es in dem zweiten von Yantis et al. (1991) berichteten Experiment aus, in dem Wort-Nonwort-Entscheidungen zu treffen waren, wobei die Wörter durch semantisch hoch assoziierte Hinweiswörter angekündigt (Bedingung 2 und 3) bzw. nicht angekündigt wurden (Bedingung 1). Für diesen Fall ließen sich die partiell vorbereiten Antwortzeiten der Bedingung 3 nicht als Mischung der „basis distributions" aus den Bedingung 1 und 2 darstellen. Die Er-

gebnisse sprechen hier eher für ein graduelles Modell der lexikalen Bahnung und gegen ein Alles -oder-Nichts-Modell (Yantis et al., 1991).

6.8.2 Modelle mit additiven Antwortzeiten

Eine naheliegende Alternative zur Idee zustandsabhängiger Antwortzeitverteilungen ist das Konzept sequentiell-additiver Antwortzeiten: Jeder Ast eines Ereignisbaumdiagramms entspricht demnach einer zeitlichen Sequenz von L kognitiven Teilprozessen, deren „Abarbeitung" jeweils eine Zeitspanne $T_l = t_l$ in Anspruch nimmt. Ist ein Teilprozeß abgeschlossen, so beginnt unmittelbar der darauf folgende Teilprozeß, bis der gesamte Ast abgearbeitet ist. Faßt man die einzelnen Zeitkomponenten T_l als Zufallsvariablen mit existierenden Erwartungswerten $E(T_l)$ auf, so ist die Gesamtdauer T, die mit dem kompletten Ast assoziiert ist, ebenfalls eine Zufallsvariable $T = T_1 + T_2 + ... + T_L$ mit Erwartungswert $E(T) = E(T_1) + E(T_2) + ... + E(T_L)$.

Eine einfache Situation ist gegeben, wenn wir es mit einem degenerierten „einästigen Ereignisbaum" zu tun haben, der nur in eine einzige Antwortkategorie (z.B. „Aufgabe gelöst") einmündet. In diesem Fall interessieren nicht die Wahrscheinlichkeiten der involvierten Teilprozesse — diese sind zwangsläufig gleich eins, weil jeder Teilprozeß nur einen Folgeprozeß zuläßt —, sondern ausschließlich deren Latenzen T_l. Man kann die Erwartungswerte $E(T_l)$ z.B. mit der Methode additiver Faktoren von Sternberg (z.B. 1998) schätzen, indem man die Dauer der einzelnen Prozesse selektiv durch bestimmte experimentelle Manipulationen beeinflußt. Werden $k \leq L$ experimentelle Faktoren, welche jeweils die Dauer eines Teilprozesses selektiv beeinflussen, komplett miteinander gekreuzt, so muß aufgrund der behaupteten Additivität der Reaktionszeitkomponenten ein reines Haupteffektmodell ohne jede Interaktion für die Antwortzeitmittelwerte des k-faktoriellen Versuchsplans gelten.[77] Ist dies erfüllt, lassen sich die Erwartungswerte $E(T_l^{(j)})$, die mit der j-ten Stufe des Faktors zum l-ten Teilprozeß assoziiert sind, aus dem Gesamtmittelwert und den Haupteffekten schätzen. Allerdings ist das Gesamtmodell nur dann identifizierbar, wenn zu einigen dieser Erwartungswerte Apriori-Annahmen vorausgesetzt werden.

Ein einfacher handhabbarer Spezialfall der additiven Faktorenmethode ist gegeben, wenn — wie bei der auf Donders zurückgehenden Subtraktionsmethode — einzelne Teilprozesse nicht nur verlangsamt oder beschleunigt, sondern komplett eliminiert werden können. Beispielsweise hat Bredenkamp (1990, S. 64 ff) in seinen Untersuchungen eines Rechenkünstlers u.a. sechs verschiedene Versuchs-

[77] Die statistische Prüfung dieser Hypothese ist allerdings nicht ganz unproblematisch, da die Fehlerverteilungsannahmen des Allgemeinen Linearen Modells offenkundig verletzt sind. Lösungsvorschläge findet man bei Townsend und Ashby (1983) und Sternberg (1998).

bedingungen *A* bis *F* realisiert, die sich hinsichtlich Art und Anzahl der erforderlichen Teilprozesse bis zur Antwort auf eine Kopfrechenaufgabe unterschieden. Die Vp sollte möglichst schnell entscheiden, ob eine vorgegebene dreistellige Zahl die Lösungsendung der 137. Wurzel aus einer 1000-stelligen Zahl ist. Die auf dem Prinzip der kognitiven Ökonomie aufbauende Analyse möglicher Rechenwege durch Bredenkamp (1990) hatte zuvor gezeigt, daß zur Ableitung der korrekten Antwort immer nur relativ wenige elementare Regelanwendungen erforderlich sind. Er stellte daher die Hypothese auf, daß sich die Antwortzeiten des Rechenkünstlers als Summen der Latenzzeiten jener Regelanwendungen darstellen lassen, die zur Bewältigung der Aufgabe mit der minimalen Regelmenge erforderlich sind. Da einige Regeln in mehreren Versuchsbedingungen benötigt wurden, konnten die Erwartungswerte der entsprechenden Latenzzeiten über die betreffenden Versuchsbedingungen hinweg gleichgesetzt werden. Das vollständige Modell für alle sechs Bedingungen *A* bis *F* lautete:

$$\begin{aligned}
E_A(T) &= E(T_0) \\
E_B(T) &= E(T_0) + E(T_1) \\
E_C(T) &= E(T_0) + E(T_1) + E(T_2) \\
E_D(T) &= E(T_0) + E(T_3) \\
E_E(T) &= E(T_0) + E(T_1) + E(T_4) \\
E_F(T) &= E(T_0) + E(T_1) + E(T_2) + E(T_5)
\end{aligned} \qquad (6.8.2\text{-}1)$$

Schätzbar sind die durchschnittlichen Regelanwendungslatenzen ($E(T_0)$ bis $E(T_5)$) z.B. mit dem Kleinst-Quadrat-Methode im Rahmen des Regressionsmodells:

$$T = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + b_4 \cdot X_4 + b_5 \cdot X_5 + e. \qquad (6.8.2\text{-}2)$$

Hierbei sind die Prädiktoren X_j, j = 1, ..., 5, Dummy-Variablen, die bei allen Aufgaben den Wert X_j = 1 annehmen, bei denen der j-te Teilprozeß (bzw. die Anwendung der j-ten Regel) involviert ist; bei allen anderen Aufgaben ist X_j = 0 zu setzen. Mit dieser Wahl der Dummy-Variablen geht das Regressionsmodell (6.8.2-2) in das additive Antwortzeitmodell (6.8.2-1) über, wobei für die Regressionskoeffizienten $b_j = E(T_j)$, j = 0, ..., 5, gilt. Wenn die Dummy-Variablen nicht linear abhängig sind, ist das Modell identifizierbar und die Modellparameter sind somit schätzbar. Die Modellprüfung ist — wie immer bei Anwendungen der additiven Faktorenmethode — nicht ganz unproblematisch, da das Fehlermodell für *e* im Rahmen des Allgemeinen Linearen Modells inadäquat ist (vgl. auch Fußnote 77). Dieses technische Problem läßt sich jedoch mit robusten Alternativen zur Regressionsanalyse in den Griff bekommen.

Das von Bredenkamp (1990) benutzte Verfahren der Antwortzeitmodellierung kann auf Ereignisbaummodelle mit mehreren Ästen pro Versuchsbedingung übertragen werden, wenn man voraussetzen kann, daß die Latenzen der einzelnen

Prozeßkomponenten von ihren Wahrscheinlichkeiten unabhängig sind. Unproblematisch ist das Verfahren dann, wenn die Parameter des zugehörigen MVB-Modells bekannt sind bzw. als bekannt vorausgesetzt werden können. Der Einfachheit halber soll das Prinzip anhand des 1HT-Modells der Rekognition illustriert werden, welches im letzten Abschnitt bereits den Ausgangspunkt zur Ableitung des Fast-Guess-Modells bildete. Während dort davon ausgegangen wurde, daß Antworten aus dem Ratezustand heraus besonders schnell erfolgen, wird hier allerdings angenommen, daß — im Einklang mit dem Ereignisbaumdiagramm zum 1HT-Modell — immer zuerst ein Rekognitionsprozeß mit Latenz T_r abläuft. Nur wenn dieser scheitert, folgt zusätzlich ein Rateprozeß mit Latenz T_b. Obwohl im Rahmen sequentiell-additiver Modelle nicht zwingend notwendig, soll der Einfachheit halber vorausgesetzt werden, daß erfolgreiche und nicht erfolgreiche Rekognitionsprozesse die gleiche Latenzverteilung aufweisen. Dies ist z.B. dann plausibel, wenn man annimmt, daß Rekognitionsprozesse auf exhaustiven Suchprozessen basieren (zur Unterscheidung selbstterminierender und exhaustiver Prozesse vgl. z.B. Townsend & Ashby, 1983, Kap. 2 und 7). Ebenso soll vorausgesetzt werden, daß die Latenz der Rateprozesse unabhängig von deren Ergebnis ist. Qualitativ ergibt sich aus den Annahmen somit eine Prognose, die dem Fast-Guess-Modell genau entgegengesetzt ist: Geratenen Antworten korrespondieren durchschnittlich längere (stochastisch größere) Antwortzeiten als rekognizierten Antworten.

Wie erhält man unter diesen Modellvoraussetzungen Kleinst-Quadrat-Schätzer der durchschnittlichen Latenzen $E(T_r)$ für Rekognitionsprozesse und $E(T_b)$ für Rateprozesse? Zunächst einmal sind die Modellgleichungen für die mittleren „Ja"- und „Nein"-Antwortzeiten abzuleiten. Beginnen wir mit den Treffer-Antwortzeiten. Treffer erfolgen mit Wahrscheinlichkeit r_{1ht} durch ein erfolgreichen Rekognitionsprozeß mit durchschnittlicher Latenz $E(T_r)$ und mit Wahrscheinlichkeit $(1-r_{1ht}) \cdot b_{1ht}$ durch einen erfolgreichen Rateprozeß nach gescheitertem Rekognitionsprozeß. Dieser hat die durchschnittliche Gesamtdauer $E(T_r) + E(T_b)$. Die durchschnittliche Antwortzeit bei *allen* Treffern ist folglich eine Mischung der folgenden Art:

$$\begin{aligned} E_{hit}(T) &= p(\text{Rekognitionspr.}|\text{hit}) \cdot E(T_r) + p(\text{Ratepr.}|\text{hit}) \cdot \big(E(T_r) + E(T_b)\big) \\ &= \frac{r_{1ht}}{r_{1ht} + (1-r_{1ht}) \cdot b_{1ht}} \cdot E(T_r) + \frac{(1-r_{1ht}) \cdot b_{1ht}}{r_{1ht} + (1-r_{1ht}) \cdot b_{1ht}} \cdot \big(E(T_r) + E(T_b)\big) \qquad (6.8.2\text{-}3) \\ &= E(T_r) + E(T_b) \cdot \frac{(1-r_{1ht}) \cdot b_{1ht}}{r_{1ht} + (1-r_{1ht}) \cdot b_{1ht}} \ . \end{aligned}$$

Völlig analog kann man die Modellgleichungen für Verpasser, falsche Alarme und korrekte Zurückweisungen ableiten. Da hierbei im Falle des 1HT-Modells je-

weils nur ein Ast in die Zielkategorie führt, ergibt sich die durchschnittliche Antwortzeit einfach als Summe der durchschnittlichen Teilprozeßzeiten:

$$\begin{aligned} E_{\text{miss}}(T) &= E(T_{\text{r}}) + E(T_{\text{b}}) \\ E_{\text{false alarm}}(T) &= E(T_{\text{r}}) + E(T_{\text{b}}) \\ E_{\text{corr. reject.}}(T) &= E(T_{\text{r}}) + E(T_{\text{b}}) \end{aligned} \qquad (6.8.2\text{-}4)$$

Man beachte, daß aus diesen Modellgleichungen unmittelbar eine testbare Prognose ableitbar ist, nämlich die Gleichheit der durchschnittlichen Antwortzeiten bei Verpassern, falschen Alarmen und korrekten Zurückweisungen.

Die Schätzung der mittleren Antwortzeitkomponenten kann in diesem Fall auf der Basis des einfachen linearen Regressionsmodells $T = b_0 + b_1 \cdot X + e$ erfolgen. In die Regressionsanalyse gehen die Antwortzeiten aller vier Itemklassen ein. Die Prädiktorvariable X ist jeweils die bedingte Wahrscheinlichkeit eines Rateprozesses. Für Verpasser, falsche Alarme und korrekte Zurückweisungen wird sie folglich auf $X = 1$ gesetzt. Für Treffer setzt man den Prädiktorwert auf $X = (1-r_{1\text{ht}}) \cdot b_{1\text{ht}} / (r_{1\text{ht}} + (1-r_{1\text{ht}}) \cdot b_{1\text{ht}})$. Damit reduziert sich das einfache Regressionsmodell auf das Modell gemäß (6.8.2-3) und (6.8.2-4), wobei $b_0 = E(T_{\text{r}})$ und $b_1 = E(T_{\text{b}})$. Die üblichen Kleinst-Quadrat-Schätzer für b_0 und b_1 sind somit zugleich Kleinst-Quadrat-Schätzer für $E(T_{\text{r}})$ und $E(T_{\text{b}})$.

Das Vorgehen läßt sich problemlos auf beliebige MVB-Modelle übertragen. Die Maximalzahl schätzbarer Antwortzeitkomponenten $E(T_l)$ entspricht gerade der Gesamtzahl der Antwortkategorien, in einem verbundenen multinomialen Modell mit K Versuchsbedingungen also $J_1 + J_2 + \ldots + J_K$. Identifizierbar ist das Antwortzeitmodell genau dann, wenn die Prädiktorvariablen im zugehörigen Regressionsmodell über alle Antwortkategorien hinweg nicht linear abhängig sind. Testbar ist jedes sequentiell-additive Antwortzeitmodell insofern, als alle Regressionskoeffizienten größer oder gleich null sein müssen; negative durchschnittliche Antwortzeiten in einem Teilprozeß sind selbstverständlich nicht zulässig. Darüber hinaus ergeben sich testbare Restriktionen bezüglich der mittleren Antwortzeiten in den verschiedenen Antwortkategorien, wenn es sich um ein nichtsaturiertes Modell mit weniger Antwortzeitkomponenten als Antwortkategorien handelt. Im oben illustrierten Beispielfall ergeben sich für den Test des Modells beispielsweise zwei Freiheitsgrade, weil vier durchschnittliche Antwortzeiten für die verschiedenen Beobachtungskategorien durch nur zwei freie Antwortzeitkomponenten beschrieben werden.

Problematisch wird die Anwendung sequentiell-additiver Antwortzeitmodelle, wenn die Parameter des zugehörigen MVB-Modells nicht als bekannt vorausgesetzt werden können. Ein möglicher Ausweg aus diesem Problem, der mei-

nes Wissens erstmals von Xiangen Hu vorgeschlagen wurde[78], besteht in einer zweistufigen Datenanalyseprozedur. Im ersten Schritt wird ein MVB-Modell an die Kategoriehäufigkeiten unter Ignorierung der Antwortzeiten angepaßt. Mit Hilfe der Parameterschätzungen aus Schritt 1 wird dann im zweiten Schritt das zugehörige sequentiell-additive Modell an die Antwortzeiten angepaßt. Man verfährt dabei so, als wären die Parameterschätzungen aus Schritt 1 mit den Parametern des Modells identisch. Kritisch einzuwenden ist natürlich, daß dieses Vorgehen strenggenommen nur bei sehr großen Stichproben zu rechtfertigen ist. Bei kleinen oder mittleren Stichprobenumfängen sind die Parameterschätzungen mit mehr oder minder großem Stichprobenfehler behaftet, der sich in den anschließenden Regressionsanalysen der Antwortzeiten verfälschend auswirken kann. Dies liegt daran, daß die Werte der Prädiktorvariablen X_j, die ja aus den Parameterschätzungen ganz oder teilweise abgeleitet werden (vgl. z.B. Gleichung (6.8.2-3)), dann ebenfalls fehlerbelastet sind. Technisch hat man es also mit dem Problem fehlerbelasteter Prädiktorvariablen in der Regressionsanalyse zu tun. Da der Fehler in den Prädiktoren eine unbekannte Verteilung aufweist und mit den Antwortzeiten auch nicht notwendig unkorreliert ist, gibt es keine vernünftige Korrekturmöglichkeit für den verfälschenden Effekt fehlerbelasteter Parameterschätzungen.

Der sauberste Weg, dieses Problem aus dem Wege zu räumen, besteht in einer simultanen Analyse der Kategoriehäufigkeiten und der zugehörigen Antwortzeiten im Rahmen eines gemeinsamen stochastischen Modells. Eine ausführliche Behandlung dieses Zugangs, der z.B. bei Luce (1986, z.B. Kapitel 6, 7, 9) und Townsend und Ashby (1983, insbesondere Kapitel 9 und 10) an mehreren Beispielen illustriert wird, würde den Rahmen der vorliegenden Arbeit sprengen. Ich will mich deshalb auf die Skizzierung einer möglichen Konkretisierung dieser Idee beschränken.

Denkbar ist die simultane Analyse beispielsweise im Rahmen eines gemeinsamen multinomialen Modells für Antworthäufigkeiten und Antwortzeiten. Ähnlich wie im letzten Abschnitt werden hierzu die Antwortzeiten zunächst bestimmten Antwortzeitkategorien (sogenannten „bins") zugewiesen, und es wird ein multinomiales Modell für die nach verschiedenen Antwortzeitkategorien aufgeschlüsselten Antworten formuliert. Gibt es J Antworten und B „bins" für Antwortzeiten in einer bestimmten Versuchsbedingung, so ergeben sich folglich $J{\cdot}B$ Beobachtungskategorien. Die Daten, auf die sich ein solches Modell bezieht, sind Häufigkeiten bestimmter Antworten in bestimmten „bins" der Antwortzeitverteilung. Anders als im letzten Abschnitt soll dieses Modell jedoch die sequentiell-ad-

[78]Persönliche Mitteilung (September 1997). Xiangen Hu hat diesen Ansatz zur Prüfung von Quellengedächtnismodellen eingesetzt. Bei Zwei-Quellenmodellen ergeben sich neun Antwortkategorien. Anhand der durchschnittlichen Antwortzeiten in diesen Kategorien können maximal neun durchschnittliche Antwortzeitkomponenten geschätzt werden, die alle größer als null sein müssen.

ditive Natur der Antwortzeiten zum Ausdruck bringen und kann deshalb kein einfaches MVB-Modell sein. Für die Anzahl *B* und die Breite der „bins" existieren keine festen Regeln. Man sollte jedoch nicht zu wenige „bins" wählen, um die Anzahl identifizierbarer Modellparameter möglichst groß zu halten. Zugleich sollte man auf hinreichend breite „bins" achten, damit die beobachteten Häufigkeiten in den Antwortzeitkategorien nicht zu klein werden. Der ideale Kompromiß zwischen möglichst vielen und möglichst breiten „bins" dürfte wesentlich durch den Gesamtstichprobenumfang mitbestimmt werden: Je größer er ist, desto schmalere und desto mehr „bins" kann man sich leisten. Verschiedene Konkretisierung dieser Intuition diskutiert Moore (1986, S. 69f).

Um die Wahrscheinlichkeiten der „bins" im Rahmen eines sequentiell-additiven Modells darstellen zu können, benötigt man zunächst ein einfaches und gleichzeitig flexibles stochastisches Modell für die Antwortzeiten einzelner Stufen oder Stadien des gesamten Prozeßmodells. Das für diesen Zweck wohl am häufigsten benutzte Verteilungsmodell ist die Exponentialverteilung (vgl. Luce, 1986; Townsend & Ashby, 1983). Grundlegend für eine Exponentialverteilung ist die Annahme, daß die bedingte Wahrscheinlichkeitsdichte eines Prozeßabschlusses zum Zeitpunkt *t*, gegeben daß der Prozeß zuvor noch nicht abgeschlossen wurde, für alle Zeitpunkte *t* gleich ist. Kurz: Ein Prozeß mit exponentialverteilter Dauer endet in jedem Zeitintervall mit der gleichen Wahrscheinlichkeit, unabhängig von der Zeitspanne, die der Prozeß bereits andauert. Man nennt Prozesse mit dieser Eigenschaft „Poisson-Prozesse". Vorstellen kann man sich einen Poisson-Prozeß als eine mehr oder minder große Anzahl von Versuchen (N), ein gegebenes Problem — z.B. das Auffinden eines Items im Gedächtnis — zu lösen. Jeder Versuch nimmt eine kleine Zeitspanne Δt in Anspruch und führt nur mit sehr kleiner und für alle Versuche gleicher Wahrscheinlichkeit p zum Erfolg. Die Gesamtzeitdauer $T = N \cdot \Delta t$ bis zum ersten Erfolg ist dann approximativ exponentialverteilt.

Exponentialverteilungen sind durch einen Parameter λ — die „Rate des Poisson-Prozesses" — definiert. Hierdurch ist einerseits der Erwartungswert einer exponentialverteilten Antwortzeitvariablen T_j festgelegt, d.h. die durchschnittliche Zeit bis zum Abschluß des j-ten Teilprozesses: $E(T_j) = 1/\lambda_j$. Auch alle anderen Momente der Verteilung ergeben sich als einfache Funktion von λ_j (vgl. Townsend & Ashby, 1983, S. 36-38). Die Dichtefunktion lautet (z.B. Townsend & Ashby, 1983, S. 36):

$$f(T_j = t;\ \lambda_j) = \lambda_j \cdot e^{-\lambda_j \cdot t}, \text{für } t \geq 0 \qquad (6.8.2\text{-}5)$$

und die zugehörige Verteilungsfunktion ist

$$F(T_j = t;\ \lambda_j) = 1 - e^{-\lambda_j \cdot t}, \text{ für } t \geq 0. \qquad (6.8.2\text{-}6)$$

Folgen mehrere Teilprozesse aufeinander, deren Latenzen T_j unabhängig λ_j-exponentialverteilt sind, so folgt die Gesamtdauer $T = T_1 + T_2 + ... + T_L$ all dieser Prozesse einer allgemeinen Gamma-Verteilung (oder Erlang-Verteilung; Townsend & Ashby, 1983, S. 209) mit Dichtefunktion[79]

$$f(T=t;\lambda_1...\lambda_L)=\sum_{j=1}^{L}\left(\prod_{\substack{l=1\\ l\neq j}}^{L}\frac{\lambda_l}{\lambda_l-\lambda_j}\right)\cdot\lambda_j\cdot e^{-\lambda_j\cdot t},\text{für } t\geq 0,\ \lambda_l\neq\lambda_j \qquad (6.8.2\text{-}7)$$

und Verteilungsfunktion

$$F(T=t;\lambda_1...\lambda_L)=1-\sum_{j=1}^{L}\left(\prod_{\substack{l=1\\ l\neq j}}^{L}\frac{\lambda_l}{\lambda_l-\lambda_j}\right)\cdot e^{-\lambda_j\cdot t},\text{für } t\geq 0,\ \lambda_l\neq\lambda_j. \qquad (6.8.2\text{-}8)$$

Die Gamma-Verteilung ist ein Spezialfall der allgemeinen Gamma-Verteilung; sie ergibt sich, wenn man alle λ_j-Parameter der Teilprozesse als gleich annimmt.

Die Formulierung eines sequentiell-additiven multinomialen Modells für Antworthäufigkeiten und Antwortzeiten bereitet nun keine Schwierigkeit mehr. Die „Bin"-Wahrscheinlichkeiten zu einzelnen Ästen eines Ereignisbaummodells ergeben sich unmittelbar aus der Verteilungsfunktion der Exponentialverteilung (sofern $L = 1$) bzw. der generellen Gamma-Verteilung (sofern $L > 1$). Sofern mehrere Äste in die gleiche Antwortkategorie einmünden, ergeben sich die „Bin"-Wahrscheinlichkeiten als Mischungen der astspezifischen „Bin"-Wahrscheinlichkeiten.

Für das sequentiell-additive 1HT-Modell der Rekognition, dessen erwartete Antwortzeiten in den Gleichung (6.8.2-3) und (6.8.2-4) dargestellt wurde, sieht die Wahrscheinlichkeiten des i-ten bins ($i = 1, ..., I$), das durch das halboffene Zeitintervall $[t_{i-1}, t_i)$ definiert wird (mit $t_0 = 0$), für die vier Antwortkategorien „hits", „misses", „false alarms" und „correct rejections" beispielsweise wie folgt aus:

$$\begin{aligned}
p_{\text{hit},i} &= r_{1\text{ht}}\cdot\left(F(T=t_i;\lambda_\text{r})-F(T=t_{i-1};\lambda_\text{r})\right)\\
&\quad+(1-r_{1\text{ht}})\cdot b_{1\text{ht}}\cdot\left(F(T=t_i;\lambda_\text{r},\lambda_\text{b})-F(T=t_{i-1};\lambda_\text{r},\lambda_\text{b})\right)\\
p_{\text{miss},i} &= (1-r_{1\text{ht}})\cdot(1-b_{1\text{ht}})\cdot\left(F(T=t_i;\lambda_\text{r},\lambda_\text{b})-F(T=t_{i-1};\lambda_\text{r},\lambda_\text{b})\right)\\
p_{\text{f. alarm},i} &= b_{1\text{ht}}\cdot\left(F(T=t_i;\lambda_\text{r},\lambda_\text{b})-F(T=t_{i-1};\lambda_\text{r},\lambda_\text{b})\right)\\
p_{\text{corr. rej.},i} &= (1-b_{1\text{ht}})\cdot\left(F(T=t_i;\lambda_\text{r},\lambda_\text{b})-F(T=t_{i-1};\lambda_\text{r},\lambda_\text{b})\right)
\end{aligned} \qquad (6.8.2\text{-}9)$$

[79] Dies kann man durch Berechnung des Faltungsintegrals (convolution integral) der Dichtefunktionen von Exponentialverteilungen zeigen (vgl. Townsend & Ashby, 1983, S. 30 ff). Die in Gleichung (6.8.2-5) und (6.8.2-6) angegebenen Formeln setzen voraus, daß die λ_j-Parameter paarweise ungleich sind. Für die Fall gleicher λ_j-Parameter müssen andere Formeln benutzt werden (vgl. Townsend & Ashby, 1983, S. 209).

Für die Verteilungsfunktionen $F(T = t; \lambda_r)$ und $F(T = t; \lambda_r, \lambda_b)$ sind jeweils die rechten Seiten der Gleichungen (6.8.2-6) bzw. (6.8.2-8) einzusetzen. Ihre Differenzen ergeben jeweils die Wahrscheinlichkeit des bins $[t_{i-1}, t_i)$ innerhalb des betreffenden Astes des zugehörigen Ereignisbaummodells. Der Parameter λ_r ist der Intensitätsparameter des Rekognitionsprozesses, der Parameter λ_b der Intensitätsparameter des Rateprozesses. Jeweils der inverse Wert korrespondiert dem Erwartungswert der zugehörigen Exponentialverteilung.

Natürlich ist das Modell in Gleichung (6.8.2-9) gegenüber dem Regressionsmodell in Gleichung (6.8.2-3) und (6.8.2-4) insofern eine Verschärfung, als nicht nur die Erwartungswerte der Antwortzeiten, sondern deren komplette Verteilung modelliert wird. Da das Verteilungsmodell der Exponential- bzw. Erlang-Verteilung von Antwortzeiten jedoch im Rahmen des sequentiell-additiven Modellrahmens gut zu motivieren ist (Townsend & Ashby, 1983), erscheint diese Verschärfung vertretbar. Der damit verbundene Gewinn ist, daß nunmehr eine simultane ML-Analyse (oder eine andere BAN-Analyse) von Antworthäufigkeiten und Antwortzeiten im Rahmen eines gemeinsamen parametrisierten multinomialen Modells möglich wird. Die Modellgleichungen (6.8.2-9) bilden zwar kein MVB-Modell mehr, können jedoch mit den Computerprogrammen PRAXIS von Gegenfurtner (1992) oder der SOLVER-Routine aus EXCEL durch Minimierung von G^2 oder eine anderen Potenz-Divergenzstatistik an empirische Daten angepaßt werden (vgl. Abschnitt 4.7). Man erhält auf diese Weise bei Modellgültigkeit asymptotisch χ^2-verteilte Modellanpassungsstatistiken und BAN-Schätzer der Modellparameter, wobei die λ_j-Parameterschätzungen der Exponentialverteilungen ggf. in Schätzungen der mittleren Teilprozeßzeiten umgerechnet werden können.[80] Die Ergebnisse ermöglichen einen deutlich strengeren Test der sequentiell-additiven Prozeßannahmen, als er mit einer isolierten Analyse von Antworthäufigkeiten unter Ignorierung von Antwortzeiten erreicht werden kann.

6.9 Zusammenfassung

In diesem Kapitel wurden zunächst verschiedene Modellfamilien vorgestellt, die gewöhnlich losgelöst von MVB-Modellen betrachtet und mit speziell darauf zugeschnittener Software analysiert werden, obwohl sie sich bei näherer Betrachtung als MVB-Modelle herausstellen oder zumindest zu bestimmten MVB-Modellen datenäquivalent sind. Behandelt wurden eine Vielzahl von log-linearen Modellen

[80] Asymptotisch χ^2-verteilte Prüfstatistiken resultieren nur dann, wenn alle Modellparameter — auch die λ_i-Parameter der Exponentialverteilungen — aus den Häufigkeiten in den verschiedenen „bins" geschätzt werden. Prinzipiell könnten die λ_i-Parameter auch aus den ungruppierten Rohdaten nach der ML-Methode geschätzt werden. Dann sind G^2 und andere Minimum-Potenz-Divergenzstatistiken, die auf der Basis beobachteter und erwarteter „Bin"-Häufigkeiten berechnet werden, unter H_0 jedoch nicht mehr asymptotisch χ^2-verteilt (vgl. z.B. Moore, 1986, S. 65ff).

und genetischen Modellen, ferner Latent-Class-Modelle, Binomialmischverteilungsmodelle, Markoff-Ketten-Modellen und deren Varianten sowie Signalentdeckungsmodelle für kontinuierliche theoretische Größen. Die Nützlichkeit dieser Modellklassen für die kognitive Psychologie wurde in einigen Fällen durch Beispiele illustriert. Die Auswertung erfolgte grundsätzlich durch benutzerfreundliche MVB-Software, was die praktischen Vorteile verdeutlicht, die mit einer Einordnung der o.g. Modelle in die Familie der MVB-Modelle verbunden sein können.

Wie eingangs bereits betont, ging es in diesem Kapitel primär darum, die Flexibilität und Vielfalt von MVB-Modellen zu illustrieren. Eine erschöpfende Darstellung von Modellen, die als Spezialfälle von MVB-Modellen aufgefaßt werden können, war nicht intendiert. Tatsächlich ließe sich die Kette der Spezialfälle leicht erweitern. Ersetzt man etwa die Annahme lokaler Binomialverteilungen innerhalb der latenten Klassen (vgl. Abschnitt 6.3) durch die Annahme lokaler geometrischer Verteilungen, Pascal-Verteilungen oder negativer Binomialverteilungen (einführend: Hays, 1973, S. 198-200), dann resultierenden ebenfalls Spezialfälle binärer MVB-Modelle. Die Ereignisbaumkonstruktion verläuft völlig analog zu Abbildung 6.3. Wenn derartige Modelle hier unerwähnt blieben, dann deshalb, weil interessante Anwendungen in der kognitiven Psychologie m.W. noch ausstehen. Die statistische Analyse solcher Modelle würde jedoch keine oder jedenfalls keine prinzipiellen Probleme aufwerfen. Allenfalls kann die Kategorienanzahl pro Variable so groß werden, daß die MVB-Modellierung wegen der großen Zahl erforderlicher Äste recht umständlich wird. Unter der Voraussetzung, daß Programme wie GPT oder AppleTree in Verbindung mit RAM-intensiver Hardware benutzt werden, sollte die Anwendung derartiger Modelle auf Datensätze allerdings auch dann noch möglich sein.

Der letzte Abschnitt dieses Kapitels war Erweiterungen der MVB-Modellklasse auf Antwortzeiten gewidmet. Der Nutzen dieser Erweiterungen besteht insbesondere darin, strengere Tests von Theorien über serielle Informationsverarbeitungsprozesse zu ermöglichen. Dies kann im günstigsten Fall bedeuten, daß serielle Informationsverarbeitungsmodelle empirisch diskriminierbar werden, die bei Beschränkung auf die Modellierung von Antwortwahrscheinlichkeiten datenäquivalent sind. Unterschieden wurden Modelle mit zustandsabhängigen Antwortzeiten und solche mit sequentiell-additiven Antwortzeiten. Probleme einer rein deskriptiven Analyse mittlerer Antwort- bzw. Reaktionszeiten im Rahmen sequentiell-additiver Antwortzeitmodelle wurden aufgezeigt. Als potentieller Lösungsansatz wurde die Modellierung von Antwortzeitverteilungen über Mischungen von Exponentialverteilungen bzw. generalisierten Gamma-Verteilungen in Betracht gezogen. Dieser Ansatz ermöglicht u.a. eine simultane ML-Analyse von Antworthäufigkeiten und Antwortzeiten im Rahmen eines gemeinsamen

multinomialen Modells — der statistisch wohl sauberste Weg der Erweiterung von MVB-Modellen auf Antwortzeiten.

7. Überblick über bisherige Anwendungen von MVB-Modellen

Im letzten Kapitel lag der Fokus auf verschiedenen Klassen von stochastischen Modellen, die sich unter das Dach der MVB-Modelle subsumieren lassen. Substanzwissenschaftlich motivierte Modelle hatten lediglich illustrierenden Charakter. In diesem Kapitel soll ein eher substanzwissenschaftlicher Zugang zu MVB-Modellen gewählt werden. Gefragt werden soll, welche *inhaltlichen Fragestellungen* mit MVB-Modellen bereits bearbeitet wurden und in welche Bereiche der Psychologie sie fallen. Da die Anzahl entsprechender Modellanwendungen sehr groß ist, muß eine formale Detailbetrachtung der einzelnen Modelle unterbleiben. Es soll lediglich summarisch dargestellt werden, wie die Datenbasis (d.h. die empirische Kategorienstruktur) zu einem Modell aussieht, welche Interpretationen für die Modellparameter gewählt wurden und welche Befunde dazu bereits vorliegen. Für Leserinnen und Leser, die die Modelle kritisch auf ihre Brauchbarkeit hin hinterfragen wollen, wird diese summarische Darstellung sicher nicht ausreichen. Sie seien auf die jeweils zitierte Primärliteratur verwiesen.

Ein in verschiedener Hinsicht hilfreicher Ausgangspunkt ist das jüngst von Batchelder und Riefer (1999) vorgelegte Sammelreferat zu kognitionspsychologisch motivierten MVB-Modellen. Angesichts der Tatsache, daß dieses Sammelreferat erst kürzlich erschienen ist, stellt sich die Frage, ob ein erneutes Überblickskapitel in dieser Arbeit überhaupt erforderlich ist. Drei Gründe sprechen für ein Überblickskapitel auch in dieser Arbeit. Erstens ist das Sammelreferat von Batchelder und Riefer ursprünglich schon 1997 erstellt worden und deshalb nur noch partiell aktuell. Zweitens klammert es deutschsprachige Arbeiten aus. Gerade im deutschsprachigen Raum sind aber seit 1990 eine ganze Reihe von Anwendungen multinomialer Modelle publiziert worden. In dem vorliegenden Übersichtskapitel werden demgegenüber deutschsprachige Arbeiten berücksichtigt. Auch zahlreiche neuere Arbeiten werden diskutiert, die von Batchelder und Riefer noch nicht erfaßt werden konnten. Insgesamt umfaßt das vorliegende Kapitel 187 Arbeiten zur multinomialen Modellierung, wovon nur 92 von Batchelder und Riefer (1999) besprochen wurden. Um Redundanzen zu vermeiden, werden die von Batchelder und Riefer ausführlich besprochenen Arbeiten hier kürzer behandelt oder sogar ganz ausgeklammert. Die Schnittmenge wird also so klein gehalten, daß eine parallele Lektüre des vorliegenden Kapitels und des Überblicksreferats von Batchelder und Riefer lohnend sein sollte.

Der dritte und aus meiner Sicht wichtigste Grund für das vorliegende Kapitel ist der Fokus der Analyse. Batchelder und Riefer (1999) haben vorwiegend auf die Darstellung formaler Aspekte der Modelle bzw. der Modellkonstruktion Wert gelegt, weniger auf die substanzwissenschaftliche Bedeutung der Modellanwendungen. Hier soll stärker der Versuch unternommen werden, verschiedene For-

men der Anwendung von MVB-Modellen aus substanzwissenschaftlicher Perspektive zu beleuchten. Dabei werden auch solche Modelle erneut kurz aufgegriffen, die in den vorstehenden Kapiteln dieser Arbeit bereits geschildert wurden. Hier wird primär gefragt, welche *Funktion* die stochastische Modellierung im substanzwissenschaftlichen Kontext hat (z.B. Ableitung von Prozeßmaßen, Theorieprüfung, Theorievergleich, exploratorische Funktion, annahmenkritische Funktion etc.). Ferner wird gezielt nach Problemen und Schwächen bisheriger Anwendungen von MVB-Modellen gesucht, so daß erkennbar wird, wie MVB-Modellierungen in Zukunft gewinnbringender eingesetzt werden können.

Die folgenden Abschnitte sind thematisch gegliedert. Ich beginne mit gedächtnispsychologischen Anwendungen (Abschnitt 7.1) und schreite dann zu lernpsychologischen (Abschnitt 7.2), denkpsychologischen (Abschnitt 7.3), sprachpsychologischen (Abschnitt 7.4), wahrnehmungspsychologischen (Abschnitt 7.5), sozialpsychologischen (Abschnitt 7.6), entwicklungspsychologischen (Abschnitt 7.7) und psychometrischen Anwendungen fort (Abschnitt 7.8). Im abschließenden Abschnitt 7.9 wird eine Zusammenfassung der wichtigsten Anwendungsformen und der größten Anwendungsprobleme multinomialer Modellierungen in der kognitiven Psychologie versucht.

7.1 Gedächtnispsychologische Anwendungen

Psychologische Anwendungen von MVB-Modellen sind in erster Linie im Bereich der Gedächtnispsychologie zu finden. Dies dürfte vor allem drei Ursachen haben. Erstens hat die Gedächtnispsychologie einen besonderen Bedarf nach stochastischen Meßmodellen, weil sich hier das Dekomponierungsproblem sehr massiv stellt. Nicht ohne Grund wurde das Dekomponierungsproblem in Abschnitt 2.6 an einem gedächtnispsychologischen Beispiel illustriert. Es herrscht inzwischen weitgehend Einigkeit darüber, daß herkömmliche Maße des Gedächtnisses wie Rekognition, freie und geförderte Reproduktion sowie diverse indirekte Gedächtnistests (Puff, 1982; Richardson-Klavehn & Bjork, 1988) niemals „prozeßreine" (Jacoby, 1991) oder „systemreine" (Strack & Förster, 1995) Indikatoren bestimmter Gedächtnisprozesse oder -systeme sein können. Durch Vergleich verschiedener Maße zwischen experimentellen Bedingungen kann man dem Ziel einer Erfassung spezieller Gedächtnisprozesse gelegentlich näher kommen, aber hierdurch ergeben sich neue Probleme, die mit der Vergleichbarkeit und Sensitivität von Gedächtnismaßen zu tun haben. So kann beispielsweise das Ausbleiben eines Effekts in einem Gedächtnismaß A bei gleichzeitiger Signifikanz des Effekts in Gedächtnismaß B („einfache Dissoziation") einfach darauf beruhen, daß die Maße unterschiedliche Reliabilitäten aufweisen. Der Schluß von einem Dissoziationsbefund auf unter-

schiedliche Gedächtnisprozesse, die den Maßen zugrunde liegen, ist also keineswegs zwingend (Buchner & Wippich, in press).

Ein zweiter Grund für die besondere Bedeutung der MVB-Modellierung in der Gedächtnispsychologie dürfte die kategoriale Natur der meisten Daten sein, die bei Gedächtnistests gewöhnlich anfallen. Sofern quantitative Daten verwendet werden (z.B. Reproduktions- und Rekognitionsraten), sind diese — von Antwortzeiten einmal abgesehen — zumeist aus diskreten Daten abgeleitet. Wichtiger noch dürfte der dritte Grund sein: Sehr viele theoretische Konstrukte der Gedächtnispsychologie — beispielsweise Gedächtnisspeicher und -zustände — sind diskreter Natur, so daß eine auf Finite-State-Annahmen basierende multinomiale Modellierung auch konzeptuell sehr plausibel ist.

Die wichtigsten Anwendungsfelder von MVB-Modellen in der Gedächtnispsychologie sollen im folgenden kurz angesprochen werden.

7.1.1 Wiedererkennen und Raten

Der Vollständigkeit halber müssen die schon mehrfach als Beispiele herangezogenen MVB-Modelle für Rekognitionstests hier erneut erwähnt werden. Ausgehend von Arbeiten aus den fünfziger und sechziger Jahren, die sich — mehr oder minder explizit — sowohl auf Signalentdeckungs- als auch auf Rekognitionsexperimente bezogen, bildet dieses Gebiet gewissermaßen die historische Wurzel der MVB-Modellierung in der kognitiven Psychologie (z.B. Blackwell, 1953; 1963; Egan, 1958; Krantz, 1969; Luce, 1963b).

Die empirische Domäne von Rekognitionsmodellen sind Trefferraten und Raten falschen Alarms in Rekognitionstests, typischerweise über mehrere Antworttendenzbedingungen hinweg manipuliert, so daß eine empirische „memory operating characteristic curve" (MOC-Kurve) entsteht. Die Modellparameter, mit denen derartige MOC-Kurven erklärt werden sollen, zerfallen in die beiden Gruppen „Gedächtnisparameter" und „Antworttendenzparameter" (zu Hochschwellen-Modellen vgl. speziell Abschnitt 3.5.2). In erster Linie geht es also darum, möglichst reine Maße des Rekognitionsgedächtnisses abzuleiten, die von Antworttendenzeffekten (Ja- bzw. Nein-Sage-Tendenzen) dekontaminiert sind.

In manchen Anwendungen kann allerdings auch der Antworttendenzparameter von primärem Interesse sein. So unterscheiden sich beispielsweise Wörter, die Bedrohungen ausdrücken, von Kontrollwörtern in erster Linie durch eine höhere Ja-Sage-Tendenz in Rekognitionstests (Windmann & Krüger, 1998). Dies kann als Ausdruck adaptiver Urteilsstrategien bei bedrohlichen Reizen interpretiert werden (Minimierung von Verpassern). Umgekehrt weisen Depressive im Vergleich zu Kontrollpersonen eine stärkere Nein-Sage-Tendenz auf, während Unterschiede in den Gedächtnisparametern vernachlässigbar sind (Miller & Lewis,

1977). Dies weist auf konservative, zu vorsichtige bzw. zu wenig selbstbewußte Urteils- und Bewertungsstrategien bei Depressiven hin, die möglicherweise auch kausal an der Störung beteiligt sind. Derartige Ergebnisse können Konsequenzen für die Differentialdiagnose und Therapie der Depression haben.

Anwendungen von Rekognitionsmodellen fallen typischerweise in die Kategorie der *Meßmodellanwendungen* von MVB-Modellen: Nicht die Theorieprüfung oder der Theorievergleich stehen im Vordergrund, sondern die Ableitung möglichst unverfälschter Maße spezifischer kognitiver Zustände oder Prozesse.

7.1.2 Itemgedächtnis und Quellengedächtnis

Quellengedächtnistests sind einfache Verallgemeinerungen von Rekognitionstests für den Fall, daß die alten Items zwei oder mehr verschiedenen „Quellen" entstammen. Hierbei kann es sich um unterschiedliche Versuchsleiter oder auch unterschiedliche Darbietungsmodalitäten bzw. Darbietungszeitpunkte handeln. Im Regelfall interessiert dann nicht nur die Fähigkeit der Vp, alte und neue Items zu diskriminieren (Itemgedächtnis), sondern zusätzlich auch ihre Fähigkeit, als alt rekognizierte Items korrekt ihrer Quelle zuzuordnen (Quellengedächtnis).

Batchelder und Riefer (1990) haben ein BVB-Modell für Quellengedächtnistests mit zwei Quellen vorgeschlagen, das auf der Grundidee des 1HT-Modells für Ja-Nein-Rekognitionstests aufbaut (1HTSM-Modell, vgl. Abschnitt 4.5). Alte Items der Quelle i (i = 1, 2) werden demnach mit Wahrscheinlichkeit D_i als alt rekogniziert bzw. mit der komplementären Wahrscheinlichkeit $(1–D_i)$ nicht rekogniziert. Im erstgenannten Falle kann zusätzlich mit Wahrscheinlichkeit d_i auch noch die Quelle korrekt erinnert werden, was zu einem „Alt"-Urteil und zusätzlich zu einem korrekten Quellenurteil führt. Wird die Quelle nicht erinnert, so wird sie geraten, was mit Wahrscheinlichkeit a ebenfalls zu einem korrekten Quellenurteil führen kann. Bei nicht als alt rekognizierten Items wird zunächst „alt" (Wahrscheinlichkeit b) oder „neu" geraten und im erstgenannten Fall zusätzlich noch das Quellenurteil (mit Wahrscheinlichkeit g bzw. $1–g$).

Neue Items werden völlig analog zu Distraktoritems in Ja-Nein-Rekognitionstests behandelt: Die angenommenen Rateprozesse entsprechen exakt denen bei nicht rekognizierten alten Items. Insgesamt umfaßt das Modell damit in seiner allgemeinsten Form zwei Itemgedächtnisparameter (D_1 und D_2), zwei Quellengedächtnisparameter (d_1 und d_2) und drei Rateparameter (a, g, b). Durch Gleichheitsrestriktionen über die Parameter dieses Modells lassen sich Submodelle erzeugen, die identifizierbar und testbar sind (vgl. Batchelder und Riefer, 1990).

Kinchla (1994) kritisierte die Modellierung des Itemgedächtnisses im 1HTSM-Modell von Batchelder und Riefer (1990), die konzeptuell auf dem fragwürdigen Ein-Hochschwellen-Modell für Rekognitionstests aufbaut (vgl. Ab-

schnitt 4.5). Als Konsequenz aus dieser Kritik wurden verschiedene Modifikationen des 1HTSM-Modells vorgeschlagen. Darunter sind MVB-Modelle, die auf Zwei-Hochschwellen-Konzeptionen (2HTSM-Modell) oder Niedrigschwellen-Konzeptionen (LTSM-Modell) des Itemgedächtnisses basieren (Batchelder, Riefer & Hu, 1994; Bayen et al., 1996; Erdfelder et al., 1995), und auch Signalentdeckungsmodelle für Quellengedächtnistests (Slotnick, Klein, Dodson & Shimamura, 1999). Auch Erweiterungen des 1HTSM-Modells auf mehr als zwei Quellen wurden entwickelt und erfolgreich eingesetzt, insbesondere zur Klärung der Frage, ob bestimmte experimentelle Manipulationen den Quellengedächtnisparameter oder Quellenratetendenzen affizieren (Batchelder, Hu & Riefer, 1994; Riefer et al., 1994). Innerhalb des 1HTSM-Modells ist diese Frage schwer zu entscheiden, da die Sechs-Parameter-Submodelle 6b (gleiche Quellengedächtnisparameter $d_1 = d_2$, jedoch beliebige Quellenratetendenzen a bzw. g) und 6c (beliebige Quellengedächtnisparameter d_1 bzw. d_2, jedoch gleiche Quellenratetendenzen $a = g$) nahezu datenäquivalent sind und folglich fast immer gleichermaßen gut auf die Daten passen (vgl. Abschnitt 4.4). In der Erweiterung des Modells auf drei Quellen sind die Quellengedächtnisparameter demgegenüber auch ohne Restriktion der Quellenrateparameter identifizierbar. Riefer et al. (1994) fanden mit Hilfe dieses Modells heraus, daß das Quellengedächtnis für selbstgenerierte Informationen besser ist als für verschiedene Formen passiv rezipierter Informationen (Experiment 1) und daß der Gedächtnisvorteil von Bildern im Vergleich zu gelesenen oder gehörten Wörtern nicht nur das Item-, sondern auch das Quellengedächtnis betrifft (Experiment 2).

Eine interessante Verallgemeinerung des 1HTSM-Modells stammt von Dodson, Holland und Shimamura (1998). Wenn mehr als zwei Quellen zu diskriminieren sind, kann u.U. partielles Quellengedächtnis verfügbar sein (z.B. über das Geschlecht einer Person), nicht aber spezifisches Quellengedächtnis (z.B. über die konkrete Person, die eine Aussage getätigt hat). Sie entwickelten deshalb ein 1HTSM-Modell für multiple Quellen, welches sich von dem Modell von Batchelder, Hu und Riefer (1994; Riefer et al., 1994) dadurch unterscheidet, daß im Fall einer „Alt"-Rekognition *ohne* spezifisches Quellengedächtnis (Wahrscheinlichkeit $D \cdot (1-d)$) nicht automatisch geraten wird, sondern mit Wahrscheinlichkeit P partielle Quelleninformation (z.B. das Geschlecht der Person) verfügbar sein kann, so daß nur noch eine von den Quellen geraten wird, auf die die partielle Quelleninformation zutrifft. Die Parameter d und P wurden von Dodson et al. (1998) durch experimentelle Manipulationen der Verfügbarkeit partieller Quelleninformationen (Sprecher unterschiedlichen Geschlechts vs. gleichen Geschlechts) und der Aufmerksamkeit in der Abrufphase (divided attention versus undivided attention) validiert. Das Standard-1HTSM-Modell (mit $P = 0$) paßte auf die Bedingung mit Sprechern gleichen Geschlechts, nicht jedoch auf die Bedingung mit Sprechern ungleichen Geschlechts, für die $P > 0$ angenommen werden mußte (Experiment 1).

In Experiment 2 affizierte die Aufmerksamkeitsmanipulation spezifisches Quellengedächtnis *d* (für die Stimme einer Person), nicht jedoch partielles Quellengedächtnis *P* (für das Geschlecht des Sprechers) und das Itemgedächtnis. Alt-Neu-Entscheidungen und Entscheidungen über partielle Quelleninformationen verlangen also offenbar wenig oder keine Aufmerksamkeit, sie können relativ automatisch getroffen werden.

Unabhängig von Dodson et al. (1998) haben Klauer und Wegener (1998) ein sehr ähnliches Quellendiskriminationsmodell für mehr als zwei Quellen entwickelt und ausgiebig validiert (vgl. Abschnitt 4.5). Ihr Modell basiert auf einer Zwei-Hochschwellen-Annahme, läßt also anders als das Modell von Dodson et al. (1998) die Möglichkeit der Distraktor-Entdeckung zu. Außerdem beruht es auf einer etwas anderen Parametrisierung der Ratewahrscheinlichkeiten bei fehlendem Quellengedächtnis. Da das Klauer-Wegener-Modell bislang nur auf Fragestellungen zur sozialen Kategorisierung angewandt wurde, soll es an dieser Stelle nur kurz erwähnt werden. Ausführlicher komme ich in Abschnitt 7.6.1 darauf zurück.

Eine Verallgemeinerung des 2HTSM-Modells auf multiple Quellen, die in mehreren Dimensionen variieren, wurde von Meiser und Bröder (1999) vorgelegt. Im einfachsten Fall verlangt ihr Modell Items aus vier Quellen, die auf zwei Dimensionen in jeweils zwei Stufen variieren. Beispielsweise können Items auf einem Computerbildschirm an zwei verschiedenen Orten (z.B. linke versus rechte Hälfte des Monitors) und in jeweils zwei verschiedenen Größen (z.B. 1.5 versus 0.6 cm) präsentiert werden. Anschließend müssen Probanden alte und neue Items in einem Vier-Quellen-Diskriminationstest bearbeiten. Für Items, die als „alt" bezeichnet werden, muß angegeben werden, ob sie (a) links oder rechts und (b) in kleinem oder großem Font dargeboten wurden. Das Modell von Meiser und Bröder umfaßt neben den üblichen Itemgedächtnis- und Alt-Neu-Rateparametern zwei Klassen von Quellengedächtnis- und Quellenrateparametern. Im geschilderten Beispiel wären dies Maße für das Quellengedächtnis und den Ratebias bezüglich (a) der Lokalisationsdimension und (b) der Fontgrößendimension im Itemmaterial. In zwei Validierungsexperimenten fanden Meiser und Bröder gute bzw. sehr gute Ergebnisse des G^2-Modellanpassungstests. Sie zeigten, daß die Parameter ihres mehrdimensionalen Modells den korrespondierenden Parametern des eindimensionalen 2HTSM-Modells sehr gut entsprechen, wenn für letzteres jeweils nur eine Itemdimension in die Analyse einbezogen wird. Außerdem zeigten sie, daß sich Manipulationen der Quellenähnlichkeit, die auf einzelne Dimensionen beschränkt sind, selektiv auf die diesen Dimensionen zugeordneten Quellengedächtnisparameter auswirken. Geringere räumliche Distanz in der Lokalisationsdimension reduziert beispielsweise nur das Quellengedächtnis für den Darbietungsort, nicht aber das Quellengedächtnis für die Itemgröße. Eine erste Anwendung des Quellengedächtnismodells für gekreuzte Kategorien auf das Phänomen

illusorischer Korrelationen in der Stereotypenforschung haben Meiser und Hewstone (1999) präsentiert (vgl. dazu Abschnitt 7.6.2). Die Verallgemeinerung dieses Modells auf multiple Quellen, die mehr als zwei Dimensionen und/oder mehr als zwei Ausprägungen pro Dimension aufspannen, liegt auf der Hand.

Sind nur zwei Quellen zu diskriminieren, so sind viele Submodelle der 1HTSM-, 2HTSM- und LTSM-Modellfamilien datenäquivalent oder „annähernd datenäquivalent" (Erdfelder et al., 1995). Um so wichtiger sind systematische Validierungsstudien zu den verschiedenen Modellen. Bislang favorisieren sie recht eindeutig das 2HTSM-Modell für Quellendiskriminationstests (Bayen et al., 1996). Allerdings betreffen die vorliegenden Validierungsbefunde nur die Parameter für das Itemgedächtnis und das Quellengedächtnis. Inwieweit sich Manipulationen der Antworttendenzen ausschließlich in den Rateparametern und nicht in den Gedächtnisparametern niederschlagen, wurde noch nicht überprüft. Im nächsten Kapitel werde ich auf diesen Punkt zurückkommen und anhand einer Reanalyse publizierter Daten zeigen, daß das 2HTSM-Modell auch Antworttendenzen besser abbildet als das 1HTSM- und das LTSM-Modell.

Bisherige Anwendungen der geschilderten Modelle konzentrieren sich im wesentlichen auf das 1HTSM-Modell. Batchelder und Riefer (1990) haben u.a. Daten zur Realitätsüberwachung (reality monitoring) von Johnson, Foley und Leach (1988) reanalysiert. Sowohl das Item- als auch das Quellengedächtnis war für vorgestellte Informationen („imagine") besser als für gehörte Informationen („listen"), und zwar um so mehr, je mehr sich die Stimme, in der man sich die Information vorzustellen hatte, von der Stimme der gehörten Informationen unterschied. Quellenähnlichkeit reduziert also das Quellengedächtnis auch dann, wenn es sich bloß um vorgestellte Attribute einer Quelle handelt.

Eine Reanalyse klinischer Daten zur Realitätsüberwachung (Harvey, 1985) durch Batchelder und Riefer (1990) zeigte, daß Schizophrene mit Wahnstörungen im Vergleich zu anderen psychiatrischen Gruppen einen extrem schlechten Quellengedächtnisparameter (d) für ausgesprochene („say") und bloß gelesene Wörter („think") aufweisen. Dieser Befund wurde von Keefe, McCourtney, Bayen und Harvey (1999) mit dem 2HTSM-Modell repliziert: 28 Patienten mit schizophrenem Wahnerleben wiesen ein schlechteres Quellengedächtnis als 19 Kontrollpersonen auf, die hinsichtlich Alter und Ausbildung parallelisiert worden waren. Das Quellengedächtnisdefizit Schizophrener war auf allen Ebenen beobachtbar: bei der Diskriminierung (a) externer Quellen (männliche versus weibliche Sprecher), (b) interner Quellen (vorgestellte eigene versus vorgestellte fremde Aussagen) sowie (c) interner versus externer Quellen (gehörte oder ausgesprochene versus bloß vorgestellte Items). Im letztgenannten Fall bestand bei Schizophrenen darüber hinaus ein stärkerer Ratebias zugunsten externer Quellen, sofern die Quelleninformation

vergessen worden war. Dies *kann* ein kausaler Faktor im schizophrenen Wahnerleben sein, der auf anderem Wege nur schwer diagnostizierbar sein dürfte.

Batchelder und Riefer (1990) haben darüber hinaus Daten zum Quellengedächtnis für die Sprache einer Originalinformation bei mehrsprachigen Versuchspersonen mit dem 1HTSM-Modell reanalysiert. Saegert, Hamayan und Ahmar (1975) boten ihren mehrsprachigen Vpn englische und französische Wörter dar, einmal isoliert und einmal im Satzkontext. Das Gedächtnis für die Sprache der Originalinformation (Quellengedächtnis) war für isolierte Wörter und Wörter im Satzkontext gleichermaßen gut ($\hat{d} \approx .90$), während das Itemgedächtnis für isolierte Wörter klar besser war. Ein ganz ähnliches Ergebnis resultierte bei der Reanalyse der Daten von Rose, Rose, King und Perez (1975), die ihren mehrsprachigen Vpn englische und spanische Sätze dargeboten hatten, welche entweder unzusammenhängend oder thematisch zusammenhängend aufgebaut waren. Das Quellengedächtnis für die Sprache der Originalinformation war bei unzusammenhängenden und zusammenhängenden Sätzen gleichermaßen gut, während das Itemgedächtnis für unzusammenhängende Sätzen erneut besser war. Beide Befunde passen gut zu einem Multiple-Code-Modell des mehrsprachigen Gedächtnisses, das für jede Sprache einen eigenständigen Gedächtniscode vorsieht. Single-Code-Konzeptionen haben dagegen mit den Befunden Probleme, da das Quellengedächtnis für die Sprache der Originalinformation auch dann sehr gut bleibt, wenn die Satzinhalte eine Enkodierung der Sprache, in der die Information dargeboten wurde, eher unwahrscheinlich erscheinen lassen.

Johnson, Kounios und Reeder (1994) ließen ihre Vpn unter mehr oder minder starkem Zeitdruck Quellengedächtnistests bearbeiten (300 bis 1500 Millisekunden nach Darbietung des Testitems). Sie fanden, daß der *D*-Parameter des 1HTSM-Modells schon wenige Millisekunden nach Darbietung des Testitems signifikant größer als null ist, während der *d*-Parameter erst später ansteigt. Das Itemgedächtnis kann also offenbar schneller beurteilt werden als das Quellengedächtnis, was für einen seriellen Prozeß spricht (erst wird das Itemgedächtnis beurteilt, danach das Quellengedächtnis).

Inzwischen liegen auch zahlreiche Befunde zur Dissoziierbarkeit der Itemgedächtnis- und Quellengedächtnisparameter des 1HTSM-Modells vor. Mulligan (1996) zeigte beispielsweise, daß die Präsenz interferierender Stimuli während der Enkodierung das Itemgedächtnis erhöhen kann (*D*-Parameter des 1HTSM-Modells), jedoch keinen Einfluß auf das Quellengedächtnis zu haben scheint. Hu, Marks und Isenberg (1996) fanden umgekehrt, daß die Mitteilung, eine soeben gelernte Itemliste sei irrelevant für den folgenden Gedächtnistest (directed forgetting), das Quellengedächtnis für diese Itemliste im Vergleich zu einer Kontrollgruppe drastisch reduziert, nicht jedoch das Itemgedächtnis.

Begg, Anas und Farinacci (1992) fanden, daß das Quellengedächtnis für glaubwürdige versus unglaubwürdige Quellen besser ist, wenn die Vpn *vor* der Darbietung der Originalinformation über die Glaubwürdigkeit der Quellen informiert werden statt danach. Die gleiche Manipulation hat praktisch keinen Einfluß auf das Itemgedächtnis. Begg et al. (1992) verwendeten das 1HTSM-Modell allerdings nur als Ergänzung zu traditionellen Analysen der Quellenverwechslungsraten. Da in ihren Datensätzen das Itemgedächtnis fast perfekt war ($.80 < \hat{D} < 1$), ließ sich der Schätzer für d gut durch eine lineare Transformation der Quellenverwechslungsraten approximieren (vgl. Batchelder & Riefer, 1990, S. 554, Gleichung 22; Begg et al., 1992, S. 451, Fußnote 3). Auf eine modellbasierte Datenanalyse kann unter diesen Randbedingungen verzichtet werden.

Henkel und Franklin (1998) verwendeten eine Realitätsüberwachungsaufgabe (reality monitoring task), bei der zuvor wirklich wahrgenommene Items (Quelle A), zuvor bloß imaginierte Items (Quelle B) und neue Items zu diskriminieren waren. Sie beobachteten eine doppelte Dissoziation von Item- und Quellengedächtnis: Imaginierte Items wiesen im Vergleich zu wahrgenommenen Items eine besseres Itemgedächtnis, jedoch ein schlechteres Quellengedächtnis auf, und zwar um so mehr, je physikalisch ähnlicher die imaginierten Items im Vergleich zu wahrgenommenen Items waren. Allerdings basieren ihre Ergebnisse auf dem Modell 6c von Batchelder und Riefer (1990), welches dem Modell 6b (freie Quellenrateparameter, jedoch gleichgesetzte Quellengedächtnisparameter) datenäquivalent ist. Folglich können die Quellengedächtniseffekte alternativ auch als Quellenrateeffekte interpretiert werden (Murnane & Bayen, 1998). Ein Erweiterung der Analyse auf ein Drei-Quellen-Modell (Riefer et al., 1994) hätte diesbezüglich mehr Klarheit gebracht.

Häufig wurde das 1HTSM-Modell zur Untersuchung der Gedächtnisentwicklung eingesetzt, typischerweise mit dem Ergebnis, daß ältere Personen im Vergleich zu jüngeren (Ferguson, Hashtroudi & Johnson, 1992; Henkel, Johnson & DeLeonardis, 1998; Jacoby, 1999; Light, LaVoie, Valencia-Laver, Albertson Owens & Mead, 1992) und Kinder im Vergleich zu Erwachsenen (Lindsay, Johnson & Kwon, 1991) insbesondere hinsichtlich des Quellengedächtnisses benachteiligt sind. Tendenziell scheinen die Altersunterschiede im Quellengedächtnis mit der perzeptuellen Ähnlichkeit der Quellen zuzunehmen. Sind wirklich wahrgenommene und bloß imaginierte Items in einer Realitätsüberwachungsaufgabe zu diskriminieren, so sind die Altersunterschiede im Quellengedächtnis für imaginierte Items sehr groß, jedoch für die wirklich wahrgenommenen Items vernachlässigbar (Henkel et al., 1998).

Henkel et al. (1998) haben die für einzelne Vpn geschätzten Parameter des 1HTSM-Modells zu den Ergebnissen neuropsychologischer Testbatterien und Intelligenztests in Beziehung gesetzt. Sie fanden bedeutsame Korrelationen zwischen

den Schätzwerten des Quellengedächtnisparameters und psychologischen Tests, welche frontalhirn- und temporalhirnspezifische Leistungen erfassen. Allerdings sind diese Ergebnisse mit Vorsicht zu genießen, da Henkel et al. (1998) die individuellen Parameterschätzungen auf der Basis des Modells 4 von Batchelder und Riefer (1990) ermittelt haben, welches gleiche Itemgedächtnisparameter und gleiche Quellengedächtnisparameter für wahrgenommene und imaginierte Items voraussetzt. Diese Modellwahl ist problematisch, da die Autoren die Inadäquatheit der beiden Annahmen in der gleichen Arbeit eindrucksvoll nachgewiesen haben. Vermutlich steht hinter der Modellwahl einfach der Wunsch, globale Maße des Item- und des Quellengedächtnisses in die Korrelationsanalysen eingehen zu lassen. Genau dies liefert die Analyse jedoch nicht: Wenn in Wahrheit für jede Vp das Modell 6c mit getrennten Itemparametern D_1 und D_2 und getrennten Quellengedächtnisparametern d_1 und d_2 gilt, die Daten aber dennoch mit Modell 4 analysiert werden, so ist der Schätzer des Quellengedächtnisparameters d in Modell 4 keineswegs nur von d_1 und d_2 abhängig. Wie Tabelle 7.1 anhand einiger Parameterkonstellationen exemplarisch zeigt, kann der Schätzer für d sowohl durch die Differenz der Itemgedächtnisparameter D_1 und D_2 in Modell 6c als auch durch den Quellenrateparameter g massiv beeinflußt werden.

Tabelle 7.1:
ML-Parameterschätzungen und G^2-Anpassungsstatistiken zum 1HTSM-Quellengedächtnismodell 4 von Batchelder und Riefer (1990) für residuenfreie Datensätze zum 1HTSM-Quellengedächtnismodell 6c mit fünf unterschiedlichen Parametervektoren. Die Stichprobenumfänge (jeweils N = 1680) und die Parameterwerte des ersten Datensatzes sind durch die Tabellen 1 und 2 bei Henkel et al. (1998, S. 257) inspiriert. Obwohl die beiden Quellengedächtnisparameter d_1 und d_2 in allen fünf Datensätzen identische Werte annehmen, variiert der d-Parameterschätzer des angepaßten Modells 4 in Abhängigkeit vom Quellenratebias $g = a$ und von der Differenz der Itemgedächtnisparameter D_1 und D_2.

Parameter des gültigen Modells 6c						Schätzer des angepaßten Modells 4				
D_1	D_2	d_1	d_2	b	$a = g$	$\hat{D}$	$\hat{d}$	$\hat{b}$	$\hat{a} = \hat{g}$	$G^2_{(2)}$
.400	.710	.520	.070	.230	.560	.555	.217	.230	.613	48.29
.400	.710	.520	.070	.230	.100	.555	.426	.230	.120	45.16
.400	.710	.520	.070	.230	.900	.555	.096	.230	.915	44.57
.100	.900	.520	.070	.230	.560	.500	.122	.230	.574	299.0
.900	.100	.520	.070	.230	.560	.500	.353	.230	.636	311.0

Als Konsequenz aus den in Tabelle 7.1 dargestellten Zusammenhängen bleibt nur die Feststellung, daß Korrelationen des d-Schätzers aus dem statistisch inadäquaten Modell 4 nicht sinnvoll interpretierbar sind. Derartige Korrelationen können zwar Zusammenhänge zwischen dem Quellengedächtnis und neuropsychologischen Indikatoren indizieren; ebensogut können sie jedoch auch Zusammenhänge zwischen Quellenratetendenzen oder Itemgedächtnisunterschieden

D_1–D_2 mit neuropsychologischen Indikatoren reflektieren. Wenn beispielsweise alle älteren Vpn nahezu identische Quellengedächtnisparameter d_1 und d_2 und Itemgedächtnisparameter D_1 und D_2, jedoch unterschiedliche Quellenratewahrscheinlichkeiten g aufweisen, würde eine bedeutsame Korrelation zwischen dem d-Schätzer aus Modell 4 und neuropsychologischen Tests in Wahrheit eine Korrelation zwischen g und diesen Tests zum Ausdruck bringen. Der Schluß, daß das Quellengedächtnis mit Frontal- oder Temporalhirnfunktionen zusammenhängt, ist aufgrund dieser Befunde also nicht berechtigt. Einmal mehr wird deutlich, daß Parameterschätzungen substanzwissenschaftlich oftmals[81] wertlos sind, wenn begründete Zweifel an der Gültigkeit des den Schätzern zugrundeliegenden Modells bestehen.

Bayen und Murnane (1996) haben die Hypothese von Ferguson et al. (1992) überprüft, wonach das Altersdefizit im Quellengedächtnis im wesentlichen auf der mangelhaften simultanen Nutzung unterschiedlicher Merkmale der verschiedenen Quellen beruht. Anders als die zuvor genannten Autoren verwendeten sie das 2HTSM-Modell anstelle des 1HTSM-Modells. Im Gegensatz zur Hypothese von Ferguson et al. (1992) zeigte sich in ihrer Untersuchung, daß ältere Menschen stärker von der Erhöhung der Distinktivität perzeptueller und temporaler Merkmale der Quellen profitieren als jüngere. Dieses Ergebnis stützt eher die „Environmental-Support-Hypothese" von Craik (1994), wonach ältere Menschen im Vergleich zu jüngeren reduzierte Verarbeitungsressourcen aufweisen, was jedoch durch günstig konstruierte „Lernumwelten" partiell kompensiert werden kann.

Das 2HTSM-Modell wurde ferner zur Untersuchung der Wechselwirkung zwischen Emotion und Gedächtnis eingesetzt (Johnson, Nolde & De Leonardis, 1996). Die Instruktion, während eines Dialogs zwischen zwei beobachteten Personen darauf zu achten, wie man sich selbst fühlt, erhöht beim späteren unerwarteten Quellengedächtnistest das Itemgedächtnis, reduziert jedoch das Quellengedächtnis. Umgekehrt bewirkt die Instruktion, während des Dialogs darauf zu achten, wie sich jede der beiden interagierenden Personen fühlt (oder wie man selbst über die Gefühle dieser Personen denkt) eine Reduktion des Itemgedächtnisses bei gleichzeitiger Verbesserung des Quellengedächtnisses. Die Befunde von Johnson et al. (1996) zeigen somit, daß die Verteilung der Verarbeitungsressourcen beim inzidentellen Lernen durch den emotionalen Fokus beeinflußt werden kann. Weitere Anwendungen des 2HTSM-Modells betreffen das Gedächtnis für schema- bzw.

[81] Die Einschränkung „oftmals" ist notwendig, weil die Wahl eines falschen Analysemodells in einigen Fällen auch unschädlich sein kann. Im Beispielfall ist der b-Parameter aus Modell 4 interpretierbar, auch wenn in Wirklichkeit Modell 6c gelten sollte. Ob sich die Zugrundelegung eines falschen Modells auf die Interpretierbarkeit eines Parameters schädlich auswirkt, kann erst nach genauer Inspektion der Modellgleichungen entschieden werden. Im allgemeinen muß immer damit gerechnet werden, daß die Zugrundelegung eines inadäquaten Modells Parameterschätzungen uninterpretierbar macht.

skriptbasierte Informationen (Bayen, Nakamura, Dupuis & Yang, in press; Erdfelder & Bredenkamp, 1998) sowie illusorische Korrelation (Klauer & Meiser, in press), worauf jedoch an dieser Stelle noch nicht eingegangen werden soll (vgl. dazu die Abschnitte 7.1.6 und 7.6.2).

Einen anderen modellbasierten Zugang zum Problem des Quellengedächtnisses haben Kliegl und Lindenberger (1993) gewählt. Ihr Modell bezieht sich auf eine Sequenz von vier Paar-Assoziations-Lerndurchgängen, bei denen immer die gleichen Stimuli und Responsewörter verwendet werden, wobei jedoch die Zuordnung von Responsewörtern zu Stimuli von Liste zu Liste wechselt. In den vier Lernphasen werden die zusammengehörigen Stimulus-Response-Paare präsentiert; in den anschließenden Testphasen werden die Stimulusitems vorgegeben und die Vpn müssen die für den betreffenden Durchgang richtigen Antworten nennen (Lern-Prüf-Methode des PAL). Ab dem zweiten Durchgang können neben korrekten Reproduktionen zwei Arten von Fehlern unterschieden werden: Intrusionen (falsche Antworten, die auf einer der vorherigen Listen richtig waren) und andere Fehler. Kliegl und Lindenberger unterschieden für jedes Responsewort drei Zustände: „gespeichert mit Listenmarkierung" (Wahrscheinlichkeit a), „gespeichert ohne Listenmarkierung" (Wahrscheinlichkeit b) und „nicht gespeichert" (Wahrscheinlichkeit $1–a–b$). Das entspricht einem Quellengedächtnismodell, bei dem die Listenmarkierung die Rolle der Quellenerinnerung übernimmt. Für die Übergänge zwischen den drei Zuständen über die Durchgänge hinweg nahmen die Autoren ein stationäres Markoff-Ketten-Modell an: Die Wahrscheinlichkeit, daß eine Listenmarkierung für Durchgang i (i = 1, ..., 3) beim nächsten Durchgang (i + 1) noch erhalten ist, beträgt p; und die Wahrscheinlichkeit, daß das Item noch gespeichert ist, falls keine Listenmarkierung vorhanden ist, beträgt q. Bei jeder neuen Liste werden die Antworten mit Wahrscheinlichkeit a inklusive Markierung für die aktuelle Liste und mit Wahrscheinlichkeit b ohne Listenmarkierung gespeichert. Es wird ferner angenommen, daß die Vpn eine optimale Antwortstrategie verfolgen:

(1) Ist zu einem Stimulus eine Antwort mit Markierung für die aktuelle Liste vorhanden, so wird diese Antwort genannt.

(2) Fehlt eine Antwort mit passender Listenmarkierung, so wird eines der zum Stimulus gespeicherten Responsewörter ohne Listenmarkierung zufällig ausgewählt.

(3) Gibt es keine zu dem Stimuluswort gespeicherten Responsewörter, so wird ein beliebiges Responsewort ausgewählt, das keine Markierung aufweist. Die Wahrscheinlichkeit einer korrekten Antwort ist in diesem Fall allerdings vernachlässigbar, sofern eine hinreichend lange Itemliste verwendet wird.

Aus diesen Annahmen lassen sich die Wahrscheinlichkeiten von korrekten Reproduktionen, von Intrusionen aus jeder der vorhergehenden Listen und von an-

deren Fehlern ableiten. Das umfangreiche binäre MVB-Modell ist in Tabelle 6 bei Kliegl und Lindenberger (1993, S. 622) zusammengefaßt.

Die Autoren haben das Modell auf Stichproben von älteren und jüngeren Erwachsenen angewandt. Besonderes Gewicht wurde auf die adäquate Instruktion der Probanden gelegt, wodurch die Gültigkeit der o.g. Modellannahmen zur Antwortstrategie wahrscheinlich gemacht werden sollte. Darüber hinaus wurden unterschiedliche Item-Response-Darbietungszeiten für ältere und jüngere Probanden gewählt, so daß sich der Prozentsatz korrekter Reproduktionen zwischen den Altersgruppen nicht unterschied. Die Modellanpassung wurde nach der ML-Methode vorgenommen. Leider resultierte für die Gesamtstichprobe eine schlechte Modellanpassung, die jedoch im wesentlichen auf einige wenige Vpn der jüngeren Stichprobe zurückzuführen war (Experiment 1).[82] Nach Ausschluß der Daten dieser Vpn ergab sich eine befriedigende Datenanpassung. Da sich die Parameterschätzungen durch den Ausschluß dieser Vpn nur unwesentlich veränderten, wurden die statistischen Analysen nur für den Gesamtdatensatz durchgeführt.[83] Die Ergebnisse lassen vermuten, daß sich die Vergessensfunktionen (*p*- und *q*-Parameter) für ältere und jüngere Probanden nicht unterscheiden, wohl aber die *a*- und *b*-Parameter des Modells: Ältere hatten deutlich geringere *a*-Parameterschätzungen (gespeicherte Responsewörter mit Listenmarkierung) und höhere *b*-Parameterschätzungen (gespeicherte Responsewörter ohne Listenmarkierung) als jüngere Probanden. Als Folge der längeren Item-Response-Darbietungszeiten für ältere Probanden war die Summe der *a*- und *b*-Parameterschätzungen für beide Altersgruppen annähernd gleich. Interessanterweise führte diese künstliche Egalisierung des Itemgedächtnisses aber nicht zur Angleichung der Quellengedächtnisraten: Die Wahrscheinlichkeit einer adäquaten Listenmarkierung, gegeben daß eine gespeicherte Antwort vorliegt, wurde bei jüngeren Probanden deutlich höher geschätzt. Dies steht im Einklang mit den Ergebnissen der bereits geschilderten gerontologischen Untersuchungen, die das 1HTSM- oder das 2HTSM-Modell verwendet haben.

Systematische Vergleiche zwischen Ergebnissen, die mit verschiedenen Quellengedächtnismodellen für gegebene Daten erzielt werden können, sind bislang rar. Batchelder, Riefer und Hu (1994) fanden beim Vergleich des 1HTSM- und

---

[82]Kliegl und Lindenberger (1993, S. 630f) haben auch eine plausible Modellerweiterung vorgestellt, die zu einer deutlichen Verbesserung der Anpassungsgüte führt.

[83]Obwohl dieses Vorgehen für die Parameterschätzungen selbst durchaus vertretbar ist, ist es für die berichteten G^2-Differenzentests zur Prüfung von Parameternullhypothesen problematisch. Diese setzen nämlich voraus, daß das Ausgangsmodell, für das bestimmte Parameterrestriktionen geprüft werden sollen, zutrifft (vgl. Abschnitt 4.3.2.2). Ist diese Annahme nicht erfüllt — wie im Falle des Modells von Kliegl und Lindenberger (1993) — so ist über die Verteilung der G^2-Differenzenstatistik nichts bekannt. Im Falle der Daten von Kliegl und Lindenberger (1993) sind die deskriptiv-statistischen Ergebnisse allerdings so eindeutig, daß deren Interpretation durch ein veränderte inferenzstatistische Strategie wohl kaum wesentlich tangiert werden würde.

des LTSM-Modells für die von Batchelder und Riefer (1990) berichteten Daten ebenso wenig bemerkenswerte Unterschiede wie Erdfelder und Bredenkamp (1998), die ihr Experiment 1 sowohl mit dem 2HTSM-Modell als auch mit dem 1HTSM-Modell ausgewertet haben. Slotnick et al. (1999) verglichen das 1HTSM-Modell mit mehrdimensionalen signalentdeckungstheoretischen Modellen für Quellengedächtnistests und stellten ebenfalls hohe Ähnlichkeiten der Ergebnisse fest, sieht man einmal von Fällen ab, in denen die Item- und Quellengedächtnisparameter des 1HTSM-Modells extreme Werte annehmen. Auf der anderen Seite zeigen die Ergebnisse von Bayen et al. (1996), daß Unterschiede zwischen verschiedenen modellbasierten Analysen durchaus bestehen können. Vor allem für die Arbeiten, die das in Validierungsstudien bislang erfolgreichste 2HTSM-Modell nicht verwendet haben, scheint deshalb eine Reanalyse der Daten angebracht zu sein. Denkbar ist, daß sich einige Befunde — speziell zu den Itemgedächtnisparametern — im Rahmen des 2HTSM-Modells anders darstellen.

7.1.3 Explizites und implizites Gedächtnis

Seit etwa Mitte der siebziger Jahre beschränkt sich die Analyse von Rekognitionstests nicht mehr allein auf die „Dekontaminierung" der beobachteten Trefferraten von Antworttendenzeffekten bzw. Rateeinflüssen, um auf diese Weise möglichst reine Wiedererkennensmaße zu gewinnen (vgl. Abschnitt 7.1.1). Vielmehr wurde der Prozeß des Wiedererkennens selbst zunehmend zum Gegenstand der Modellbildung. Die MVB-Modellierung erwies sich insbesondere im Bereich der sogenannten Zwei-Prozeß-Modelle der Rekognition als sehr nützlich, die einen expliziten Erinnerungsprozeß („bewußte Erinnerung") und einen eher impliziten Prozeß („Vertrautheit") unterscheiden.

Mandler (1980) hat ein einflußreiches Zwei-Prozeß-Modell vorgestellt, das als einfaches binäres MVB-Modell formuliert werden kann: Ein altes Item wird mit Wahrscheinlichkeit F als alt beurteilt, weil es ein starkes Vertrautheitsgefühl auslöst; wenn es kein hinreichend starkes Vertrautheitsgefühl auslöst (mit Wahrscheinlichkeit $1–F$) kann es u.U. aufgrund von Kontextinformation dennoch bewußt erinnert werden (mit Wahrscheinlichkeit R), was ebenfalls zu einem „Alt"-Urteil führt. Die Rekognitionswahrscheinlichkeit folgt folglich dem einfachen binären Modell $p(\mathrm{Rg}) = F + (1–F) \cdot R$. Verallgemeinerungen dieses Modells auf die Rekognition von Wortpaaren sind leicht möglich. Derartige Verallgemeinerungen ermöglichen die Schätzung des Parameters F anhand von Paar-Assoziations-Lernexperimenten: A-B-Itempaare werden gelernt, und anschließend ist entweder ein A-Item allein, ein B-Item allein oder ein A-B-Itempaar neben korrespondierenden Distraktoritems daraufhin zu beurteilen, ob das betreffende Item bzw. Itempaar alt oder neu ist. Damit das o.g. Modell identifizierbar wird, werden die R-Parameter

des Modells von Mandler nicht statistisch geschätzt, sondern durch die beobachteten Reproduktionsraten in geförderten Reproduktionstests (mit dem A- bzw. dem B-Item als Hinweisreiz) approximiert. Nach Mandler (1980) fließen in Rekognitionstests also immer auch Reproduktionsvorgänge ein, die prinzipiell beobachtbar sind bzw. beobachtbar gemacht werden können. Trotz der strengen Annahmen, auf denen das Modell für PAL-Experimente beruht, paßt es deskriptiv recht gut auf die von Mandler (1980) berichteten Daten. Modellanpassungsstatistiken wurden allerdings nicht berichtet.

Humphreys und Bain (1983) haben das Zwei-Prozeß-Modell von Mandler (1980) mit einem eigenen Zwei-Komponenten-Modell der Rekognition verglichen. Auch ihr Modell ist als binäres MVB-Modell formulierbar. Es nimmt zunächst an, daß ein Zielwort unabhängig von seinem Kontext enkodiert und abgerufen werden kann. Zusätzlich kann jedoch auch noch die Beziehung zwischen Item und Kontext abgespeichert werden, was einen zweiten Weg zu einer erfolgreichen Rekognition eröffnet, sofern (a) der itemspezifische Prozeß scheitert und (b) während des Abrufs Kontextinformationen der Enkodierungsphase präsent sind. Wie Humphreys und Bain (1983) zeigten, sind die Modelle von Mandler und Humphreys/Bain für die von Mandler (1980) untersuchten Paradigmen empirisch nur schwer gegeneinander kontrastierbar.

Jacoby (1991) schlug ein sehr einfallsreiches Paradigma vor, um die Parameter des Mandler-Modells identifizierbar zu machen, ohne auf die problematische Gleichsetzung des *R*-Parameters mit geförderten Reproduktionsraten zurückgreifen zu müssen. In der Rekognitionsvariante seiner Prozeßdissoziationsprozedur (PDP) bekommen Probanden in der Testphase drei Gruppen von Items vorgelegt: A-, B-, und N-Items. Die A- und B-Items sind zwei Gruppen alter Items, die N-Items sind neue Items (Distraktoren). Die Items werden unter zwei Bedingungen beurteilt, die entweder intra- oder interindividuell variiert werden: In der Inklusionsbedingung soll sowohl auf A- als auch auf B-Items mit „ja" geantwortet werden; nur auf Distraktoren soll mit „nein" reagiert werden. In der Exklusionsbedingung soll dagegen nur auf B-Items mit „ja" reagiert werden; A- und N-Items sind mit „nein" zu beantworten. Beispielsweise kann A eine gelesene und B eine gehörte Liste von Items sein, die zeitlich separiert dargeboten werden. In der Inklusionsbedingung sind Items beider Listen als „alt" zu beurteilen, in der Exklusionsbedingung dagegen nur B-Items. Die Idee ist, daß A-Items, die in der Exklusionsbedingung dennoch als „alt" beurteilt werden, nicht bewußt erinnert worden sein können; sie müssen also aufgrund ihrer Vertrautheit für alt gehalten werden. Im Gegensatz dazu kann ein A-Item in der Inklusionsbedingung — wie von Mandler (1980) angenommen — sowohl aufgrund von Vertrautheit als auch aufgrund von bewußter Erinnerung „alt" genannt werden. Kurz: Für die „Alt"-Beurteilungswahrscheinlichkeit von A-Items gilt in der Inklusionsbedingung die Modellglei-

chung $p_{Ai} = R + (1-R) \cdot F$, die dem Mandlerschen Modell äquivalent ist, und in der Exklusionsbedingung die Gleichung $p_{Ai} = (1-R) \cdot F$,. Beide Modellgleichungen zusammen liefern ein saturiertes Modell, für das ML-Schätzer der Parameter R und F ganz einfach aus den Trefferraten in der Inklusions- und der Exklusionsbedingung abgeleitet werden können (Jacoby, 1991; Buchner et al., 1995).

Trotz der zahlreichen erfolgreichen Anwendungen dieses Modells, in denen R und F durch verschiedene experimentelle Manipulationen immer wieder in prognostizierbarer Weise selektiv beeinflußt werden konnten, gab es auch recht häufig Kritik, die u.a. die mangelnde Berücksichtigung von Antworttendenzen im Modell von Jacoby (1991) betraf (Graf & Komatsu, 1994; Komatsu, Graf & Uttl, 1995; Roediger & McDermott, 1994; Vaterrodt-Plünnecke, 1994). Vaterrodt-Plünnecke (1994) zeigte die Möglichkeit der Einbeziehung von Rateparametern anhand eines MVB-Modells zu einem Wort-Nonwort-Entscheidungsparadigma auf. Buchner et al. (1995) schlugen eine einfache Erweiterung des Modells von Jacoby (1991) auf die Distraktoritems vor, so daß anhand der falschen Alarme Ratewahrscheinlichkeiten g_i und g_e für die Inklusions- bzw. die Exklusionsbedingung geschätzt werden können. Analog zum 1HT-Modell der Rekognition lassen sich damit Rateprozesse für A-Items modellieren, indem man annimmt, daß Vpn mit Wahrscheinlichkeit g_i bzw. g_e „alt" raten, wenn ein A-Item weder bewußt erinnert wird noch vertraut erscheint. Buchner et al. zeigten außerdem die Überlegenheit dieses erweiterten Meßmodells (EMM) auf, indem sie in drei Experimenten Antworttendenzen manipulierten. Diese Manipulationen affizierten die Gedächtnisparameter des Jacoby-Modells stark, die Parameter des EMM jedoch kaum oder gar nicht.

Yonelinas, Regehr und Jacoby (1995) favorisierten einen alternativen Zugang zur Modellierung von Antworttendenzen in der PDP. Sie nahmen an, daß „Vertrautheit" keine diskrete Größe mit zwei möglichen Ausprägungen ist (vertraut vs. nicht vertraut), sondern wie im Modell der Signalentdeckungstheorie stetig verteilt ist, wobei die Vertrautheiten von A-Items und N-Items jeweils Normalverteilungen folgen, die um den Betrag d' gegeneinander verschoben sind. In ihrem „dual-process signal detection model" (DPSDM) wird angenommen, daß A-Items mit Wahrscheinlichkeit R bewußt erinnert werden, was in der Inklusionsbedingung zu einem „Alt"-Urteil und in der Exklusionsbedingung zu einem „Neu"-Urteil führt. Wird ein A-Item mit Wahrscheinlichkeit $(1-R)$ nicht bewußt erinnert, so erfolgt die Alt-Neu-Entscheidung vertrautheitsbasiert wie im Signalentdeckungsmodell angenommen (vgl. Abschnitt 3.3, Gleichungen (3.3-1) und (3.3-2)). Der Vertrautheitsparameter ist in diesem Modell also d', der Antworttendenzparameter ist c. Yonelinas, Kroll, Dobbins, Lazzara und Knight (1998) haben dieses Modell benutzt, um Daten zum Gedächtnisdefizit von Amnestikern zu reanalysieren. Demnach zeichnen sich Amnestiker insbesondere durch ein starkes Defizit in der expliziten (bewußten) Erinnerung aus. Anders als beim normalen Alterungs-

prozeß konnte jedoch darüber hinaus ein zusätzliches Defizit im Bereich impliziter (vertrautheitsbasierter) Gedächtnisprozesse festgestellt werden, das jedoch deutlich kleiner als das Defizit im expliziten Gedächtnis ausfiel. Dieser Befund steht im Gegensatz zu den Ergebnissen, die von Verfaellie und Treadwell (1993) für die gleichen Daten mit dem Originalmodell von Jacoby (1991) ermittelt wurden. Ursache für die Diskrepanz ist offensichtlich die Nichtberücksichtigung von Antworttendenzen im Jacoby-Modell.

Yonelinas und Jacoby (1996a; 1996b) haben das DPSDM und das EMM anhand von memory operating characteristics (MOC) für die Inklusions- und die Exklusionsbedingung deskriptiv miteinander verglichen, wobei die MOC-Daten mittels Ratingskalenmethodik gewonnen wurden (Yonelinas, 1994). Erdfelder und Buchner (1998b) haben entgegnet, daß dieser Vergleich nicht zulässig ist, da das EMM explizit als Modell für Ja-Nein- bzw. Alt-Neu-Entscheidungen formuliert wurde. Wenn man es auf Ratingskalenurteile anwenden will, muß man das Modell zunächst entsprechend erweitern. Eine entsprechende Erweiterung des EMM auf Ratingskalen (EMM-RS) paßte ausgesprochen gut auf die Ratingskalendaten von Yonelinas (1994), während das DPSDM zwar deskriptiv recht gut zu passen schien, auf der Grundlage eines G^2-Modellanpassungstests jedoch klar verworfen werden muß (Erdfelder & Buchner, 1998b). Weitere Befunde, die für das DPSDM sehr kritisch sind, haben jüngst Glanzer und Mitarbeiter berichtet (Glanzer, Hilford, Kim & Adams, 1999; Glanzer, Kim, Hilford & Adams, 1999).

Aus der Überlegenheit des EMM-RS bei Ratingskalenuntersuchungen kann nicht auf die generelle Überlegenheit des EMM als Meßmodell für die PDP geschlossen werden. Tatsächlich gibt es triftige Gründe, die gegen die generelle Adäquatheit dieses Modells sprechen, insbesondere wegen der problematischen Annahme einer einzigen hohen Schwelle, die in das Modell eingeht (Bredenkamp, 1998a; Erdfelder & Buchner, 1998b; Vaterrodt-Plünnecke et al., 1996). Diese Probleme lassen sich jedoch mit Zwei-Hochschwellen-Modifikationen des Modells ähnlich gut lösen wie bei einfachen Rekognitionsmodellen und Quellendiskriminationsmodellen (siehe letzter Abschnitt). Hierzu ist lediglich ein Distraktorentdeckungsparameter d zusätzlich in das Modell aufzunehmen. Er reflektiert die Möglichkeit, daß ein Distraktoritem als „sicher neu" erkannt wird, z.B. aufgrund von metakognitiven Urteilsstrategien (Strack & Bless, 1994). Damit das Gesamtmodell identifizierbar wird, kann der d-Parameter mit dem Parameter für bewußte Erinnerung gleichgesetzt werden (Vaterrodt-Plünnecke et al., 1996). Liegen mehrere Klassen alter Items vor, kann d mit dem gewichteten Mittel ihrer Erinnerungsparameter gleichgesetzt werden (Erdfelder & Buchner, 1995). Möglich ist auch, d ohne jede Koppelung an die Erinnerungsparameter dadurch identifizierbar zu machen, daß man Antworttendenzen manipuliert und d über die verschiedenen Antworttendenzbedingungen hinweg als konstant annimmt (Erdfelder &

Buchner, 1995). Erfreulicherweise führten alle geschilderten Wege in den bisherigen Anwendungen zu ähnlichen Schätzungen von *d*. Die Wahl der Methode scheint also die primär interessierenden Parameter für explizites und implizites Gedächtnis nur marginal zu affizieren. Unbefriedigend bleibt jedoch, daß es meines Wissens kein Argument gibt, mit dem man die Gleichsetzung von *d* und bewußten Erinnerungswahrscheinlichkeiten theoretisch begründen könnte. Wenn möglich sollte man daher auf diese Gleichsetzungsrestriktion verzichten und stattdessen *d* eher über eine Antworttendenzmanipulation identifizierbar machen.

Eine weitere Frage, die gegenwärtig heftig diskutiert wird, ist die nach der Beziehung zwischen expliziten (bewußten) und impliziten (unbewußten) Gedächtnisprozessen, die durch Meßmodelle zur Prozeßdissoziationsprozedur erfaßt werden (Buchner & Erdfelder, 1996; Curran & Hintzman, 1997; Dehn & Engelkamp, 1997; Hillstrom & Logan, 1997; Hintzman, 1993; Hintzman & Curran, 1997; Hirshman, 1998a; Jacoby, Begg & Toth, 1997; Jacoby & Shrout, 1997; Jacoby, Yonelinas & Jennings, 1997; Joordens & Merikle, 1993; Krüger, 1999; Krüger & Vaterrodt-Plünnecke, 1997; Russo & Andrade, 1995; Russo, Cullis & Parkin, 1998; Wainwright & Reingold, 1996). Jacoby (1991) verband sein Meßmodell ausdrücklich mit einer stochastischen Unabhängigkeitsannahme für die Beziehung zwischen bewußter Erinnerung und Vertrautheit. Buchner et al. (1995) zeigten, daß diese Annahme nicht notwendig ist, wenn man F als *bedingte* Wahrscheinlichkeit eines Vertrautheitsgefühles im Falle fehlender bewußter Erinnerung interpretiert. Einige Autoren haben diese Möglichkeit des Verzichts auf die Unabhängigkeitsannahme aufgegriffen (z.B. Stolz & Merikle, 1995). Macht man die Unabhängigkeitsannahme dennoch, so handelt es sich um eine Zusatzannahme, die innerhalb von Jacobys Modellansatz nicht testbar ist. Die Annahme bezieht sich auf die Gleichheit der bedingten Wahrscheinlichkeiten für Vertrautheit, gegeben eine bewußte Erinnerung bzw. gegeben keine bewußte Erinnerung. Da nur die letztgenannte bedingte Wahrscheinlichkeit identifizierbar ist, kann man jede beliebige Annahme zu der erstgenannten machen, ohne daß dies irgendwelche Konsequenzen für die Modellanpassungsergebnisse und die Werte der anderen Modellparameter hat. Man kann also auch sagen, daß die Unabhängigkeitsannahme innerhalb des Jacoby-Modells metaphysisch ist. Die zahlreichen Befunde, die Jacoby und Mitarbeiter zur Stützung ihrer Unabhängigkeitsannahme zusammengestellt haben (z.B. Jacoby, Yonelinas & Jennings, 1997), ändern an der Korrektheit dieses Arguments nichts. Sie belegen nur die selektive Beeinflußbarkeit des Erinnerungs- und des Vertrautheitsparameters und demonstrieren damit Kontextunabhängigkeit oder „funktionale Unabhängigkeit" der Parameter, nicht jedoch „Prozeßunabhängigkeit", d.h. stochastische Unabhängigkeit bewußter und unbewußter Prozesse, um die es in der Debatte eigentlich geht (Erdfelder & Buchner, 1995; 1998b; Hillstrom & Logan, 1997; Hirshman, 1998a). Lediglich unter Hinzufügung von Zusatzannahmen ver-

raten Dissoziationsbefunde etwas über die Prozeßunabhängigkeit; diese Zusatzannahmen sind jedoch modellimmanent nicht überprüfbar (zusammenfassend: Krüger, 1999).

Es gibt Auswege aus dem Identifizierbarkeitsproblem, allerdings nur um den Preis, daß man das zugrunde liegende empirische Paradigma und das Meßmodell erweitert. Damit sind natürlich zwangsläufig weitere Annahmen verbunden, die potentiell falsch sein können. Der Vorteil dieses Ansatzes besteht jedoch darin, daß die Annahmen in testbare Modelle einmünden und somit nicht mehr metaphysisch sind wie die Unabhängigkeitsannahme im Jacoby-Modell. Ein solcher Ansatz stammt von Krüger und Vaterrodt-Plünnecke (Krüger, 1995; Krüger & Vaterrodt-Plünnecke, 1997). Der Ansatz basiert auf einer Erweiterung der PDP um eine indirekte Testbedingung, von der angenommen werden kann, daß sie nur unbewußte Gedächtnisprozesse reflektiert. Zu allen drei Testbedingungen (Inklusion, Exklusion, indirekte Bedingung) formulierten Krüger und Vaterrodt-Plünnecke ein verbundenes MVB-Modell, innerhalb dessen die beiden bedingten Wahrscheinlichkeiten für bewußte Erinnerungsprozesse, gegeben Vertrautheit bzw. gegeben keine Vertrautheit identifizierbar sind. Somit wird die stochastische Unabhängigkeitsannahme über eine Gleichheitsrestriktion testbar (vgl. Abschnitt 4.3.2.1).

Ein zweiter Ansatz, der allerdings nicht in die Klasse der MVB-Modelle fällt und auch noch nicht auf empirische Daten angewandt wurde, stammt von Erdfelder und Buchner (1995). Er basiert auf dem DPSDM-Modell von Yonelinas et al. (1995), welches ebenfalls auf einer Unabhängigkeitsannahme aufbaut. Wenn man diese Unabhängigkeitsannahme fallen läßt, entsteht aus dem DPSDM das „correlated processes signal-detection model" (CPSDM), welches im Gegensatz zum DPSDM zuläßt, daß bewußt erinnerte Items am oberen oder am unteren Rand der Vertrautheitsverteilung liegen können. Das Modell ist anhand von MOC-Kurven prinzipiell ebenso testbar wie das DPSDM. Es enthält das DPSDM von Yonelinas et al. (1995) als Spezialfall.

Die Befunde von Krüger und Vaterrodt-Plünnecke (1997) schließen extreme Formen der Abhängigkeit von bewußten und unbewußten Prozessen zwar aus, werfen jedoch zugleich Zweifel an der Gültigkeit der Unabhängigkeitsannahme auf. Dies harmoniert mit Überlegungen von Erdfelder und Buchner (1995) zur Plausibilität verschiedener Korrelationen zwischen bewußten und unbewußten Prozessen innerhalb des CPSDM. Extrem positive Korrelationen (d.h. die bewußt erinnerten Items sind zugleich die vertrautesten) und extrem negative Korrelationen (d.h. die bewußt erinnerten Items sind gerade die mit den niedrigsten Vertrautheitswerten) implizieren nämlich auf dem Hintergrund des CPSDM partiell „flache" (d.h. mit der Rate falschen Alarms nicht ansteigende) MOC-Kurven. Solche MOC-Kurven sind aber empirisch nie gefunden wurden.

MVB-Modelle zur PDP wurden häufig — wie schon in der Originalarbeit von Jacoby (1991) — im Zusammenhang mit Rekognitionstests angewendet, beispielsweise zur Untersuchung der Frage, ob der Generierungseffekt (Rekognitionsvorteil generierter im Vergleich zu gelesenen Items; vgl. Buchner et al., 1995) oder der Handlungseffekt (Rekognitionsvorteil ausgeführter im Vergleich zu gelesenen Items, vgl. Dehn & Engelkamp, 1997; Dehn, 1998; Meiser, 2000) auf bewußten Erinnerungsprozessen, auf Vertrautheitsprozessen oder auf beidem beruht. Daneben gibt es jedoch auch zahlreiche Anwendungen in Verbindung mit Wort-Nonwort-Entscheidungen (Ott, Curio & Scholz, 2000; Scholz, Ott & Müller-Sinik, 1997; Vaterrodt-Plünnecke, 1994; Windmann & Krüger, 1998) mit Wortstammergänzungen (Bröder & Bredenkamp, 1996; Krüger, 1999; McBride & Dosher, 1999; Schmitter-Edgecombe, 1999; Vaterrodt-Plünnecke et al., 1996) und mit Berühmtheitsurteilen (Buchner & Wippich, 1996; Steffens et al., 1999). Als im klinischen Kontext besonders fruchtbar hat sich das Modell von Vaterrodt-Plünnecke (1994) zu Wort-Nonwort-Entscheidungen herausgestellt. In der klinischen Psychologie ist ein modellbasierter Zugang von besonderer Bedeutung, da sich traditionelle Ansätze der Untersuchung impliziter Informationsverarbeitungsprozesse als methodisch und konzeptuell fragwürdig herausgestellt haben (Scholz, 1997). Scholz et al. (1997) und Ott et al. (2000) schlugen deshalb als alternative Erfassungsmöglichkeit für implizite Informationsverarbeitungsprozesse das Modell von Vaterrodt-Plünnecke vor. Sie boten Probanden mit und ohne Tendenzen zu somatoformen Störungen (d.h. körperliche Beschwerden mit unklaren medizinischen Ursachen) zunächst bedrohliche und neutrale Wörter mit der Instruktion dar, sie zu einem Satz zu ergänzen. In der anschließenden Testphase wurde alte und neue Wörter sowie daraus durch Phonemvertauschung konstruierte Nonwörter akustisch maskiert dargeboten, und die Probanden mußten entscheiden, ob es sich um ein Wort oder um ein Nonwort handelt. Das Modell von Vaterrodt-Plünnecke (1994) geht davon aus, daß ein Wort bzw. Nonwort entweder bewußt erkannt werden kann oder nicht. Wird es nicht bewußt erkannt, kann bei zuvor dargebotenen (alten) Wörtern und den dazu konstruierten Nonwörtern ein impliziter Vertrautheitseffekt zu einer „Wort"-Entscheidung führen, und zwar über die Ratewahrscheinlichkeit für „Wort" hinaus. Es zeigte sich, daß der Parameter für implizite Erfahrungsnachwirkungen bei bedrohlichen Wörtern wesentlich stärker ausfiel als bei neutralen Wörtern (Ott et al., 2000), und zwar insbesondere bei Probanden mit Tendenzen zu somatoformen Störungen (Scholz et al., 1997). Unterschiede in den Parametern für explizites Erkennen und Raten fielen demgegenüber vergleichsweise gering aus. Signifikante Korrelationen zwischen den getrennt für jede einzelne Vp geschätzten Parametern und klinischen Symptomstärkemaßen konnten nicht festgestellt werden, was jedoch an den hohen Standardschätzfehlern der individuellen Parameterschätzungen liegen kann

(vgl. Abschnitt 5.4.1). Tendenziell sprechen die Ergebnisse klar für einen Zusammenhang zwischen der Stärke somatoformer Störung und impliziten Erfahrungsnachwirkungen bedrohlicher Wörter.

Vergleicht man die Ergebnisse von Scholz et al. (1997) mit denen von Windmann und Krüger (1998), so wird deutlich, daß man bei Verallgemeinerungen über klinischen Gruppen hinweg zur Zeit noch vorsichtig sein muß. Obwohl Windmann und Krüger (1998) ein sehr ähnliches Paradigma und dasselbe Analysemodell benutzt haben, fanden sie keinen Effekt im impliziten Gedächtnisparameter, stattdessen jedoch einen Effekt in den Ratewahrscheinlichkeiten zugunsten bedrohlicher Wörter. Denkbar ist, daß das an der von ihnen untersuchten Probandengruppe liegt (Panikpatienten). Möglich ist auch, daß die visuelle (statt akustische) Darbietung der Wort-Nonwort-Entscheidungsaufgabe ein wichtige Rolle spielt. Zwar wurden die Wörter von Windmann und Krüger (1998) ebenfalls nur sehr kurz und visuell maskiert dargeboten; dennoch kann der akustische Maskierungseffekt in der Studie von Scholz et al. (1997) effektiver gewesen sein, wofür auch der niedrigere Parameter für bewußtes Erkennen bei bedrohlichen Wörtern spricht (Ott et al., 2000).

Inzwischen sind auch einige Verbesserungs- bzw. Optimierungsmöglichkeiten für das experimentelle Vorgehen innerhalb der PDP vorgeschlagen worden. Die Ergebnisse von Krüger (1999) legen eine Vereinfachung der PDP für Wortstammergänzungsaufgaben nahe: Auf die Exklusionsbedingung kann im Prinzip verzichtet werden, wenn man bei Wörtern, die in der Inklusionsbedingung ergänzt werden, zusätzlich erfragt, ob das betreffende Wort in der Akquisitionsphase vorgekommen ist oder nicht. Krüger (1999) hat seine modifizierte Prozedur für Wortstammergänzungen experimentell validiert und mit der Standardversion der PDP für Wortstammergänzungen verglichen. Er wies nach, daß die modifizierte Prozedur bei weniger Aufwand Ergebnisse liefert, die der Standard-PDP weitgehend gleichwertig sind. Buchner, Erdfelder et al. (1997) haben eine ähnliche Vereinfachungsmöglichkeit der PDP für Rekognitionsurteile aufgezeigt: Statt separater Inklusions- und Exklusionstestbedingungen kann einfach ein Quellendiskriminationstest durchgeführt werden, bei dem die Probanden für als „alt” beurteilte Items angeben sollen, ob sie der A-Liste oder der B-Liste entstammen. Die Autoren zeigten, daß sich die Daten einer gewöhnlichen Inklusions- versus Exklusionsaufgabe aus der Quellendiskriminationsaufgabe nahezu perfekt vorhersagen lassen. Dies wurde durch Steffens, Buchner, Martensen und Erdfelder (in press) mit veränderten Inklusions- und Exklusionsinstruktionen repliziert. Auch für die PDP in Verbindung mit Berühmtheitsurteilen ließ sich eine vereinfachte Quellengedächtnisaufgabe finden, die die Daten einer herkömmlichen Inklusions- versus Exklusionsaufgabe sehr gut vorhersagen konnte. Insgesamt spricht somit viel für die Annahme, daß die PDP Prozesse aktiviert, die denen eines Quellengedächtnistests

sehr ähnlich sein können (vgl. auch Mulligan & Hirshman, 1997). Ich werde im folgenden Kapitel darauf zurückkommen.

Es gibt einige Hinweise darauf, daß nicht alle Varianten der PDP mit demselben Meßmodell adäquat analysiert werden können (siehe bereits Vaterrodt-Plünnecke et al., 1996). Selbst wenn man die Aufgabe für die Vp konstant hält, jedoch die Instruktion leicht variiert, kann das für die Adäquatheit eines Meßmodells wichtig sein. Jacoby (1998) wendete das EMM von Buchner et al. (1995) beispielsweise erfolgreich auf eine Wortstammergänzungsaufgabe an, bei der die Probanden zunächst instruiert wurden, einen Wortstamm als Abrufhilfe für ein zuvor dargebotenes Wort zu benutzen (direct-retrieval instruction). In der Inklusionsbedingung sollte dieses Wort dann genannt werden bzw. nur dann ein anderes Wort ergänzt werden, wenn der Abrufversuch scheitert. In der Exklusionsbedingung sollte das eventuell abgerufene Wort dagegen nicht ergänzt werden, sondern ein anderes Wort, das nicht alt war. Bei dieser Instruktion wirkte sich eine Aufmerksamkeitsmanipulation (Experiment 1) und eine Darbietungszeitmanipulation (Experiment 3) in der Akquisitionsphase hypothesengemäß nur auf den Parameter für explizite Erinnerung aus, nicht dagegen auf den impliziten Gedächtnisparameter.[84] Gibt man stattdessen beim Abruf die Instruktion, den Wortstamm zunächst zu einem beliebigen Wort zu ergänzen, welches in der Inklusionsbedingung immer genannt werden soll, in der Exklusionsbedingung dagegen nur dann, wenn es neu ist, dann wird der EMM-Parameter für implizites Gedächtnis durch die Aufmerksamkeits- und Darbietungszeitmanipulation plötzlich affiziert. Dieses hypothesenwidrige Ergebnismuster verschwindet dagegen im Rahmen einer Modifikation des EMM, bei der bewußte Erinnerungsprozesse — im Einklang mit der Instruktion — auf der letzten statt auf der ersten Ebene des Verarbeitungsbaummodells angesiedelt werden (vgl. Jacoby, 1998, S. 25).

Eine interessante Alternative zur PDP haben jüngst Brainerd, Reyna und Mojardin (1999) vorgeschlagen. Ihr Ansatz der „verbundenen Rekognition" (conjoint recognition, CR) hat den Vorteil, daß die empirische Basis für die Modellbildung breiter ist als im Falle der PDP, so daß viele Annahmen, die in die PDP unüberprüft eingehen, im Rahmen des CR-Paradigmas testbar sind. Brainerd et al. stellen ihr CR-Modell explizit als „second-generation model" vor, das aus den Problemen der PDP die richtigen Lehren zu ziehen versucht. Im CR-Paradigma lernen Probanden zunächst eine Liste von Items. Im anschließenden Gedächtnistest werden — anders als in gewöhnlichen Ja-Nein-Rekognitionstests — drei Klassen von Items vorgelegt: alte Items (targets oder „T-items"), neue Items, die den alten Items semantisch ähnlich sind („R-Items"), und schließlich neue Items, die keine Bezie-

[84] Jacoby (1998) nennt diesen Test der selektiven Beeinflußbarkeit eines Modellparameters „Test der Unabhängigkeitsannahme". Aus den oben genannten Gründen darf dies jedoch nicht im Sinne eines Tests der Prozeßunabhängigkeitsannahme mißverstanden werden.

hung zu den alten Items aufweisen („U-Items"). Alt-Neu-Rekognitionsurteile werden unter drei Rekognitionsinstruktionen erhoben: (a) nur T-Items sind „alt" zu nennen (T-Instruktion), (b) nur R-Items sind alt zu nennen (R-Instruktion) und (c) T- oder R-Items sind alt zu nennen, U-Items dagegen neu (T+R-Instruktion). Die Kreuzklassifikation der drei Testitemklassen mit den drei Instruktionen ergibt neun Beobachtungskategorien mit $3 \cdot 2 = 6$ unabhängigen Kategoriewahrscheinlichkeiten.[85]

Brainerd et al. schlugen ein MVB-Meßmodell zu dieser Testprozedur vor, das drei Klassen von Prozessen annimmt: Identitätsfeststellung, Ähnlichkeitsfeststellung und Raten. Das erste Konstrukt „Identitätsfeststellung" ist dem der bewußten Erinnerung bei Mandler (1980) und Jacoby (1991) sehr ähnlich. Unter T- und T+R-Instruktionen führt dieser Prozeß für T-Items — wie in der Inklusionsbedingung der PDP — zu einem „Alt"-Urteil, unter R-Instruktionen dagegen — wie in der Exklusionsbedingung der PDP — zu einem „Neu"-Urteil. Identitätsfeststellungen dominieren also Gedächtnisurteile im Conjoint-Recognition-Paradigma (CR-Paradigma) ähnlich wie bewußte Erinnerungsprozesse in der PDP.

Demgegenüber sind die Konstrukte „Ähnlichkeitsfeststellung" bei Brainerd und „Vertrautheitsgefühl aufgrund vorheriger Verarbeitung" bei Jacoby kaum gleichsetzbar, da nach dem Modell von Brainerd et al. auch neue R-Items Ähnlichkeitsfeststellungen auslösen können. Ähnlichkeitsfeststellungen können im CR-Modell nur dann zu „Alt"-Urteilen führen, wenn zuvor die Identität eines Testitems mit einem alten Item negiert wurde. Wird sowohl die Identität als auch die Ähnlichkeit negiert, kann über Rateprozesse dennoch ein „Alt"-Urteil resultieren. Die Ratewahrscheinlichkeiten für die verschiedenen Instruktionen werden ähnlich wie im EMM von Buchner et al. (1995) aus den unrelatierten Distraktoren (U-Items) geschätzt.

Das CR-Modell wurde in drei Experimenten validiert, die unterschiedliche R-Items in der Testphase verwendeten (Experiment 1: Synonyme, Experiment 2: Antonyme, Experiment 3: Oberkategoriebezeichnungen). Erwartungsgemäß erwiesen sich die Gedächtnisparameter über die drei Instruktionsbedingungen (T, R und T+R) hinweg als invariant, die Rateparameter dagegen nicht. Umgekehrt wirkte sich die Häufigkeit der Darbietung eines Items in der Akquisitionsphase erwartungsgemäß auf den Identitätsfeststellungsparameter aus. Der Ähnlichkeitsfeststellungsparameter für T-Items variierte weder zwischen den experimentellen Bedingungen noch zwischen den Experimenten; demgegenüber sank der Ähnlich-

[85] Auch hier liegt natürlich die Vermutung nahe, daß man die experimentelle Prozedur durch einen einheitlichen Itemklassenbeurteilungstest vereinfachen kann: Die Probanden müßten — ähnlich wie in einem Zwei-Quellen-Diskriminationstest — zunächst beurteilen, ob das Item semantisch zur alten Itemliste paßt, und wenn ja, ob das Item in der alten Liste wörtlich enthalten war. Diese Vereinfachungsmöglichkeit wurde von Brainerd et al. noch nicht exploriert.

keitsfeststellungsparameter für R-Items mit der Lernintensität — d.h. mit zunehmender Darbietungshäufigkeit eines Items in der Lernphase — ab. Letzteres ist eine Vorhersage des zugrundeliegenden „Fuzzy-Trace"-Gedächtnismodells von Brainerd und Mitarbeitern. Sie nehmen an, daß für jedes gelernte Item zwei Gedächtnisrepräsentationen angelegt werden: Eine wörtliche Repräsentation (verbatim trace) und eine semantische Repräsentation (gist trace). Von der relativen Stärke der wörtlichen und semantischen Repräsentationen hängt die Rate falsche Alarms bei R-Distraktoren ab: Ist die wörtliche Repräsentation stark, wird sie eher durch den R-Distraktor aktiviert als die semantische Repräsentation, was tendenziell zu wenig Ähnlichkeitsurteilen führt; ist das „verbatim trace" dagegen vergleichsweise schwach, wird durch einen R-Distraktor eher ein „gist trace" aktiviert, was sehr wahrscheinlich ein Ähnlichkeitsurteil zur Folge hat.

Insgesamt liegt somit ein zufriedenstellendes Validierungsergebnis vor. Erste entwicklungspsychologische Anwendungen des Modells auf den „false memory effect" (extrem hohe Raten falschen Alarms bei Distraktoren, die den Zielitems ähnlich sind, vgl. Abschnitt 7.1.9) sowie auf Effekte nachträglicher Falschinformation (vgl. dazu Kapitel 10) wurden bereits publiziert (Brainerd & Reyna, 1998a). Abgesehen von zusätzlichen Befunden zur Validität der Modellparameter präsentierten die Autoren Evidenz zugunsten der Hypothese einer globalen Verbesserung der Gedächtnisparameter „Identitätsfeststellung" und „Ähnlichkeitsfeststellung" im Kindesalter. Interessanterweise ist dieser Entwicklungsprozeß in den Parameterschätzungen klar zu erkennen, nicht aber in den falschen Rekognitionsraten verschiedener Altersgruppen. Dies liegt daran, daß „Identitätsfeststellung" und „Ähnlichkeitsfeststellung" antagonistische Effekte auf falsche Alarme ausüben: Verbesserte Identitätsfeststellung reduziert falsche Alarme, während verbesserte Ähnlichkeitsfeststellung falsche Alarme bei verwandten Distraktoren fördert. Beide Effekte können sich in der Wirkung aufheben, so daß auf der Rohdatenebene der falsche Eindruck entsteht, als würden die kognitiven Prozesse, welche falsche Rekognitionen generieren, keinem Entwicklungstrend unterliegen.

Die bisherigen Anwendungen des CR-Modells von Brainerd und Mitarbeitern sind interessant und vielversprechend. Es bleibt abzuwarten, ob sich das Modell für die Gedächtnispsychologie als ähnlich fruchtbar wie die PDP herausstellen wird.

7.1.4 Speicherung und Abruf

Ein bekanntes und äußerst stabiles Phänomen bei der freien Reproduktion ist „Clustering": Zusammengehörige Items — beispielsweise Wörter einer semantischen Kategorie — werden tendenziell unmittelbar nacheinander reproduziert, auch wenn sie in einer Lernliste zufällig angeordnet oder sogar gezielt separiert

dargeboten wurden (Bredenkamp & Wippich, 1977b). Batchelder und Riefer (1980; 1986) sahen in diesem Clustering-Phänomen einen Ansatzpunkt für die Separierung von Speicher- und Abrufprozessen: „Clustering" ist für sie Ergebnis des Zusammenwirkens einer erfolgreichen gemeinsamen Speicherung der zusammengehörigen Items und des anschließenden erfolgreichen Abrufs dieser gemeinsamen Gedächtnisspur nach dem Alles-oder-Nichts-Prinzip: Das Cluster ist nur komplett oder aber gar nicht abrufbar, was in einer gemeinsamen Reproduktion aller zusammengehörigen Item (E1) bzw. im völligen Mißlingen der Reproduktion (E4) seinen Ausdruck findet. Durch Kontrastierung dieser beiden Ereigniskategorien mit dem erfolgreichen, aber nicht gemeinsamen Abruf aller Items einer Kategorie (E2) bzw. dem nur partiell erfolgreichen Abruf (E3) erhält man eine Datenbasis, die differenziert genug ist, um ein identifizierbares Modell zu konstruieren, das die Schätzung der Cluster-Enkodierungswahrscheinlichkeit, der Cluster-Abrufwahrscheinlichkeit, und der Einzelitem-Reproduktionswahrscheinlichkeit ermöglicht.

Das auf diesen Ideen beruhende Enkodierung-Abruf-Modell (EA-Modell) von Batchelder und Riefer und dessen Anwendungen wurde in Abschnitt 3.5.1 bereits ausführlich behandelt, so daß sich eine erneute Besprechung an dieser Stelle erübrigt. Insgesamt dürfte es sich — abgesehen von den Schwellenmodellen des Rekognitionsgedächtnisses — um das bislang einflußreichste MVB-Modell handeln. Trotz der bislang schon recht zahlreichen Arbeiten zum EA-Modell bieten sich weitere Anwendungen zur Klärung der Frage an, welchen Anteil Enkodierung (verstanden als die Fähigkeit, Information als Cluster zu speichern) und Abruf an bestimmten Gedächtnisphänomenen haben. Exemplarisch sei die Diplomarbeit von Schilken (1998) genannt, die aufzeigte, daß Epileptiker mit linkstemporalem Fokus tendenziell mehr Schwierigkeiten beim Abruf geclusterter Informationen haben als Epileptiker mit rechtstemporalem Fokus.

Die meisten Arbeiten zum EA-Modell fallen in die Kategorie *exploratorischer Anwendungen* von MVB-Modellen: Ein empirisch adäquates und als psychologisch gültig angesehenes Modell wird benutzt, um Antworten auf Fragen zu bestimmten Gedächtnisphänomenen zu finden (z.B. „Ist das episodische Gedächtnisdefizit im Alter primär enkodierungs- oder primär abrufbedingt?"). Eine Apriori-Hypothese zur korrekten Antwort muß nicht unbedingt existieren. Jede Antwort ist im Prinzip denkbar; scheitern kann die Modellanwendung allein an schlechten Werten der Modellanpassungsstatistik.

Neben dem EA-Modell sind weitere MVB-Modelle zur Separierung von Speicher- und Abrufprozessen vorgeschlagen worden, die auf anderen Annahmen basieren und deshalb nicht notwendig vergleichbare Speicher- und Abrufprozesse abbilden. Das „Trace Susceptibility Model" (Chechile, 1987; Chechile & Meyer, 1976) basiert im Kern auf einem Zwei-Phasen-Modell der Reproduktion, wonach Re-

produktion neben erfolgreicher Speicherung auch erfolgreichen Abruf erfordert, während erfolgreiche Speicherung für die Rekognition hinreichend ist. Empirisch angewandt wird diese Idee auf ein Paradigma, in dem reine Reproduktionstests in zufälligen Abständen durch Rekognitionstests mit Konfidenzratings ersetzt werden. Für Reproduktions-, Rekognitions- und Konfidenzdaten wird ein gemeinsames MVB-Modell formuliert, das neben Speicherungs- und Abrufparametern einige Antworttendenz- bzw. Rateparameter für die Rekognitionstests umfaßt. Das Modell ist ausgiebig validiert und auf viele unterschiedliche Gedächtnisphänomene angewandt worden, vorwiegend jedoch auf das sogenannte Brown-Peterson-Paradigma zur Untersuchung des Arbeitsgedächtnisses. Im Verlaufe des Forschungsprogramms wurden mehrere Phänomene identifiziert, die nur auf Speicherprozessen bzw. nur auf Abrufprozessen basieren (zusammenfassend: Batchelder & Riefer, 1999; Brainerd, 1985). Eine leicht modifizierte Version des Modells wurde jüngst von Chechile und Soraci (1999) zur Untersuchung des Generierungseffekts (generation effect) eingesetzt. Selbst generierte Information wird nach ihren Resultaten später vor allem deshalb besser erinnert als gelesene Information, weil die Generierung die Speicherung der Information verbessert. Die Abrufbarkeit der Information wird durch die Generierung dagegen nur unwesentlich beeinflußt, tendenziell sogar negativ. Diese Ergebnisse passen gut zu den Vorhersagen von Zwei- und Drei-Faktoren-Theorien des Generierungseffekts (vgl. z.B. Hirshman & Bjork, 1988; McDaniel, Waddill & Einstein, 1988).

Interessant ist, daß sich das Chechile-Modell insgesamt sehr gut bewährt hat, obwohl die zugrundeliegende Idee eines Zwei-Phasen-Modells der Reproduktion aus heutiger Sicht sehr fragwürdig ist (siehe z.B. Baddeley, 1997, S. 202f; vgl. auch Abschnitt 7.1.8). Auflösen läßt sich dieser Widerspruch, wenn man das Zwei-Phasen-Modell der Reproduktion nicht als paradigmenunabhängige Theorie, sondern als spezifische Theorie zum Reproduktions-Rekognitions-Paradigma von Chechile auffaßt. Obwohl das Zwei-Phasen-Modell als allgemeine Theorie zurückgewiesen werden muß, kann sie als spezifische Theorie zu einem speziellen Paradigma dennoch approximativ gültig sein. Man muß hierfür lediglich sicherstellen, daß im Reproduktionstest praktisch keinerlei Abrufhilfen präsent sind, während im Rekognitionstest der Enkodierungskontext nahezu kopiert wird, so daß der Abruf der alten Information extrem leicht fallen muß, sofern sie enkodiert wurde. Werden diese Bedingungen eingehalten, so werfen die Annahmen des Chechile-Modells gedächtnispsychologisch keine Probleme auf.

Humphreys und Greeno (1970) haben erstmals vorgeschlagen, das aus der mathematischen Lerntheorie bekannte Markoff-Ketten-Modell mit drei Zuständen (three-state Markov model, vgl. Greeno, 1968) als Speicher-Abruf-Modell für wiederholte Reproduktionstests in Lern-Test-Sequenzen zu verwenden. Wird dieses Modell gedächtnispsychologisch interpretiert, so ist die Grundidee dem Dual-Code-

Markoff-Modell aus Abschnitt 6.5 durchaus verwandt. Anders als das Chechile-Modell bezieht es sich primär nicht auf kurzfristige Speicherung und unmittelbaren Abruf, sondern auf Speicherung und Abruf im bzw. aus dem episodischen Langzeitgedächtnis. Das episodische Langzeitgedächtnis wird jedoch nicht — wie im Falle des Dual-Code-Markoff-Modells — in ein verbales und ein imaginales Subsystem zerlegt, sondern als homogenes System konzipiert, in dem ein Item entweder gespeichert (Zustand 2 oder 3) oder nicht gespeichert ist (Zustand 1). Ein weiterer Unterschied zum Dual-Code-Markoff-Modell besteht darin, daß gespeicherte Items entweder zuverlässig abrufbar sein können (Zustand 3: „algorithmic retrieval") oder aber nur gelegentlich abrufbar sein können (Zustand 2: „heuristic retrieval"). Das Modell kann sich prinzipiell auf beliebig lange Lern-Test-Sequenzen beziehen. Wie auch das modifizierte Dual-Code-Markoff-Modell aus Abschnitt 6.5.2 ist es kein stationäres MK-Modell, sondern läßt unterschiedliche Übergangswahrscheinlichkeiten im ersten Lerndurchgang und in den folgenden Durchgängen zu. Es umfaßt insgesamt 11 Parameter, darunter zwei Speicherungswahrscheinlichkeiten (im 1. bzw. in den folgenden Durchgängen), ein Vergessensparameter (storage loss), vier Abrufperformanzparameter für Zustand 2 und vier Abruflernparameter für den Übergang in Zustand 3 (Brainerd, 1985, S. 166 ff.). Trotz der Vielzahl der Parameter ist das Modell insgesamt befriedigend getestet und validiert; auch die Unterscheidung zwischen Abrufperformanz- und Abruflernparametern hat sich insofern bewährt, als beide Parameterklassen selektiv durch bestimmte experimentelle Manipulationen beeinflußt werden konnten. Entwicklungspsychologische Anwendungen zeigen ziemlich klar, daß alle drei Parameterklassen im Kindesalter profitieren, die Speicher- und Abrufperformanzparameter jedoch etwas stärker als die Abruflernparameter (zusammenfassend: Brainerd, 1985, S. 173 ff).

Eine Variante dieses Modells wurde von Batchelder, Chosak-Reiter, Shankle und Dick (1997) zur Differentialdiagnose von Gedächtnisdefiziten bei verschiedenen Formen und Schweregraden der Demenz eingesetzt. Ihnen gelang es, ein stationäres Drei-Zustands-Modell an wiederholte freie Reproduktionsdaten von Patienten mit vaskulärer Demenz und Morbus Alzheimer anzupassen, allerdings unter Preisgabe der Forderung, daß die Abrufwahrscheinlichkeit im dritten Gedächtniszustand („algorithmic retrieval") perfekt sein muß. Ihr Modell macht die schwächere Annahme, daß die Abrufwahrscheinlichkeit im ersten Zustand („unlearned state") 0, im zweiten Zustand t und im dritten $l \geq t$ ist.[86] Während Morbus Alzheimer (MA) und die vaskuläre Demenz (VD) varianzanalytisch nicht differenzierbar waren, sofern der Schweregrad der Krankheit konstant gehalten

[86]Zur Realisierung dieser Ungleichheitsrestriktion im Rahmen von MVB-Modellen vgl. Abschnitt 4.3.2.1.

wurde, offenbarten die Gruppen unterschiedliche Parameterschätzwerte im Drei-Zustands-Markoff-Modell. Patienten mit beginnender MA zeigten im Vergleich zu Patienten mit beginnender VD und Kontrollpersonen deutlich reduzierte Übergangswahrscheinlichkeiten in den dritten Zustand (langfristige Speicherung mit hoher Abrufwahrscheinlichkeit), während Hinweise auf reduzierte Abrufwahrscheinlichkeiten *l* und *t* bei Alzheimer-Patienten im Vergleich zu VD-Patienten nicht gegeben waren.[87] Insgesamt stützen diese Ergebnisse somit eher strukturelle Theorien der Alzheimer-Krankheit, nach denen das kognitive Defizit im wesentlichen aus beschädigten oder zerstörten Hirnstrukturen herrührt, die für eine langfristige Speicherung von Information unumgänglich sind. Abrufbeeinträchtigungstheorien der AD finden demgegenüber keine Stützung.

Wilkinson und Mitarbeiter (Wilkinson, DeMarinis & Riley, 1983; Wilkinson & Koestler, 1983) haben Speicher-Abruf-Markoff-Modelle für wiederholte Reproduktionsdaten vorgelegt, bei denen zwischen den Reproduktionstests keine zusätzlichen Lerndurchgänge stattfinden. Das einfachste Modell (strengthen-and-discard model, vgl. Wilkinson et al., 1983) nimmt drei Prozesse an: Speicherung, Abruflernen und Abrufvergessen. Erfolgreiche Reproduktionen können nur dann vorkommen, wenn die Speicherung im ersten (und einzigen) Lerndurchgang gelingt. Gelingt die Speicherung nicht oder kann ein gespeichertes Item niemals erfolgreich abgerufen werden, sind Mißerfolge in allen Tests die Konsequenz. Die Abrufwahrscheinlichkeiten gespeicherter Items sind über die Durchgänge hinweg nicht konstant, sondern hängen davon ab, ob der vorhergehende Abruf erfolgreich war (dies bewirkt Abruflernen, d.h. einen Anstieg der Abrufwahrscheinlichkeiten) oder nicht erfolgreich war (dies bewirkt Abrufvergessen, d.h. einen Abfall der Abrufwahrscheinlichkeiten). Im Gegensatz zu den o.g. Modellen legen Untersuchungen auf der Grundlage des Wilkinson-Modells die Vermutung nahe, daß nur die Informationsspeicherung von der Gedächtnisentwicklung bei Kindern profitiert, während Abrufwahrscheinlichkeiten im wesentlichen altersinvariant sind bzw. im Verlaufe der Entwicklung nur minimal ansteigen. Brainerd (1985, S. 192) wies darauf hin, daß die Invarianz der Abrufvergessenswahrscheinlichkeiten im Wilkinson-Modell keinen Widerspruch zu den Ergebnissen bedeutet, die mit anderen Modellen erzielt wurden; durch keines der anderen Modelle wird nämlich Abrufvergessen aus dem Langzeitgedächtnis erfaßt. Die Invarianz des Abruflernens mit dem Alter steht jedoch in Konflikt zu den Ergebnissen, die mit dem Drei-Zu-

[87]Dies kann allerdings auch an den sehr großen Standardschätzfehlern der Abrufwahrscheinlichkeiten liegen, die wiederum eine Konsequenz niedriger Speicherungswahrscheinlichkeiten *a* sind. Die Teststärke des Tests zum Vergleich von Abrufwahrscheinlichkeiten ist als sehr gering. Da die Speicherungswahrscheinlichkeiten von Alzheimerpatienten nicht optimierbar sind, können die in Abschnitt 4.3.3.6 (Maßnahme c) angesprochenen Möglichkeiten der Teststärkeoptimierung hier keine Anwendung finden. Allein die Erhöhung des Stichprobenumfangs kommt als Maßnahme der Poweroptimierung in Frage.

stands-Markoff-Modell ermittelt wurden. Erklären kann man den Widerspruch eventuell damit, daß zusätzliche Lerndurchgänge im Drei-Zustands-Modell vorgesehen sind, im Modell von Wilkinson dagegen nicht.

Kail, Hale, Leonard und Nippold (1984) haben das Strengthen-and-Discard-Modell zur Untersuchung der Gedächtnisdefizite von Kindern mit Sprachstörungen eingesetzt. In ihrer Untersuchung wurde deutlich, daß das Gedächtnisdefizit sprachgestörter Kinder im Vergleich zu Gleichaltrigen, die keine Sprachstörung aufweisen, den Speicherparameter *und* die Abruflernparameter des Modells betrifft, nicht jedoch die Abrufvergessensparameter. Man kann also die Sprachentwicklungsstörung nicht allein auf ein beeinträchtigtes oder reduziertes mentales Lexikon zurückführen; auch der schnelle und korrekte Zugriff auf vorhandene Lexikoneinträge scheint bei sprachgestörten Kindern nur begrenzt trainierbar zu sein.

Die Untersuchung von Kail et al. (1984) macht im übrigen ein Problem bisheriger Anwendungen des Strengthen-and-Discard-Modells deutlich: Nach der zugrundeliegenden Theorie müssen sowohl die Abruflern- als auch die Abrufvergessensparameter über die Reproduktionsdurchgänge hinweg ansteigen bzw. sie dürfen zumindest nicht abfallen. Dies impliziert Ordnungsrestriktionen bezüglich der betreffenden Modellparameter, die jedoch in den berichteten statistischen Analysen keine Berücksichtigung fanden. Als Konsequenz daraus ist die errechnete Modellanpassungsstatistik in einigen Fällen eine Unterschätzung der korrekten Anpassungsstatistik, da hypothesenwidrige nichtmonotone Verläufe der Parameterschätzungen über die Reproduktionsdurchgänge hinweg zu beobachten sind (vgl. Kail et al., 1984, S. 45, Tabelle 4). Mit den in Abschnitt 4.3.2.1 vorgestellten Methoden der Prüfung von Ordnungshypothesen läßt sich dieses Problem bei zukünftigen Modellanwendungen umgehen.

Ein weiteres Speicher-Abruf-Modell, welches Annahmen des Drei-Zustands-MK-Modells und des Strengthen-and-Discard-Modells miteinander verbindet, wurde von Howe und Brainerd vorgelegt (Brainerd, Reyna, Howe & Kingma, 1990; Howe, 1991; 1995; Howe & Brainerd, 1989). Ihr „Spurenintegritätsmodell" (trace-integrity model) bezieht sich ebenfalls auf das Paradigma mehrfach wiederholter Reproduktionstests ohne eingeschobene Lernphasen. Es wird davon ausgegangen, daß das Lernmaterial zunächst einmal gut gelernt wurde (2 oder 3 gelungene Reproduktionen pro Item), bevor nach einem Retentionsintervall mindestens vier erneute Reproduktionsdurchgänge folgen, auf die sich das Modell bezieht (Howe & Brainerd, 1989, S. 321). Das Spuren-Integritätsmodell läßt im Gegensatz zum Modell von Wilkinson Spurenzerfall (storage failure) und Spurenreintegration (trace redintegration) neben Abruflernen und Abrufvergessen zu. Während Spurenzerfall und Abrufvergessen zum Scheitern der Reproduktion zuvor reproduzierbarer Items beitragen können, können Spurenreintegration und Abruflernen

umgekehrt zu Reminiszenzphänomenen führen, d.h. zur erfolgreichen Reproduktion von Items, die in einem vorgeschalteten Testdurchgang nicht reproduziert wurden.

Anwendungen des Spurenintegritätsmodells auf Reproduktionsdaten von Kindern zeigen, daß Altersunterschiede primär in der Spurenzerfallswahrscheinlichkeit, weniger in den Abruf- und Spurenreintegrationsparametern bestehen. Auch Effekte nachträglicher Falschinformationen (vgl. Kapitel 10) schlagen sich — sofern sie bei Kindern überhaupt auftreten — ebenfalls in erster Linie in den Spurenzerfallswahrscheinlichkeiten nieder (Howe, 1991), sind allerdings auch in diesem Parameter nur schwach ausgeprägt (Marche & Howe, 1995). Leichter nachweisbar sind klassische retroaktive Hemmungseffekte, die bei Kindern ebenfalls vorwiegend den Spurenzerfallsparameter betreffen (Howe, 1995). Im Gegensatz dazu reduzieren konsistente Nachinformationen bei Vorschulkindern Spurenzerfallswahrscheinlichkeiten und fördern zusätzlich Abrufprozesse (Howe, Courage & Bryant-Brown, 1993). Ein kürzlich von Howe und O´Sullivan (1997) vorgelegtes Sammelreferat über Anwendungen des Spurenintegritätsmodells auf Kinder und Erwachsene zeigt, daß „Vergessen" in dem Paradigma, auf das sich das Modell bezieht, in erster Linie durch „storage failure" und weniger durch „retrieval failure" hervorgerufen wird, während Reminiszenzphänomene eher durch Abruflernen (retrieval recovery) als durch Spurenreintegration (trace redintegration) bedingt sind. Altersunterschiede schlagen sich nicht nur bei Kindern, sondern auch bei Erwachsenen in erster Linie im Spurenzerfallsparameter nieder: Er ist in allen Versuchsbedingungen bei 71-Jährigen höher als bei 18-Jährigen und bei 7- bis 8-Jährigen höher als bei 10- bis 11-jährigen Kindern. Die anderen Modellparameter zeigen eher Unterschiede zwischen verschiedenen Versuchsbedingungen als zwischen verschiedenen Altersgruppen. Soweit Alterseffekte auftreten, zeigen sich jedoch auch hier Vorteile Jugendlicher und jüngerer Erwachsener gegenüber älteren Erwachsenen und Kindern im Grundschulalter (Howe & O´Sullivan, 1997).

Bender, Wallsten und Ornstein (1996) haben das in Abschnitt 5.1 bereits besprochene Modell von Greeno et al. (1978) zur proaktiven Hemmung als Speicher-Abruf-Modell eingesetzt, indem sie es nicht auf Paar-Assoziations-Lernen von A-B- und A-C-Itempaaren, sondern auf zwei zeitlich versetzte Reproduktionstests für Alltagsereignisse anwendeten. Kinder sollte die Ereignisse, die während eines Kinderarztbesuches passiert sind, direkt nach dem Arztbesuch und dann noch einmal einen Tag später wiedergeben. Jedes kritische Ereignis kann für jeden der beiden Zeitpunkte als erfolgreich oder nicht erfolgreich reproduziert klassifiziert werden, so daß vier Datenkategorien zur Verfügung stehen. Die Idee des Modells ist, daß (a) erfolgreiche Reproduktion die Speicherung und den Abruf des Ereignisses voraussetzt und (b) der Abruf, gegeben eine erfolgreiche Speicherung, bei beiden Reproduktionstests konditional unabhängig erfolgt, wenn auch nicht not-

wendig mit gleichen Abrufwahrscheinlichkeiten. Es zeigte sich, daß ältere Kinder verglichen mit jüngeren sowohl bessere Speicher- als auch bessere Abrufparameter aufwiesen. Im Gegensatz zu den oben erwähnten Befunden von Howe, Wilkinson und anderen scheinen in *diesem* Paradigma und für *dieses* spezielle Gedächtnismaterial (Alltagsereignisse) Abrufwahrscheinlichkeiten über die Altersgruppen hinweg allerdings stärker anzusteigen als Speicherwahrscheinlichkeit.

Schließlich haben Riefer und Rouder (1992) ein weiteres Speicher-Abruf-Modell für PAL-Experimente vorgestellt, das ähnlich wie die Modelle von Chechile (1987), Riefer und Batchelder (1995) und Bender et al. (1996) auf dem Kontrast zweier Gedächtnistests basiert. In der Lernphase des PAL-Experiments werden Wortpaare dargeboten. In der Abrufphase wird zunächst ein freier Reproduktionstest für alle Wörter und danach ein geförderter Reproduktionstest für das zweite Wort eines Paares zu vorgegebenem ersten Wort durchgeführt (free then cued recall). Riefer und Rouder (1992) nahmen an, daß die freie Reproduktion eines kompletten Wortpaares sowohl Speicher- als auch Abrufprozesse reflektiert, während einzelne Wörter eines Paares auch ohne Wortpaarspeicherung mit Wahrscheinlichkeit s abgerufen werden können. Ist ein Wortpaar gespeichert (mit Wahrscheinlichkeit a), dann erfolgt der freie Abruf des kompletten Paares und der geförderte Abruf des zweiten Wortes unabhängig voneinander mit den Erfolgswahrscheinlichkeiten r_1 bzw. r_2, wobei natürlich im Regelfall $r_1 < r_2$ gelten sollte.[88]

Das Modell von Riefer und Rouder (1992) wurde bislang nur zur Klärung der Frage eingesetzt, inwieweit der häufig, aber nicht immer zu beobachtende Gedächtnisvorteil bizarrer Informationen im Vergleich normalen Informationen speicher- oder abrufbedingt ist. Es zeigt sich, daß bizarre Informationen besser abgerufen werden als normale Informationen, hinsichtlich der Speicherung jedoch keinen Vorteil (Riefer & Rouder, 1992) bzw. sogar einen Nachteil gegenüber gewöhnlichen Informationen aufweisen (Riefer & LaMay, 1998).

Rouder und Batchelder (1998) haben kürzlich eine sehr allgemeine hierarchische MVB-Modellklasse zum gleichen Paradigma („free then cued recall") vorgelegt. Obwohl ihre Modellklasse viele konzeptuelle Ähnlichkeiten zum Riefer-Rouder-Modell aufweist, beinhaltet sie letzteres Modell nicht als Spezialfall. Im Gegensatz zu Riefer und Rouder (1992) nehmen Rouder und Batchelder (1998) an, daß beide Wörter eines Paares — ähnlich wie im EA-Modell von Batchelder und Riefer (1980; 1986) — prinzipiell auch als Einzelwörter unabhängig voneinander erfolgreich abgerufen werden können, d.h. selbst dann, wenn keine Wortpaarspeicherung vorliegt. Darüber hinaus wurden mehrere Gleichsetzungsrestriktionen aufgegeben, die beim Riefer-Rouder-Modell Identifizierbarkeit und Testbarkeit si-

[88] Ausnahmen sind denkbar, wenn irreführende an Stelle förderlicher „Abrufhilfen" für das zweite Wort dargeboten werden.

cherstellen. Insgesamt gelingt Rouder und Batchelder (1998) mit ihrer flexibleren Modellklasse eine etwas bessere Anpassung der Daten von Riefer und Rouder (1992). Die substanzwissenschaftlichen Schlußfolgerungen aus ihren Ergebnissen decken sich jedoch mit den von Riefer und Rouder (1992).

7.1.5 Retroaktive und proaktive Hemmung

Das EA-Modell kann in Verbindung mit freier Reproduktion von Itempaaren und Einzelitems auch zur Untersuchung von proaktiven und retroaktiven Hemmungseffekten eingesetzt werden. Man muß hierzu lediglich vor bzw. nach der zu reproduzierenden Zielliste zusätzliche Itemlisten darbieten und dann untersuchen, wie sich Art bzw. Anzahl dieser interferierenden Listen auf die Parameter des EA-Modells auswirken. Für die retroaktive Hemmung ist dies bereits recht ausgiebig geschehen (Bäuml, 1991a; 1991b; 1991c; Riefer & Batchelder, 1988; vgl. Abschnitt 3.5.1), für die proaktive Hemmung interessanterweise bislang noch nicht. Dies ist sicherlich ein naheliegendes Forschungsfeld für die Zukunft.

Einen alternativen Zugang zum Phänomen der proaktiven Hemmung ermöglicht das Modell von Greeno et al. (1978), welches in Abschnitt 5.1 und im letzten Abschnitt bereits erwähnt wurde. Es bezieht sich speziell auf proaktive Hemmungsprozesse beim Paar-Assoziations-Lernen (PAL): Zunächst ist eine (interferierende) Itemliste mit A-B-Paaren und dann eine Itemliste mit A-C-Paaren gelernt wird, bevor die Vpn auf einen A-Hinweisreiz sowohl das B- als auch C-Zielitem nennen sollen. Die vier möglichen Reproduktionsresultate bezüglich beider Zielitems bilden die Datenbasis des Modells. Grundlegend für das Modell ist die Annahme, daß beide Zielitems entweder gemeinsam mit dem zugehörigen A-Item abgespeichert werden (mit Wahrscheinlichkeit s) oder nicht. Im letzten Falle kann weder das B- noch das C-Zielitem abgerufen werden, im erstgenannten Falle werden sie dagegen unabhängig voneinander mit den Wahrscheinlichkeiten q und r abgerufen.

Empirisch stellte sich heraus, daß die Anzahl der A-B-Darbietungen vor der A-C-Liste zwar die Abrufwahrscheinlichkeit q für B-Items erhöht, die Abrufwahrscheinlichkeit r für C-Items jedoch untangiert läßt (Greeno et al., 1978, Kapitel 7; Riefer & Batchelder, 1988, S. 327f). Dies spricht gegen die Interferenztheorie proaktiver Hemmung (Postman & Underwood, 1973), wonach eine Stärkung der A-B-Assoziation eine Schwächung der A-C-Assoziation impliziert. Es stützt dagegen die Vorhersagen der Theorie unabhängiger Abrufprozesse, da eine experimentelle Manipulation von q offenbar keine Auswirkungen auf r hat (Greeno et al., 1978)

Diese Anwendung der MVB-Modellierung ist exemplarisch für *hypothesentestende Anwendungen* der MVB-Modellierung: Ein auf die Daten passendes und als psychologisch valide akzeptiertes Modell wird benutzt, um psychologische Hy-

pothesen zu den Ursachen bestimmter empirischer Phänomene (hier: proaktive Hemmung) zu testen. Die Anwendung des Modells auf die Daten von Greeno et al. (1978) bestätigen die Theorie unabhängiger Abrufprozesse und stehen im Widerspruch zu einer Interferenztheorie der proaktiven Hemmung.

7.1.6 Schemata und Skripts

Ross und Bower (1981) prüften verschiedene Modelle der Repräsentation schemabasierter Informationen. Sie ließen ihre Vpn vier oder fünf Wörter lernen, die über ein gemeinsames Schema verknüpft sind (z.B. Taxi, Bahn, Schalter, Fahrkarte). Anschließend wurde getestet, wieviel Wörter einer Gruppe reproduziert werden können, wenn man ein, zwei oder drei Wörter einer Gruppe zuvor als Abrufhilfen darbietet. Jedes der n nicht dargebotenen Wörter in einer Versuchsbedingung kann als erfolgreich reproduziert bzw. nicht reproduziert klassifiziert werden, wodurch sich allein für die betreffende Bedingung 2^n Antwortkategorien ergeben. Betrachtet man alle Versuchsbedingungen simultan, ergeben sich also hinreichend viele Beobachtungskategorien.

Ross und Bower haben drei binäre Zwei-Parameter-MVB-Modelle konstruiert, die als Formalisierungen dreier verschiedener Modelle des assoziativen Gedächtnisses aufgefaßt werden können. Eines dieser Modelle konnte die Häufigkeiten erfolgreicher Reproduktionen über drei Experimente hinweg eindeutig am besten erklären: das Schema-Modell. Nach diesem Modell sind thematisch assoziierte Informationen nicht komplett vernetzt repräsentiert (wie im „horizontalen Modell"), sondern nur indirekt über das gemeinsame Schema verknüpft. Direkte Assoziationen bestehen nach dem Schema-Modell zwar jeweils zwischen den Items und dem zugehörigen Schema, nicht jedoch zwischen thematisch verwandten Informationen untereinander. In der Abrufsituation ermöglicht eine Abrufhilfe mit Wahrscheinlichkeit a Zugang zum richtigen Schema, was mit Wahrscheinlichkeit r zum Abruf einer mit dem Schema assoziierten Information führt. Erwartungsgemäß lag nicht nur die Modellanpassungsstatistik in allen drei Experimenten im akzeptablen Bereich; auch die Parameterschätzungen unterschieden sich zwischen Vpn, die über das gemeinsame Thema zuvor informiert wurden (mittlere Schätzung für a = .618, für r = .680), und nichtinformierten Kontrollpersonen (mittlere Schätzung für a = .531, für r = .654) in der erwarteten Richtung (Experiment 1).

Die Arbeit von Ross und Bower (1981) ist ein Beispiel für die *theorievergleichende Anwendung* von MVB-Modellen: Es werden gezielt mehrere Modelle konstruiert, die als Formalisierungen verschiedener psychologischer Theorien aufgefaßt werden können. Von vornherein besteht die Erwartung, daß nicht alle Modelle gut auf die Daten passen können: Es geht gerade darum, das bestpassende

Modell herauszufinden und schlechtpassende Modelle — und damit die dahinterstehenden psychologischen Ideen — auszusondern. Die Parameterschätzungen bzw. Parametervergleiche zwischen verschiedenen experimentellen Bedingungen sind von untergeordneter Bedeutung. Im Vordergrund steht eindeutig die Modellanpassungsstatistik.

Die Problematik dieses Vorgehens liegt auf der Hand: Überzeugend ist die Widerlegung einer psychologischen Hypothese nur dann, wenn keinerlei Zweifel daran bestehen kann, daß das zugehörige Modell die Theorie richtig formalisiert. Dies ist bei verbal formulierten psychologischen Hypothesen aber fast immer debattierbar. In die Modellkonstruktion gehen in der Regel technische Zusatzannahmen ein, die der eigentliche Grund für schlechte Modellanpassungsergebnisse sein können. Man wird ein psychologische Hypothese aber nur dann empirisch überzeugend „belasten" können, wenn man sie anhand eines Paradigmas prüft, das minimale technische Zusatzannahmen erforderlich macht. Ob das — wie vom theorievergleichenden Modellierungsansatz gefordert — für *verschiedene* psychologische Hypothesen das *gleiche* Paradigma sein kann, ist im allgemeinen fraglich.

Erdfelder und Bredenkamp (1998) haben Submodelle des 2HTSM-Quellendiskriminationsmodells (vgl. Abschnitt 7.1.2) zur Prüfung zweier Theorien der Speicherung und des Abrufs skriptbasierter Informationen eingesetzt. Beide Theorien zielen auf eine Erklärung des „Atypikalitätseffekts" im Rekognitionsgedächtnis ab. Damit ist das bessere Itemgedächtnis für skript-atypische Informationen im Vergleich zu skript-typischen Informationen gemeint. Wenn beispielsweise eine Geschichte über einen Restaurantbesuch präsentiert wird, rekognizieren Vpn später atypische Ereignisse („Ein Haar war in der Suppe") wesentlich besser als typische Ereignisse („Der Kellner brachte die Karte"). Die Aufmerksamkeitselaborationshypothese (AEH) erklärt dies mit der tieferen itemspezifischen Verarbeitung atypischer Ereignisse in der Enkodierungsphase. Die Script-Copy-Plus-Tag-Hypothese (SCTH) nimmt demgegenüber an, daß in der Enkodierungsphase eine Kopie des kompletten Skripts (im Beispiel: „Restaurantbesuch") inklusive der assoziierten hochtypischen Ereignisse gespeichert wird. Atypische Ereignisse werden demgegenüber als zusätzliche „tags" an der Skript-Kopie markiert, welche den späteren Abruf atypischer Informationen erleichtern (zusammenfassend: Graesser & Nakamura, 1982).

Erdfelder und Bredenkamp (1998) boten skript-typische und skript-atypische Ereignisse entweder in vollständiger Form oder als Lückenwörter dar. Beide unabhängigen Variablen wurden innerhalb einer Geschichte orthogonal zueinander variiert. Nach der AEH sollten Lückenwörter ebenfalls intensive itemspezifische Verarbeitungsprozesse anregen und deshalb den Rekognitionsvorteil atypischer Information nivellieren oder gar eliminieren: Der aufmerksamkeitserhöhende Effekt atypischer Ereignisse sollte bei Lückenwortdarbietung also redundant werden.

Nach der SCTH sollte der Effekt der Lückenwortdarbietung dagegen im wesentlichen additiv sein: Die Lückenwortdarbietung mag zwar einen generell förderlichen Effekt auf das Itemgedächtnis haben, welcher jedoch an dem Rekognitionsvorteil der durch „tags" markierten atypischen Ereignisse nichts ändert.

Erdfelder und Bredenkamp (1998) fanden tatsächlich eine Nivellierung (Experiment 1) bzw. sogar eine Eliminierung des Itemgedächtnisvorteils atypischer Informationen, wenn zwei Sätze mit Lückenwörtern direkt aufeinander folgen (Experiment 2). Der geschätzte Quellengedächtnisparameter *d* für die Entscheidung, ob das Item ursprünglich vollständig oder als Lückenwort dargeboten wurde, lag im mittleren Bereich, was ebenfalls eher mit der AEH als mit der SCTH vereinbar ist (vgl. dazu Erdfelder & Bredenkamp, 1998). Die bei skript-typischen Items höhere Antworttendenz *b* zugunsten von „Alt"-Antworten läßt sich prinzipiell mit der AEH und auch mit der SCTH erklären. Obwohl Möglichkeiten bestehen, die SCTH so zu modifizieren, daß sie mit dem kompletten Ergebnismuster beider Experimente konsistent wird, lassen sich die Ergebnisse von Erdfelder und Bredenkamp (1998) insgesamt zwangloser im Rahmen der AEH interpretieren.

Bayen et al. (2000) haben ebenfalls das 2HTSM-Quellengedächtnismodell verwendet, um Itemgedächtnis, Quellengedächtnis und Ratetendenzen für typische und atypische Details verbal präsentierter Szenen zu untersuchen. Sie präsentierten zunächst zwei Szenen A und B (A = Schlafzimmer , B = Badezimmer), wobei einige Items jeder Szene für die Szene A eher typisch und für die Szene B atypisch waren (z.B. Bücher), andere dagegen für Szene A eher atypisch und für Szene B typisch waren (z.B. Haarwaschmittel). Die Vpn mußte anschließend für alte und neue Items angeben, ob sie ursprünglich in Szene A bzw. Szene B dargeboten wurden oder neu sind.

Anders als bei Erdfelder und Bredenkamp (1998) ergab sich in Experiment 1 von Bayen et al. kein Atypikalitätsvorteil im Itemgedächtnis. Noch überraschender im Lichte der AEH und SCTH ist vielleicht, daß auch der Quellengedächtnisparameter für typische und atypische Items nicht signifikant unterschiedlich ausfiel. Dies kann prinzipiell auf zu geringer Teststärke des χ^2-Differenzentests beruhen. Deskriptiv gehen die Parameterschätzer für die Quellengedächtnisparameter nämlich in die richtige Richtung, d.h. das Quellengedächtnis scheint für atypische Informationen geringfügig besser zu sein als für typische Informationen. Diese Erklärung kann jedoch nicht voll überzeugen, da der Atypikalitätseffekt gewöhnlich sehr stark ist; ein vergleichbar starker Effekt hätte in der vorliegenden Untersuchung sicherlich zu einer signifikanten Teststatistik führen müssen. Plausibler erscheint die Erklärung, daß starke Schema-Effekte auf die Gedächtnisparameter deshalb ausblieben, weil die beiden präsentierten Szenen jeweils zu 50% aus atypischen Items bestanden. Keine der beiden Szenen hat also möglicherweise die zugehörigen Schemata voll aktiviert. Begünstigend kommt hinzu, daß der Gedächt-

nistest — anders als bei Erdfelder und Bredenkamp (1998) — zuvor angekündigt wurde. Vielleicht haben die Vpn also in der Enkodierungsphase gezielt Itemlisten gelernt statt die Items schemabasiert verarbeitet. Es wäre folglich interessant, Experiment 1 von Bayen et al. unter inzidentellen Lernbedingungen mit nur wenigen atypischen Items pro Szene zu replizieren. Man könnte beispielsweise die Vpn instruieren, sich einen Eindruck von einer Person zu bilden, deren Badezimmer und Schlafzimmer durch bestimmte Einrichtungsgegenstände gekennzeichnet ist. Sollten Atypikalitätseffekte auf die Item- und Quellengedächtnisparameter des 2HTSM-Modells auch unter diesen Bedingungen ausbleiben, so wäre das sowohl für die AEH als auch für die SCTH ein kritischer Befund.

Im Einklang mit den Ergebnissen von Erdfelder und Bredenkamp (1998) steht der Befund von Bayen et al., daß Schema-Typikalität von Testitems Ratetendenzen affiziert, und zwar unabhängig davon, ob die Schema-Manipulation bereits während der Enkodierung (Experiment 1) oder erst kurz vor dem Abruf (Experiment 2) stattfindet. Items, die beispielsweise für Quelle A typisch sind, werden also im Zustand der Quellenunsicherheit auch eher als aus Quelle A stammend geraten.

Als vorläufiges Fazit bleibt festzuhalten, daß Schema-Effekte auf Ratetendenzen in Gedächtnistests inzwischen gut gesichert sind, während Schema-Effekte auf das Item- und das Quellengedächtnis weiterer experimenteller Klärung bedürfen.

7.1.7 Unmittelbare Reproduktion

Schweikert (1993) hat ein sehr einfaches MVB-Modell der unmittelbaren Reproduktion aus dem Arbeitsgedächtnis vorgeschlagen: Die Gedächtnisspur eines Items ist beim Abrufversuch mit Wahrscheinlichkeit i intakt und führt zur sofortigen Reproduktion. Wenn sie mit Wahrscheinlichkeit $(1 - i)$ nicht intakt ist, kann sie mit der bedingten Wahrscheinlichkeit r rekonstruiert werden, was ebenfalls zu einer erfolgreichen Reproduktion führt. Andernfalls scheitert der Abruf. Die Reproduktionswahrscheinlichkeit ist folglich $p = i + (1–i)\cdot r$, was bei nur zwei Kategorien (erfolgreiche vs. nicht erfolgreiche Reproduktion) keine identifizierbare Modellgleichung sein kann. Durch selektive Manipulation der Modellparameter in einem komplett gekreuzten mehrfaktoriellen Versuchsplan kann jedoch Identifizierbarkeit hergestellt werden (vgl. Abschnitt 4.6).

Schweikert (1993) schlug vor, den Parameter i z.B. über die Länge eines Wortes zu manipulieren: Je länger ein Wort, desto länger ist notwendigerweise das Retentionsintervall bis zur Reproduktion und desto kleiner wird folglich i. Den Parameter r kann man seiner Meinung nach über die Lexikalität eines Items beeinflussen: Pseudohomophone, d.h. Nonwörter, die phonetisch Wörtern entsprechen

(z.B. „churtch" anstelle von „church"), sollten leichter rekonstruierbar sein als andere leicht aussprechbare Nonwörter (z.B. „chitch"). Reanalysen publizierter Datensätze bestätigten die Interaktionshypothese, die aus der o.g. Modellgleichung folgt: Der Effekt der Wortlänge auf die Reproduktionsraten ist bei Pseudohomophonen geringer als bei anderen Nonwörtern (Schweikert, 1993, S. 171).

Die Interpretation der Befunde in Termini von intakten und rekonstruierten Gedächtnisspuren ist nicht zwingend. Eine Modellgleichung, die der von Schweikert (1993) äquivalent ist, liegt nämlich auch der Theorie von Waugh und Norman (1965) zugrunde. In ihrer Theorie stehen die Parameter für die Abrufwahrscheinlichkeit aus dem primären (*i*) und aus dem sekundären Gedächtnis (*r*). Während *i* also in beiden Modellen eine sehr ähnliche Interpretation erfährt (Abruf aus dem Arbeitsgedächtnis), unterscheiden sich die Interpretationen für *r* wesentlich. Der oben geschilderte Befund — höhere *r*-Parameter bei Pseudohomophonen — ist auch aus dem Modell von Waugh und Norman ableitbar, wenn man annimmt, daß die Lexikalität eines Items die Speicherung und den Abruf aus dem sekundären Gedächtnis (d.h. dem episodischen Langzeitgedächtnis) begünstigt — eine sehr plausible Annahme, wie mir scheint. Welche Interpretation des *r*-Parameters richtig ist, kann nur nach umfangreichen Validierungsuntersuchungen entschieden werden (vgl. Abschnitt 4.5). Bislang liegt hierzu noch keine ausreichende Datenbasis vor.

Ungeachtet dieses Problems haben Hulme, Roodenrys und Schweikert (1997) das Modell benutzt, um die Determinanten des Worthäufigkeitseffekts im Kurzzeitgedächtnis zu untersuchen. Häufige Worte scheinen demnach besser unmittelbar reproduzierbar zu sein, weil der Rekonstruktionsprozeß bei zerfallener bzw. beschädigter Gedächtnisspur — erfaßt durch Parameter *r* — dort besser funktioniert als bei seltenen Wörtern. Natürlich ist auch hier eine Alternativinterpretation im Sinne des Modells von Waugh und Norman (1965) denkbar; demnach wären häufige Wörter besser bzw. schneller langfristig speicherbar als seltene Wörter.

7.1.8 Scheitern der Rekognition (Recognition Failure)

Wenn semantisch plausible Wortpaare gelernt werden (z.B. „Wissenschaftler: Darwin") und später ein Rekognitionstests für das zweite Wort (Wurde „Darwin" dargeboten?) gefolgt von einem geförderten Reproduktionstest für das zweite Wort durchgeführt wird (Was folgte auf „Wissenschaftler"?), so zeigt sich nicht selten das „Recognition-Failure-Phänomen": Ein Wort, das nicht rekognizierbar ist, kann u.U. dennoch gefördert reproduzierbar sein. Die Beziehung zwischen der bedingten Rekognitionswahrscheinlichkeit, gegeben erfolgreiche Reproduktion, und der unbedingten Rekognitionswahrscheinlichkeit wurde von Tulving und Wiseman (1975) quantitativ beschrieben und hat seitdem zahlreiche Erklärungs-

versuche angeregt, darunter auch solche, die in Termini von MVB-Modellen formuliert bzw. mittels solcher Modelle geprüft wurden.

Die Tulving-Wiseman-Funktion scheint zunächst einmal im Widerspruch zu den Rekognitionsmodellen von Mandler (1980) sowie Humphreys und Bain (1983) zu stehen, nach denen die Reproduktionswahrscheinlichkeit eine Untergrenze für die Rekognitionswahrscheinlichkeit ist (vgl. Abschnitt 7.1.3). Man beachte jedoch, daß in den Untersuchungen, auf die Mandler und Humphreys Bezug nehmen, Reproduktion und Rekognition *zwischen* Vpn und nicht — wie im Falle der Tulving-Wiseman-Funktion — innerhalb der Vpn variiert werden. Humphreys und Bowyer (1980) haben deshalb argumentiert, daß die Tulving-Wiseman-Funktion kein geeignetes Mittel ist, um Theorien zur Beziehung zwischen Reproduktion und Rekognition zu überprüfen. Ihrer Meinung nach ist diese Funktion in erster Linie eine Konsequenz des „Priming-Effekts" des Rekognitionstests, d.h. der abruffördernden Wirkung des Rekognitionstests auf die anschließende geförderte Reproduktion desselben Zielwortes. Sie formulierten Ihre Priming-Hypothese als binäres MVB-Modell, das (1) die Reproduktionswahrscheinlichkeit, (2) die unbedingte Rekognitionswahrscheinlichkeit und (3) die bedingte Rekognitionswahrscheinlichkeit, gegeben eine erfolgreiche Reproduktion, als Funktionen von vier Parametern ausdrückt: der Wahrscheinlichkeit einer erfolgreichen Reproduktion *ohne* Primingeffekt (a), zweier bedingter Wahrscheinlichkeiten für die erfolgreiche Rekognition mit (p) und ohne zusätzliche Reproduktion (q) sowie der Wahrscheinlichkeit eines Primingeffekts im Reproduktionstest als Folge des Rekognitionstests (x). Durch Gleichsetzung der beiden bedingten Rekognitionswahrscheinlichkeiten p und q — was der Annahme unabhängiger Abrufprozesse im Reproduktions- und im Rekognitionsprozeß entspricht — kann man dieses Modell in ein identifizierbares Modell überführen.

Riefer und Batchelder (1995) haben ebenfalls die Tulving-Wiseman-Funktion kritisiert, allerdings aus anderer Perspektive. Sie zeigten, daß zwei mathematisch äquivalente Formulierungen des Tulving-Wiseman-„Gesetzes" in der einen Form — der Standardform — sehr gute, in der anderen Form sehr schlechte Anpassungen an die gleichen Datensätze liefern. Es gibt also keinen Grund, eine Theorie so zu konstruieren, daß sie *exakt* die Tulving-Wiseman-Funktion vorhersagt. Es scheint sich mehr um eine empirische Generalisierung zu handeln, der eine adäquate Theorie zwar deskriptiv nahe kommen jedoch nicht exakt entsprechen muß.

Riefer und Batchelder (1995) haben das bereits in Abschnitt 4.6 geschilderte binäre MVB-Modell zur Untersuchung von Speicherungs- und Abrufprozessen beim „Recognition-Failure"-Phänomen eingesetzt. Formal ist dieses Modell dem in den Abschnitten 5.1 und 7.1.4 besprochenen Modell von Greeno et al. (1978) zur Untersuchung der Determinanten proaktiver Hemmung äquivalent. Die Idee die-

ses Modells ist, daß ein Wortpaar mit einer bestimmten Wahrscheinlichkeit gespeichert oder aber nicht gespeichert wird. Nur wenn es gespeichert wird, kann es im Rekognitionstest oder im anschließenden geförderten Reproduktionstest abgerufen werden, wobei der Abruf bei beiden Tests stochastisch unabhängig erfolgt. Ein zusätzlicher Primingeffekt — wie von Humphreys und Bowyer (1980) postuliert — wird nicht angenommen. Dennoch kommt dieses Modell den Vorhersagen der Tulving-Wiseman-Funktion sehr nahe, wenn man annimmt, daß Speicherungs- und Abrufwahrscheinlichkeiten typischerweise über experimentelle Bedingungen hinweg korreliert sind. Zahlreiche Datensätze zum Recognition-Failure-Paradigma, die Nilsson und Gardiner (1993) in ihrem Sammelreferat zusammengefaßt haben, werden durch das Modell gut angepaßt; zudem variieren die Parameter mit den experimentellen Bedingungen so, wie man es von Speicher- bzw. Abrufparametern erwarten sollte (vgl. Riefer & Batchelder, 1995, Appendix). Insgesamt stützt Riefer und Batchelders (1995) modellbasierte Analyse zwar die Theorie unabhängiger Abrufprozesse bei Reproduktion und Rekognition, die Flexser und Tulving (1978) zur Erklärung der Tulving-Wiseman-Funktion vorgeschlagen haben. Die Funktion selbst wird durch modellbasierte Analyse allerdings nicht gestützt. Sie scheint nicht mehr als eine empirische Generalisierung — eine grobe Deskription der Datenstruktur in einem bestimmten Diagramm — zu sein; Gesetzesstatus hat diese Funktion offenbar nicht.

Damit wird eine weitere Verwendungsform multinomialer Modelle in der kognitiven Psychologie deutlich: die *ableitungskritische Verwendung*. Dadurch, daß bestimmte psychologische Annahmen formalisiert und in einem Modell zusammengefaßt werden — beispielsweise die Annahme unabhängiger Abrufprozesse — wird formal entscheidbar, welche Theoreme oder „Gesetze" aus diesen Annahmen ableitbar sind und welche nicht. Dies schafft Klarheit über den Status von empirischen Generalisierungen: Sind sie lediglich mehr oder minder gute Datendeskriptionen oder haben sie den Status eines theoretisch erklärbaren empirischen Gesetzes? Ohne modellbasierte Analysen sind solche Fragen kaum entscheidbar.

7.1.9 Gedächtnistäuschungen

Gedächtnistäuschungen (zusammenfassend: Hell, 1993) stellen eine besondere Herausforderung für die MVB-Modellierung dar. Ein statistisch adäquates und psychologisches validiertes MVB-Modell zu einer Gedächtnistäuschung ist entweder eine Erklärung der Täuschung (im Sinne einer theoretisch-statistischen Erklärung, vgl. Groeben & Westmeyer, 1981, S. 80ff) oder zumindest ein geeigneter Rahmen, um verschiedene Erklärungen gegeneinander zu testen. Insofern ist die Zielstellung, ein MVB-Modell zu einer Gedächtnistäuschung zu entwickeln, oft-

mals anspruchsvoller und schwieriger als die der Konstruktion von Meßmodellen zu bestimmten gedächtnispsychologischen Konstrukten: Die Anzahl der empirischen Randbedingungen, mit denen Modelle zu Gedächtnistäuschungen konsistent sein müssen, ist im Regelfall recht groß.

Drei Klassen von Gedächtnistäuschungen sind bislang mit MVB-Modellen untersucht worden: Falsche Erinnerungen (false memory effect), der Falschinformationseffekt (misinformation effect) und der Rückschaufehler (hindsight bias). Beim „false memory effect" werden Ereignisse bzw. Items reproduziert oder rekogniziert, die faktisch nicht passiert sind bzw. zuvor nicht dargeboten wurden. Reliabel induzieren kann man diesen Effekt, indem man thematisch hoch assoziierte bzw. über ein gemeinsames Schema verknüpfte Items darbietet, wobei einige besonders prototypische Items ausgelassen werden. Bietet man die prototypischen Distraktoren anschließend dar, liegen die Raten falschen Alarms typischerweise im Bereich der Trefferraten für alte Items oder sogar darüber (Brainerd & Reyna, 1998b). Durch Warnhinweise an die Vpn kann man den Effekt zwar etwas reduzieren, aber nicht eliminieren (Gallo, Roberts & Seamon, 1997). Erinnerungsberichte der Vpn, die mit der Remember-Know-Prozedur (vgl. Kapitel 9) erhoben wurden, lassen darauf schließen, daß die entsprechenden Distraktoren zu einem erheblichen Teil subjektiv bewußt erinnert werden, so als wären sie wirklich dargeboten worden (Payne, Elie, Blackwell & Neuschatz, 1996).

Eine Erklärung falscher Erinnerungen liefert das in Abschnitt 7.1.3 bereits geschilderte Conjoint-Recognition-Modell von Brainerd und Reyna. Wenn die „gist traces" des gemeinsamen Themas oder Schemas stark ausgeprägt sind, die „verbatim traces" der alten Items dagegen eher schwach, aktiviert eine thematisch passender Distraktor eher die „gist traces", was ein „Alt"-Urteil aufgrund einer Ähnlichkeitsidentifizierung wahrscheinlich macht (vgl. Abschnitt 7.1.3). Die Größe dieses Effekts kann durch Variablen beeinflußt werden, die differentielle Effekte auf die Stärke von „verbatim traces" und „gist traces" haben. Hierzu gehört beispielsweise die Länge des Retentionsintervalls: Wörtliche Repräsentationen zerfallen schneller als Bedeutungsrepräsentationen, so daß falsche Erinnerungen mit der Länge des Retentionsintervalls häufiger werden sollten. Eine weitere gut bestätigte Vorhersage ist der Persistenzeffekt: Falsche Alarme für thematisch passende Distraktoren können über Zeit stabiler sein als Treffer für alte Items, weil erstere auf stabilen „gist traces" basieren, letztere dagegen vorwiegend auf zeitlich instabilen „verbatim traces".

Der Falschinformationseffekt wurde bereits in Abschnitt 2.6 kurz angesprochen. Gemeint ist damit die Beeinträchtigung der Gedächtnisleistung für ein Zielitem (z.B. ein Detail einer betrachteten Szene) durch nachträglich dargebotenen Falschinformation zu diesem Detail. MVB-Modelle zum Falschinformationseffekt wurden bislang von Wagenaar und Boer (1987) und Batchelder, Riefer und Wal-

ton (1990) entwickelt. Da in Kapitel 10 auf diese Arbeiten und zusätzlich auf neue Modelle zum Falschinformationseffekt ausführlich eingegangen wird, kann eine Darstellung an dieser Stelle unterbleiben.

Einen dritten Schwerpunkt bildeten multinomiale Modelle zum Hindsight Bias, für den Pohl (1992) die deutsche Bezeichnung „Rückschaufehler" vorgeschlagen hat. Der Rückschaufehler wurde bislang im sogenannten hypothetischen Design und im Gedächtnisdesign untersucht (Hell, 1993; Hell, Gigerenzer, Gauggel, Mall & Müller, 1988; Hertwig, Gigerenzer & Hoffrage, 1997; Pohl, 1995). Typischerweise (aber nicht notwendigerweise) bekommen die Probanden schwierige Fragen zum Allgemeinwissen vorgelegt, auf die sie numerische Antworten geben sollen (z.B. „Wie lang ist der Rhein?", „In welchem Jahr wurde Friedrich der Große geboren?", „Wieviel Prozent der deutschen Bevölkerung hatten 1990 eine KfZ-Führerschein?" usw.). Im hypothetischen Design werden die Antworten auf derartige Fragen (Kontrollbedingung) mit einer Experimentalbedingung verglichen, in der die Probanden *vor* der Antwort auf die Frage die korrekte Antwort mitgeteilt bekommen („Der Rhein ist 1320 km lang. Was hätten Sie auf die Frage nach der Länge des Rheins geantwortet, wenn ich Ihnen diese Information nicht gegeben hätte?"). Im Gedächtnisdesign geben die Probanden zunächst Antworten auf die Fragen (Originalantwort, OA), bekommen nach einem Retentionsintervall in der Experimentalbedingung die korrekten Antworten (KA) mitgeteilt, in der Kontrollbedingung dagegen nicht. Abschließend sind in Experimental- und Kontrollbedingung die Originalantworten zu erinnern (Reproduktion oder Rekognition der Originalantwort, ROA). Das typische Ergebnis in beiden Designs ist, daß die Urteile bzw. Gedächtnisurteile (ROA) in der Experimentalbedingung verglichen mit der Kontrollbedingung in Richtung KA verzerrt sind. Dieser Effekt ist äußerst robust; auch explizite Warnungen der Vpn und Aufforderungen, den Effekt zu vermeiden, ändern an seiner Stärke gewöhnlich nichts (Fischhoff, 1982; Pohl & Hell, 1996).

Erdfelder (1992a; 1991) hat ein MVB-Modell zum Gedächtnisdesign des Rückschaufehlers vorgeschlagen, welches primär darauf abzielt, Gedächtnisbeeinträchtigungen durch Darbietung der korrekten Antwort und Rekonstruktionsverzerrungen im Falle der Nichtabrufbarkeit der Originalantwort separat und von Antwort- und Ratetendenzen dekontaminiert zu erfassen. Die empirische Basis des Modells sind Häufigkeiten, mit denen die verschiedenen möglichen Rangordnung von OA, KA und ROA in der Experimental- und der Kontrollbedingung beobachtet wurden. Angenommen wird, daß die Originalantwort in der Kontrollbedingung mit Wahrscheinlichkeit r_C und in der Experimentalbedingung mit Wahrscheinlichkeit r_E abgerufen werden kann. Ein Gedächtnisbeeinträchtigungseffekt der KA-Präsentation ist nachgewiesen, wenn die Hypothese $r_E < r_C$ statistisch angenommen werden kann. Kann die OA nicht abgerufen werden, so wird

sie rekonstruiert, und zwar in der Kontrollbedingung zwangsläufig unbeeinflußt von der Kenntnis der KA. In der Experimentalbedingung kann die Originalantwort dagegen mit Wahrscheinlichkeit *b* KA-beeinflußt rekonstruiert werden, was zu einer Quellenverwechslung (ROA = KA) oder zumindest zu einer Verzerrung des ROA-Urteils in Richtung KA führt (OA ≈ ROA). Mit der komplementären Wahrscheinlichkeit 1–*b* erfolgt die Rekonstruktion KA-unbeeinflußt, d.h. in der gleichen Weise wie in der Kontrollbedingung. Eine Rekonstruktionsverzerrung ist folglich nachgewiesen, wenn die Hypothese $b > 0$ statistisch angenommen werden kann. Das Modell läßt prinzipiell unterschiedliche Verteilungen rekonstruierter ROAs in Abhängigkeit von überschätzten (OA > KA) oder unterschätzten Originalantworten (OA < KA) zu. Submodelle des kompletten 13-Parameter-Modells HB13 existieren, die strengere Symmetrieannahme über die ROA-Verteilungen machen. Theoretisch folgt es einem eklektizistischen Modell des Rückschaufehlers von Hawkins und Hastie (1990), das sowohl Gedächtnisbeeinträchtigungen durch KA-Präsentation (Fischhoff, 1982; Hell, 1993; Hell et al., 1988) als auch Urteilsprozesse bei Nichtabrufbarkeit der OA (Stahlberg & Eller, 1993) als potentielle Ursachen in Erwägung zieht.

Ein systematische Validierung des HB13-Modells wurde von Erdfelder und Buchner (1998a) durchgeführt. In Experiment 1 wurden Items kontrastiert, deren OAs kurz nach der Abgabe des Ersturteils abrufbar waren bzw. nicht abrufbar waren. Erwartungsgemäß wirkte sich dies selektiv auf die r_C- und r_E-Parameter beim späteren ROA-Test aus. Die Verkürzung des Zeitintervalls zwischen KA-Rückmeldung und ROA-Urteil erhöhte selektiv den Rekonstruktionsbias *b* (Experiment 2), die Diskreditierung der KA („Die soeben rückgemeldeten korrekten Antworten waren leider alle fehlerhaft") erhöhte r_E und reduzierte *b*, letzteres allerdings nicht signifikant. Die Manipulation der ROA-Antwortalternativen bei einem *n*-Alternativen-Forced-Choice-Rekognitionstest affizierte ausschließlich den Rateparameter des HB13-Modells (Experiment 4). Insgesamt gab es in diesen Experimenten wenig Hinweise auf Gedächtnisbeeinträchtigungseffekte (die Hypothese $r_C = r_E$ konnte nur in einer von neun untersuchten Versuchsbedingungen zurückgewiesen werden), hingegen deutliche Hinweise auf die Bedeutung von Rekonstruktionsverzerrungen beim Rückschaufehler ($b > 0$).

Das HB13-Modell wurde von Ruoß (1997) benutzt, um kognitive Täuschungen bei chronischen Schmerzpatienten zu untersuchen. Diese klinische Gruppe weist im Vergleich zu psychiatrischen und studentischen Kontrollpersonen stark erhöhte Rekonstruktionsverzerrungen auf (erfaßt durch den *b*-Parameter). Hinzu kommt ein generell niedrigeres Niveau der Erinnerungsparameter r_C und r_E, so daß insgesamt eine sehr starke Tendenz der Angleichung von ROAs and die KAs zu verzeichnen ist. Ruoß (1997) sieht in der Kombination von mangelnder Erinnerung an vergangene Ereignisse bei gleichzeitiger Angleichung der Erinnerungs-

urteile an aktuell präsente Anker einen möglichen kausalen Mechanismus bei der Chronifizierung von Schmerzen. Dieser kognitive Mechanismus erschwert das Entdecken von Befindlichkeitsänderungen sowohl in negative als auch in positive Richtung; er begünstigt somit das Gefühl, die Schmerzen in der Vergangenheit schon immer so wie in der Gegenwart erlebt zu haben.

Saur (1998) hat untersucht, ob sich Rekonstruktionsverzerrungen durch eine hypnotherapeutisch begründete Imaginationsbedingung reduzieren lassen. Probanden (Studierende des Faches Psychologie) mußten sich hierzu kurz vor der ROA imaginativ erneut in den OA-Enkodierungskontext hineinversetzen. Obwohl die *b*-Parameterschätzung in der Imaginationsbedingung im Vergleich zu einer Standardbedingung leicht reduziert war (0.26 statt 0.30; vgl. Saur, 1998, S. 89), war dieser Effekt nicht signifikant, so daß nicht von einer Beeinflußbarkeit der Rekonstruktionsverzerrungen durch Imaginationstechniken ausgegangen werden kann.

Bayen, Erdfelder und Bearden (1998) haben ältere Menschen und jüngere Erwachsene auf der Grundlage des HB13-Modells verglichen und erwartungsgemäß niedrigere Gedächtnisparameter bei älteren Erwachsenen gefunden. Interessanter als dieser Effekt war der wesentlich stärkere Rekonstruktionsbias *b* bei älteren Menschen im Vergleich zu jüngeren, der dann zu beobachten ist, wenn die KA zum Zeitpunkt des ROA-Urteils visuell präsent ist. Anders als bei chronischen Schmerzpatienten persistiert dieser Effekt jedoch nicht, wenn ein Zeitintervall zwischen KA-Darbietung und ROA verstreicht (Bearden, Zukerman, Enneking, Erdfelder & Bayen, 1999). Schon ein Zeitintervall von wenigen Sekunden reicht aus, um die *b*-Parameterschätzung für ältere Menschen unter die für jüngere Erwachsene sinken zu lassen. Wird vor der ersten ROA die gesamten Liste der korrekten Antworten dargeboten, was einem KA-ROA-Intervall von ca. fünf Minuten entspricht, ist die *b*-Parameterschätzung bei älteren Erwachsenen nicht mehr signifikant von null verschieden. Einen ähnlichen Effekt kann man bei jüngeren Erwachsenen dadurch erzielen, daß nur die Items betrachtet werden, zu denen die KA nachträglich nicht abgerufen werden kann. Dies spricht dafür, daß es für den Rekonstruktionsbias *b* wesentlich ist, daß die KA zum Zeitpunkt der ROA-Generierung bewußt abgerufen werden kann. Es liegt also kein Prozeß der „immediate assimilation" der korrekten Antwort in die Wissensbasis zu einer Frage vor, wie es Fischhoff (z.B. 1982) vermutet hat. Wäre dem so, so müßte die KA einen Einfluß auf den Rekonstruktionsprozeß behalten, auch wenn sie nicht mehr bewußt abgerufen werden kann.

Paradoxerweise scheint also das schlechtere explizite Gedächtnis ältere Menschen vor einem starken Rekonstruktionsbias zu schützen, wenn die KA zeitlich vor ROA präsentiert wurde. Unter diesen Umständen zeigen jüngere Erwachsene einen stärkeren Rückschaufehler als ältere. Das Ergebnismuster kehrt sich kom-

plett um, wenn die KA simultan zur ROA präsentiert wird. Offenbar können ältere Menschen irrelevante Information schlechter unterdrücken als jüngere (Hasher, Stoltzfus, Zacks & Rypma, 1991), allerdings nur dann, wenn diese Information zum Zeitpunkt der ROA-Erzeugung auch verfügbar ist.

Modifizierte Versionen des Modells wurden von Dehn und Erdfelder (1998), Kohnert (1996), Pfister und Fischer (1992) sowie Pohl und Gawlik (1995) eingesetzt. Die Ergebnisse konvergieren dahingehend, daß Gedächtnisverzerrungen als Folge der KA-Präsentation zwar gelegentlich nachweisbar sind (Bearden et al., 1999; Erdfelder & Buchner, 1998a; Kohnert, 1996; Pohl & Gawlik, 1995), jedoch im Vergleich zu Rekonstruktionsverzerrungen eher schwach ausfallen und somit insgesamt wenig zum gesamten Rückschaufehlereffekt beitragen. Dies gilt auch dann, wenn Bedingungen geschaffen werden, die Gedächtnisbeeinträchtigungseffekte mutmaßlich begünstigen. Induziert man etwa eine „flache" Verarbeitung der OA dadurch, daß die OA nicht generiert, sondern nur ausgewählt werden muß (multiple choice test), so reduziert das zwar erwartungsgemäß r_C und r_E, führt jedoch nicht zu einer r_C-r_E-Differenz (Dehn & Erdfelder, 1998). Auch die Manipulation des numerischen Abstandes von KA und OA hat keinen Effekt auf die r_C-r_E-Differenz und darüber hinaus auch keinen Effekt auf die Rekonstruktionsverzerrungswahrscheinlichkeit *b*. Lediglich die Stärke der Verzerrung, gegeben daß ein Verzerrungseffekt eintritt (erfaßt durch die Antworttendenzparameter g_{g3} und g_{l3} des Modells), wird durch die OA-KA-Diskrepanz affiziert (Dehn & Erdfelder, 1998). Derartige Befunde widersprechen Theorien, die KA-OA-Differenzen einen moderierenden Einfluß auf das Eintreten bzw. Nichteintreten des HB-Effekts zusprechen. Sie legen die Vermutung nahe, daß bestätigende Befunde für derartige Theorien im wesentlichen auf der Verwendung von globalen Rückschaufehlermaßen beruhen, die nicht nur für das etwaige Eintreten einer Gedächtnisbeeinträchtigung oder einer Rekonstruktionsverzerrung sensitiv sind, sondern auch für das Ergebnis einer Rekonstruktion, gegeben daß eine Rekonstruktionsverzerrung vorliegt. Damit können diese Maße Antworttendenzen und Urteils- sowie Gedächtnisbeeinträchtigungen durch die KA-Präsentation nicht mehr sauber separieren.

Die einzige Maßnahme, mit der bedeutsame r_C-r_E-Differenzen im HB13-Modell relativ zuverlässig erzeugt werden können, scheint die interindividuelle Manipulation von Experimental- und Kontrollbedingung im Kontrast zur gewöhnlich benutzten intraindividuellen Manipulation zu sein (Bröder & Erdfelder, 1997). Mit anderen Worten: Der Gedächtnisbeeinträchtigungseffekt der Experimentalbedingung ist besonders dann groß, wenn eine Gruppe von Vpn zu *allen* Items und eine zweite Gruppe zu *keinem* Item ein korrekte Antwort rückgemeldet bekommt. Der Effekt wird nivelliert, wenn alle Vpn jeweils zur Hälfte der Items Feedback und zur anderen Hälfte kein Feedback bekommen (intraindividuelle Manipulation). Dies legt die Vermutung nahe, daß der Gedächtnisbeeinträchti-

gungseffekt der korrekten Antwort kein itemspezifischer ist, d.h. ein Feedback beeinträchtigt nicht nur auf die Abrufbarkeit der OA zu dem Item, auf das sich das Feedback bezieht. Es scheint sich auch nachteilig auf die Erinnerung anderer Originalantworten auszuwirken. Damit wäre eine enge Beziehung zum Paradigma der retroaktiven Hemmung hergestellt. Wie der retroaktive Hemmungseffekt im HB-Paradigma verursacht wird (Interferenzeffekt oder Abrufeffekt?), ist allerdings derzeit noch nicht endgültig geklärt (Bröder & Erdfelder, 1997).

7.2 Lernpsychologische Anwendungen

Ein Sammelreferat zu lernpsychologischen Anwendungen von MVB-Modellen könnte leicht ein komplettes Buch füllen. Wie eingangs bereits erwähnt, bilden mathematische Lernmodelle die historische Wurzel multinomialer Verarbeitungsbaummodellierung in der kognitiven Psychologie. Die Anzahl lernpsychologisch motivierter Anwendungen von MVB-Modellen dürfte daher die Anzahl primär gedächtnispsychologischer Anwendungen möglicherweise noch überschreiten.

Es soll hier nicht der Versuch unternommen werden, einen Überblick über das gesamte Spektrum von MVB-Lernmodellen zu geben. Die frühen Anwendungen bleiben komplett ausgeklammert, da ausführliche Zusammenfassungen von Theorien, Modellen und Befunden bis hinein die siebziger Jahre existieren (Atkinson et al., 1965; Greeno & Bjork, 1973; Tack, 1976). Hier soll nur auf einige neuere Modelle zum Diskriminationslernen, zum impliziten Lernen und vor allem zum Erwerb bzw. zur Anwendung von Wissen eingegangen werden. Alle zu schildernden Modelle sind mehr oder minder explizit kognitionspsychologisch motiviert. Kognitionspsychologisch begründete MVB-Modelle gab es auch schon vor 1980 (vgl. insbesondere das Sammelreferat von Greeno & Bjork, 1973); allerdings fallen diese Anwendungen nahezu ausschließlich in die Klasse der Markoff-Ketten-Modelle.

7.2.1 Diskriminationslernen

Riefer und Batchelder (1988, S. 332 ff) haben ein binäres MVB-Modell zum Diskriminationslernen vorgestellt, das eine separate Erfassung der beiden involvierten Prozesse „Stimulusdiskriminierung" und „Stimulus-Response-Lernen" ermöglichen soll. Sie wendeten das Modell auf ein PAL-Experiment an. Prinzipiell kann es jedoch für beliebige Diskriminationslernexperimente verwendet werden, in denen Reiz-Reaktions-Folgen zu lernen sind.

Der Grundgedanke des Diskriminationslernmodells ist die Kontrastierung von eindeutigen Reizen (A) mit leicht verwechselbaren Reizen (B, B´) innerhalb eines Experiments. Während beim Lernen von Reaktion zu A-Reizen nur Reiz-

Reaktions-Lernen stattfinden muß, kommt beim Lernen der Reaktionen zu B bzw. B´ noch der Stimulusdiskriminierungsprozeß hinzu. Angenommen wird, daß mit Wahrscheinlichkeit *s* eine Reiz-Reaktions-Sequenz gelernt wird, was bei A-Reizen automatisch zur korrekten Reaktion führt, bei B- und B´-Reizen jedoch nur dann, wenn keine Stimulusverwechslung stattfindet (mit Wahrscheinlichkeit 1–*c*); andernfalls ist ein Konfusionsfehler die Konsequenz, d.h. zu B wird die Reaktion gegeben, die eigentlich zu B´ gehört, oder vice versa. Bei mit Wahrscheinlichkeit 1–*s* nicht gelernten Reiz-Reaktions-Sequenzen wird die Reaktion erraten, was mit Wahrscheinlichkeit *g* ebenfalls zu einer korrekten Reaktion führen kann.

Riefer und Batchelder (1988, S. 335 f) zeigten empirisch, daß die Reihenfolge, in der die Reiz-Reaktions-Verbindungen gelernt werden, einen Einfluß auf die beiden Teillernprozesse hat: Sowohl in bezug auf den Stimulusdiskriminierungsprozeß (geringere Verwechslungswahrscheinlichkeit *c*) als auch den Reiz-Reaktions-Lernprozeß (höhere Speicherungswahrscheinlichkeit *s*) ist zunächst das Reiz-Reaktions-Paar im Vorteil, das zuerst gelernt wird. Mit der Anzahl der Lernversuche gleichen sich die Unterschiede zwischen zuerst und zuletzt gelernten Reiz-Reaktions-Paaren dann zunehmend aus.

Bäuml (1992) hat eine Erweiterung des Diskriminationslernmodells von Riefer und Batchelder (1988) zu einem Markoff-Ketten-Modell vorgeschlagen, das die Übergangswahrscheinlichkeit zwischen den Lernzuständen über die Lerndurchgänge hinweg als stationär annimmt. In seinem Modell gibt es neben den Ratewahrscheinlichkeiten zwei Stimulus-Response-Parameter (für eindeutige und leicht verwechselbare Reize) sowie zwei Stimulusdiskriminierungsparameter (für Stimuli mit bzw. ohne bereits gelernte Reiz-Reaktions-Verbindung). Er wendete das Modell erfolgreich auf das Diskriminationslernen von aufrecht oder rotiert dargebotenen Gesichtern an, denen unterschiedliche Reaktionen zuzuordnen waren. Die leicht verwechselbaren Reize waren identische Gesichter in unterschiedlichen Rotationswinkeln. Im Einklang mit einer zuvor von Riefer und Batchelder (1988, S. 337) aufgestellten Hypothese hatte der Rotationswinkel insgesamt wenig Einfluß auf das Stimulus-Response-Lernen, jedoch einen starken Einfluß auf die Stimulusdiskriminierung: Je größer der Rotationswinkel identischer Gesichter, desto größer war auch der Diskriminationslernparameter. Dieser Befund ist insofern bedeutsam, als verschiedentlich die Hypothese geäußert wurde, daß rotierte Gesichter qualitativ anders verarbeitet werden als aufrechte Gesichter. Das Ergebnis von Bäuml (1992) spricht demgegenüber für bloß quantitative Unterschiede zwischen der Repräsentation rotierter Gesichter: Ihre Diskriminierbarkeit nimmt annähernd linear mit dem Winkel zu, um den sie gegeneinander rotiert sind.

Riefer und Batchelder (1988, S. 337) haben gezeigt, daß die Hypothese qualitativ unterschiedlicher Verarbeitungsformen aufrechter und rotierter Gesichter *fälschlich* als gestützt gelten kann, wenn statt einer modellbasierten Analyse, die

Stimulusdiskriminierung und Stimulus-Response-Lernen separat erfaßt, eine varianzanalytische Auswertung der Häufigkeit korrekter Antworten erfolgt. Dies zeigt einmal mehr, wie wichtig es in der kognitiven Psychologie ist, sich nicht allein auf varianzanalytische Auswertungen zu verlassen (vgl. auch Batchelder & Riefer, 1986; Erdfelder & Bayen, 1991)

7.2.2 Explizites und implizites Lernen

Buchner, Steffens et al. (1997) haben versucht, bewußte und unbewußte Anteile beim Lernen von Ereignissequenzen durch eine Modifikation der Prozeßdissoziationsprozedur von Jacoby (1991) zu erfassen. In ihrem Paradigma müssen die Vpn zunächst lernen, auf Ereignisse (z.B. Töne unterschiedlicher Höhe) mit bestimmten Reaktionen reagieren. Die Abfolge der Ereignisse folgt dabei einer bestimmten Regel, welche die Vp nicht kennt. Der typische Befund ist, daß bei einem Wechsel von einer regelbasierten Folge auf eine unsystematische (zufällige) Folge ein bedeutsamer Anstieg der Antwortzeiten zu beobachten ist, der jedoch keinem expliziten Sequenzwissen korrespondiert, d.h. auch dann zu beobachten ist, wenn die Vp die Regel, nach der die Ereignisfolge aufgebaut ist, weder benennen noch rekonstruieren kann.

Ein naheliegender Einwand gegen den möglichen Schluß, daß implizites Sequenzlernen und explizites Sequenzwissen von zwei unabhängigen Gedächtnissystemen gespeist werden — dem prozeduralen und dem deklarativen Gedächtnis —, ist die Fragwürdigkeit der verwendeten Maße des *erschöpfende* Maße des expliziten Sequenzwissens (Perruchet & Amorim, 1992). Um dieses Problem zu umgehen, schlugen Buchner, Steffens et al. (1997) ein multinomiales Modell zur Messung expliziten und impliziten Sequenzwissens vor. Sie boten ihren Vpn nach der Akquisitionsphase drei Klassen von Ereignissequenzen dar: alte Sequenzen der Akquisitionsphase, neue Sequenzen, die einer neuen Regel folgten, und schließlich neue Sequenzen, die zufällig aufgebaut waren. Die Vpn wurden auf zwei Versuchsbedingungen zufällig aufgeteilt: Vpn der Inklusionsbedingung sollten beurteilen, ob die Testsequenz „wohlgeformt" sind, d.h. einer bestimmten Regel folgen. Dabei war es unerheblich, ob die Regel der alten Sequenz entsprach oder nicht. Im Gegensatz dazu durfte in der Exklusionsbedingung eine Sequenz nur dann als wohlgeformt bewertet werden, wenn sie einer neuen Regel folgte, d.h. nicht der aus der Akquisitionsphase bekannten Regel entsprach.

Die Wohlgeformtheitsurteile in der Inklusions- und in der Exklusionsbedingung wurden durch ein verbundenes binäres MVB-Modell erklärt („sequence identification measurement model", SIMM), das Parameter für die primär interessierenden Konstrukte „explizites Regelwissen" (c) und „implizites Regellernen ohne explizites Regelwissen" ($u_{c\text{-}}$) beinhaltet sowie zusätzlich die Hilfsparameter

„Systematizitätsentdeckung" (*s*), „Entdeckung von Zufallsfolgen" (*d*) und „Raten" (g_i und g_e). Das Modell paßte gut auf die Daten dreier Experimente. Intentionales versus inzidentelles Lernen beeinflußte hypothesengemäß den Parameter *c*, während sich die Manipulation der „Flüssigkeit" alter Sequenzen in der Testphase (Beschleunigung versus Verlangsamung der Darbietungsgeschwindigkeit gegenüber der Akquisitionsphase) hypothesengemäß auf den u_c-Parameter auswirkte. Dies spricht dafür, daß das Modell explizites Regelwissen und implizites Regellernen (vermittelt durch perzeptuelle Flüssigkeit) valide erfassen kann.

Das hinsichtlich der Parameter *c* und u_c- erfolgreich validierte SIMM-Modell wurde von Buchner, Steffens und Rothkegel (1998) eingesetzt, um den Einfluß des Trainings auf explizites und impliziten Sequenzlernens zu untersuchen. Im Einklang mit der Hypothese von Perruchet und Amorim (1992) schlugen sich Trainingseffekte früher und stärker im expliziten Sequenzwissen (*c*) verglichen mit dem impliziten Sequenzlernen (u_c-) nieder (Experiment 1). Schon minimales Training war ausreichend für einen Anstieg von *c*, nicht jedoch für einen Anstieg von u_c-. Der Erwerb expliziten Sequenzwissens scheint als beim Sequenzlernen, wie es sich typischerweise in den Antwortzeiten niederschlägt, eine zentrale Rolle zu spielen. Grundlage hierfür ist vermutlich die Gruppierung von kurzen Subsequenzen zu leicht abrufbaren „chunks". Erleichtert man die adäquate Gruppierung durch künstlich verlängerte Zeitintervalle vor bzw. nach einem „chunk" in der Akquisitions- und Testphase gleichermaßen, so führt dies zu einem signifikanten Anstieg des expliziten Systemwissens (*c*) (Buchner et al., 1998, Experiment 2).

7.2.3 Konzeptwissen versus Beispielwissen

Wissen kann einerseits in Form von Beispielwissen und andererseits in Form von abstraktem Konzept- oder Regelwissen erworben werden. Beispielwissen wird bei einer Aufgabenbearbeitung eingesetzt, wenn (a) die gestellte Aufgabe einer Aufgabe aus der Trainingsphase exakt entspricht und (b) die Aufgabe und ihre korrekte Lösung auf der Basis einer episodischen Gedächtnisspur erinnert werden. Konzeptwissen verlangt demgegenüber keinen Rückgriff auf episodische Gedächtnisspuren aus der Lernphase: Auf der Basis des erworbenen abstrakten Regel- oder Konzeptwissens wird die Lösung zu einer Aufgabe — sei die Aufgabe nun alt oder neu — abgeleitet.

Müller (1996a) hat ein multinomiales Modell zur Separierung von Beispielwissen und Konzeptwissen entwickelt und auf ein Training zur korrekten Anwendung von Funktionsaufrufen in einer Programmiersprache angewendet. Das Modell bezieht sich einerseits auf eine Rekognitionsaufgabe, bei der Aufgaben als „alt" (d.h. aus der Trainingsphase bekannt) oder „neu" zu klassifizieren sind. Dieser Teil des Modells ist mit dem 2HT-Modell der Rekognition identisch (vgl.

Abschnitt 3.5.2). Er dient zur Schätzung des Beispielwissens (Rekognitionsparameter *i*) und der Antworttendenz *r* für „Alt"-Antworten. Darüber hinaus müssen korrekte und inkorrekte Funktionsaufrufe, die aus der Trainingsphase stammen oder neu sind, als zulässig oder unzulässig klassifiziert werden. Für alte Aufgaben wird angenommen, daß ein korrektes Urteil einerseits aufgrund der Rekognition der alten Aufgabe und ihrer Lösung mit Wahrscheinlichkeit *i* erfolgen kann. Wird die alte Aufgabe mit Wahrscheinlichkeit (1 – *i*) nicht rekogniziert oder handelt es sich um eine neue Aufgabe, so wird die Aufgabe mit Wahrscheinlichkeit *k* aufgrund des erworbenen Konzeptwissens korrekt beantwortet. Versagen sowohl Beispielwissen als auch Konzeptwissen, so bleibt immer noch die Möglichkeit, die korrekte Antwort zu erraten.

Das geschilderte Modell paßte gut auf die Daten einer Trainingsstudie mit einem versus zwei Tagen Übung. Erwartungsgemäß stiegt der Parameter *k* für Konzeptwissen mit der Trainingsdauer an. Der Parameter *i* für Beispielwissen blieb dagegen von der Dauer des Trainings unaffiziert; die Parameterschätzungen hierfür lagen auf einem generell niedrigen Niveau (unter .10).

Müller und Funke (1995) sowie Müller (1996b) haben zwei sehr ähnliche Modelle zum Erlernen von Bediensequenzen vorgeschlagen. Ihre Vpn mußten Betätigungsfolgen verschiedener Tasten auf einem Display erlernen und später alte und neue Bediensequenzen auf ihre Zulässigkeit hin beurteilen. Ihre Modelle zielen anders als das Modell von Müller (1996a) nicht auf die globale Erfassung von Beispielwissen versus Konzeptwissen ab, sondern primär auf die Separierung verschiedener Formen des Konzeptwissens: Vollständiges „Grammatikwissen" über zulässige Gesamtfolgen soll von begrenztem Phrasenwissen über zulässige Teilsequenzen unterschieden werden. Angenommen wird, daß die verschiedenen Formen des Konzeptwissens neben bloßen Systematizitätsvermutungen Einfluß auf Zulässigkeitsurteile nehmen können.

Während in der Untersuchung von Müller und Funke (1995) die Wahrscheinlichkeit vollständigen Grammatikwissens (Parameter *s*) auf null gesetzt werden konnte, war das bei Müller (1996b) in den meisten Versuchsbedingungen nicht der Fall. Kritisch für den Erwerb echten Grammatikwissens scheint die permanente Verfügbarkeit einer Bediensequenz zu sein, d.h. der Überblick über alle bis zu einem bestimmten Zeitpunkt getätigten Eingriffe. Ist die komplette Bediensequenz weder in der Trainings- noch in der Testphase verfügbar, so kann $s = 0$ gesetzt werden (Müller & Funke, 1995). Müller (1996b) zeigte, daß sich die Verfügbarkeit der kompletten Bediensequenz sowohl in der Trainings- als auch in der Testphase förderlich auf die Wahrscheinlichkeit *s* echten Grammatikwissens auswirkt. Ressourcen des Arbeitsgedächtnisses scheinen also für den Erwerb und die korrekte Anwendung von „vollständigem Grammatikwissen" über Bediensequenzen von besonderer Bedeutung zu sein. Während sich Arbeitsgedächtnisent-

lastungen in der Lernphase und der Testphase annähernd additiv auf den Parameter *s* auswirkten, hatten sie praktisch keine Effekt auf partielles Phrasenwissen (Müller, 1996b). Partielles Phrasenwissen scheint also mehr oder minder automatisch erworben zu werden, ohne Ressourcen des Arbeitsgedächtnisses zu beanspruchen.

7.2.4 Wissenszustände und Wissensräume

Falmagne, Koppen, Villano, Doignon und Johannes (1990) haben eine Theorie der Wissensräume vorgestellt, deren Hauptziel die Erfassung der Struktur möglicher Wissenszustände bezüglich einer bestimmten Wissensdomäne ist, beispielsweise Wissen über Schach oder lineare Algebra. Ist die Struktur möglicher Wissenszustände bekannt, so lassen sich daraus Hypothesen über verschiedene „Lernpfade" ableiten, auf denen zunehmend komplexeres Wissen über die betreffende Domäne erlernt werden kann.

Es wird angenommen, daß jede Wissensdomäne durch eine Menge **Q** von *n* Fragen charakterisierbar ist. Ein „Wissenszustand" *K* einer Personen ist einfach die Teilmenge aller Fragen aus **Q**, welche die betreffende Person korrekt beantworten kann. Ein „Wissensraum" K ist die Menge aller möglichen Wissenszustände zu der betreffenden Domäne. Empirisch gehaltvoll wird die Theorie der Wissensräume dadurch, daß typischerweise bestimmte Teilmengen von **Q** aus K ausgeschlossen werden, so daß K eine echte Teilmenge der Potenzmenge Pot(**Q**) ist. Praktisch bedeutet dies, daß die Antworten auf zumindest einige Fragen der Wissensdomäne aus dem Wissen über andere Fragen unmittelbar ableitbar sein müssen. Ist z.B. das Wissen über A und B aus dem Wissen über C ableitbar, so kann es keinen Wissenszustand {C} in K geben, d.h. keinen Wissenszustand, zu dem *ausschließlich* die Kenntnis der Antwort zu C ohne gleichzeitige Kenntnis der Antworten zu A und B gehört.

Ein Wissensraum K ist nun aber keine beliebige Teilmenge von Pot(**Q**). Gefordert wird, daß sowohl absolutes Nichtwissen ($\varnothing$) als auch vollständiges Wissen (**Q**) zu jedem Wissensraum gehören. Gefordert wird ferner, daß jeder Wissensraum K bezüglich der Vereinigungsbildung abgeschlossen ist, d.h. wenn $K \in \mathsf{K}$ und $K' \in \mathsf{K}$, dann ist auch $K \cup K' \in \mathsf{K}$. Dies ist plausibel, wenn man annimmt, daß Individuen mit unterschiedlichen Wissenszuständen *K* und *K´* miteinander interagieren, was potentiell zum Erlernen des Wissenszustandes des Interaktionspartners führen kann.

Ein einfaches Beispiel für einen Wissensraum ist das Modell einer Guttman-Skala (vgl. z.B. Roskam, 1996b). Das Guttman-Modell postuliert, daß alle Fragen auf einen eindimensionalen Schwierigkeitskontinuum angeordnet werden können, so daß die Kenntnis der Antwort auf eine bestimmte Frage die Kenntnis der

Antworten auf alle leichteren Fragen impliziert. Besteht **Q** aus n Fragen, so gibt es folgt n+1 Zustände im Wissensraum der Guttman-Skala.

Nicht jeder Wissensraum muß so restriktive Annahmen wie das Guttman-Modell machen (vgl. etwa das Beispiel in Falmagne et al., 1990, Abbildung 1, S. 203, oder das weiter unten dargestellte Beispiel). Dennoch bleibt das Problem, daß Wissensräume in der bislang vorgestellten Form deterministische Modelle sind, die strenggenommen durch die Beobachtung eines einzigen Wissenszustandes, der nicht zum spezifizierten Wissensraum gehört, falsifiziert werden. Sie entsprechen in dieser Hinsicht dem Guttman-Modell oder dem Modell von Levine (1966) aus Abschnitt 2.4. Lösen kann man das Problem auf die gleiche Weise, wie es in Abschnitt 2.5 für das deterministische Levine-Modell dargestellt wurde. Man postuliert einfach, daß ein bestimmter Wissensraum auf die latente Struktur „hinter" den beobachteten Antworten zutrifft. Abweichungen zwischen den wahren Wissenszuständen und den beobachteten Antworten auf die Fragen können einerseits durch „lucky guesses" zustande kommen, bei denen mit einer bestimmten Wahrscheinlichkeit v_q eine korrekte Antwort auf Frage q gelingt, obwohl sie nicht zum Wissenszustand einer Person gehört. Umgekehrt können mit Wahrscheinlichkeit β_q auch „careless errors" vorkommen, wenn die Frage q zum Wissenszustand einer Person gehört, aber dennoch versehentlich falsch beantwortet wird. Falmagne (1989) legte genau diese Annahmen für seine probabilistische Theorie der Wissensräume zugrunde und postuliert zusätzlich lokale stochastische Unabhängigkeit der Antworten für einen gegebenen Wissenszustand K. Diese Annahmen entsprechen einem Latent-Class-Modell, in dem die verschiedenen Wissenszustände als latente Klassen fungieren. Da Latent-Class-Modelle zugleich MVB-Modelle sind (vgl. Abschnitt 6.3), fällt das probabilistische Wissensraum-Modell ebenfalls in die Klasse der MVB-Modelle.

Da die Anzahl der Wissenszustände in einem Wissensraum in der Regel recht groß ist, sind probabilistische Wissensraum-Modelle ohne weitere Restriktionen normalerweise nicht identifizierbar. Falmagne et al. (1990) haben einerseits Gleichsetzungsrestriktionen für die v- und β-Parameter und andererseits Restriktionen für die Wahrscheinlichkeiten der verschiedenen Wissenszustände erwogen. Zum letztgenannten Punkt hatte bereits Falmagne (1989) eine stochastische Lerntheorie formuliert. Sie basiert auf der Grundidee, daß von Zustand des Nichtwissens ∅ zum Zustand des vollständigen Wissens **Q** verschiedene Lernpfade führen, die Wissenszustände zunehmenden Komplexitätsgrades miteinander verbinden. Zielsetzung der stochastischen Lerntheorie ist die Diagnose der möglichen Lernpfade und ihrer Wahrscheinlichkeiten in einer bestimmten Population. Hierzu erweiterte Falmagne (1989) das Latent-Class-Modell auf ein Modell, das unterschiedliche Lernraten pro Zeit für die Individuen und unterschiedliche Schwierigkeitsparameter für die verschiedenen Aufgaben berücksichtigt. Wie

Falmagne et al. (1990) an einem empirischen Beispiel demonstrierten, kann ausgehend von einem sehr allgemeinen Modell mit vielen möglichen Lernpfaden durch sukzessive Parameterrestriktionen eventuell ein recht einfaches Modell mit nur wenigen Lernpfaden gewonnen werden. In dem exemplarisch analysierten Fall von Schülerantworten auf fünf Fragen zur linearen Algebra resultierte auf diese Weise eine akzeptable Modellanpassungsstatistik von $G^2_{(2)} = 3.5$. Die Verteilung der Schüler verschiedener Klassenstufen auf die Wissenszustände des Modells entsprach den Prozeßannahmen des Modells, d.h. Schüler der jüngeren Altersstufen wiesen gehäuft elementarere Wissenszustände auf, ältere Schüler dagegen eher komplexere Wissenszustände.

Durch die unterschiedlichen Aufgabenschwierigkeitsparameter und individuellen Lernratenparameter, für die eine allgemeine Gamma-Verteilung angenommen wurde, gehört das Modell von Falmagne (1989) zwar noch zu den parametrisierten multinomialen Modellen, aber nicht mehr zur Klasse der MVB-Modelle. Das für die Schülerdaten letztlich akzeptierte Modell mit vier Lernpfaden (Falmagne et al., 1990, Abbildung 7, S. 216) ließe sich aber prinzipiell auch als identifizierbares und testbares MVB-Modell formulieren, wenn man neben den β- und *v*-Parametern auch einige Übergangswahrscheinlichkeiten zwischen den Wissenszuständen restringieren würde. Das Baummodell hierzu wäre dem Lernpfaddiagramm bei Falmagne et al. (1990, S. 216) weitgehend analog, sieht man einmal von zusätzlich vorzusehenden Ästen für die Zustandswahrscheinlichkeiten und für die möglichen Antwortfehler ab.

Zur Illustration des Vorgehens sei ein einfacheres binäres MVB-Modell zum Wissensraum $\mathsf{A} = \{\varnothing, \{A\}, \{B\}, \{A,B\}, \{A,B,C\}\}$ einer Drei-Fragen-Domäne $\mathbf{Q} = \{A, B, C\}$ dargestellt. Das Modell läßt sich am einfachsten als Mischung von zwei Guttman-Skalen in den Anteilen a und $(1-a)$ charakterisieren. Es unterscheidet sich durch einen Wissenszustand vom Modell einer homogenen Guttman-Skala: Die Fragen A und B lassen sich nicht in eine für alle Probanden gleiche Schwierigkeitsrangordnung bringen. Es gibt folglich nicht nur einen Lernpfad von ∅ nach **Q**, sondern zwei. Das Gesamtmodell für beide Lernpfade ist in Abbildung 7.1 als einfaches „Stop-and-Go-Diagramm" dargestellt. Ausgehend vom Wissenszustand ∅ kann man mit einer bestimmten Wahrscheinlichkeit (*p*-Parameter) im gegeben Zustand verharren („Stop") oder zum nächstkomplexeren Wissenszustand fortschreiten („Go"). An einer Stelle ergibt sich dabei eine Weiche: Mit Wahrscheinlichkeit a kann man nach ∅ zuerst den Wissenszustand {A}, mit der komplementären Wahrscheinlichkeit zuerst den Wissenszustand {B} erlernen. Zu beachten ist, daß — von $p_\varnothing$ einmal abgesehen — die *p*-Parameter bedingte Wahrscheinlichkeiten sind und somit nicht die Wahrscheinlichkeit des betreffenden Wissenszustandes direkt angeben. Vielmehr muß die Summe der Wahrscheinlichkeiten aller Äste berechnet werden, die in den betreffenden Wissenszustand einmünden. Fer-

ner zu beachten ist, daß der Wissenszustand {A, B} im Baummodell lediglich aus graphischen Gründen doppelt auftaucht. Faktisch handelt es sich in beiden Fällen um den gleichen Zustand {A, B}, in dem man mit der Wahrscheinlichkeit $p_{\{A,B\}}$ verharrt und aus dem man mit der komplementären Wahrscheinlichkeit $1 - p_{\{A,B\}}$ zum Zustand vollständigen Wissens fortschreitet. Die gestrichelten Kästen ganz rechts im Diagramm symbolisieren das jeweils noch zu ergänzende Antwortmodell für den betreffenden Wissenszustand, welches „lucky guesses" und „careless errors" mit Wahrscheinlichkeiten v_q bzw. β_q vorsieht.

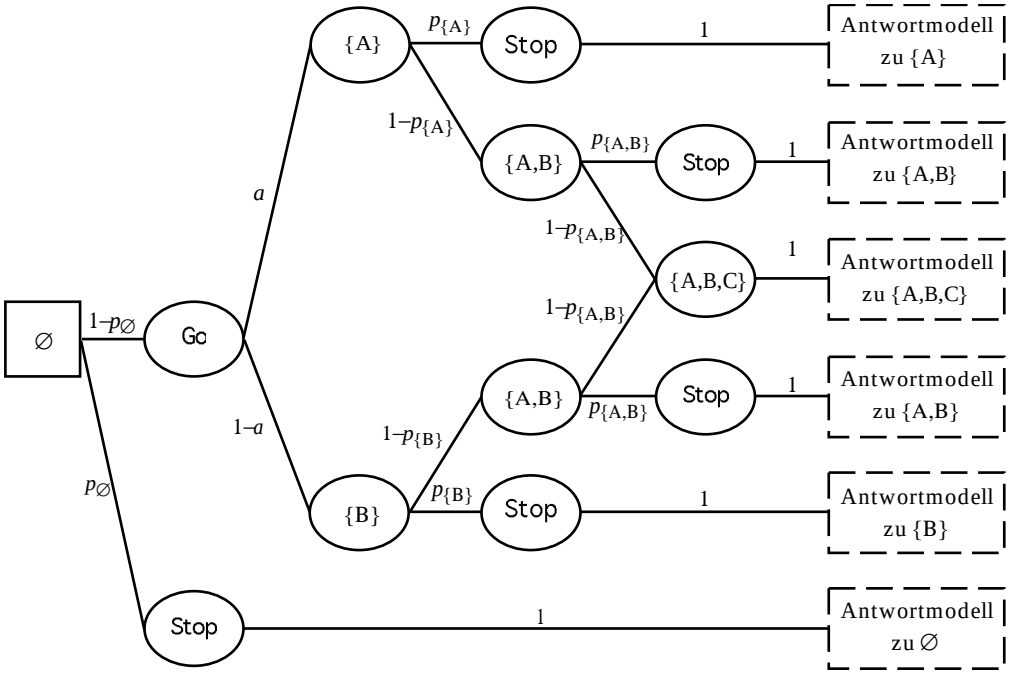

Abbildung 7.1:
Ein MVB-Modell zum Wissensraum A = {∅, {A}, {B}, {A,B}, {A,B,C}} für eine Drei-Fragen-Wissensdomäne Q = {A, B, C}. Erläuterungen im Text.

Ohne weitere Parameterrestriktionen ist das in Abbildung 7.1 skizzierte Modell für die 2×2×2-Kontingenztafel der möglichen Antworten auf die drei Fragen A, B und C nicht identifizierbar. Identifizierbarkeit kann jedoch in den meisten Anwendungskontexten auf unterschiedlichen Wegen hergestellt werden. Hierzu zählen neben Restriktionen der v_q- und β_q-Parameter vor allem Erweiterungen auf Mehr-Gruppen-Designs. Hat man etwa wie Falmagne et al. (1990) zwei oder mehr Schülergruppen unterschiedlichen Alters mit den gleichen Fragen **Q** untersucht, so ist die Annahme plausibel, daß sich die v_q- und β_q-Parameter sowie die Lernpfadwahrscheinlichkeit *a* zwischen diesen Gruppen nicht unterscheiden. Altersgruppenunterschiede sollten sich ausschließlich in den *p*-Parametern manifestieren. Kann man zusätzlich noch unterstellen, daß (a) die v_q- und die β_q-Parameter für alle Items gleich sind und (b) die bedingten Wahrscheinlichkeiten des Verbleibens in den Wissenszuständen {A} und {B} innerhalb der Altersgruppen eben-

falls gleich sind (d.h. $p\{A\} = p\{B\}$ pro Gruppe), so wird das Modell identifizierbar und testbar. Eine Modelleingabedatei für die Computerprogramme MBT, GPT oder AppleTree zum Modell aus Abbildung 7.1 — angewandt auf zwei Altersgruppen — ist in Abschnitt 9 des Anhangs wiedergegeben.

Das Beispiel demonstriert, daß es prinzipiell möglich ist, lernpsychologisch motivierte probabilistische Wissensraummodelle vollständig im Rahmen der MVB-Familie zu formulieren, auch wenn dies einen gewissen Verzicht an Detailliertheit im Vergleich zur stochastischen Lerntheorie von Falmagne (1989) mit sich bringt. Sofern das Strukturmodell gemäß Abbildung 7.1 von primärem oder ausschließlichem Interesse ist und die zum Zwecke der Identifizierbarkeitsherstellung erforderlichen Restriktionen unproblematisch sind, wiegt diese Verzicht aber nicht schwer.

7.3 Denkpsychologische Anwendungen

Neben der Lern- und der Gedächtnispsychologie ist auch die Denkpsychologie ein potentiell ergiebiges Feld für die multinomiale Modellierung. Fruchtbar war der MVB-Ansatz bislang vor allem im Bereich des schlußfolgernden Denkens. Diesem Bereich werde ich mich folgerichtig zuerst zuwenden. Anwendungen der MVB-Modellierung in anderen Gebieten der Denkpsychologie sind bislang erstaunlicherweise eher rar, obwohl sich solche Ansätze gerade in der Begriffsidentifikationsforschung (vgl. etwa das LCA-Modell zum Paradigma von Levine, 1966, in Abschnitt 6.3.1 und 6.3.2) und in der Forschung zum probabilistischen Urteilen (vgl. das Modell zum Konjunktionsfehler in Abschnitt 6.4.1 und 6.4.2) anbieten. Ein Beispiel für die MVB-Modellierung von Entscheidungsprozessen wird in Abschnitt 7.3.2 berichtet.

7.3.1 Schlußfolgerndes Denken

Drei Paradigmen zur Untersuchung schlußfolgernden Denkens sind bislang Anlaß für MVB-Modellierungen gewesen: die Selektionsaufgabe nach Wason, Beurteilungen von propositionalen Schlußfolgerungsaufgaben und Beurteilungen von syllogistischen Schlußfolgerungsaufgaben.

7.3.1.1 Die Selektionsaufgabe nach Wason

Die bekannte Selektionsaufgabe nach Wason (1968) (Wason selection task) lädt zu einem multinomialen Modellierungsansatz geradezu ein. Aus der Aufgabe ergibt sich unmittelbar eine reichhaltige Struktur von Beobachtungskategorien: Probanden bekommen vier Karten mit den Aufschriften „*p*", „*nicht-p*", „*q*" und „*nicht-q*" vorgelegt, die sie jeweils umdrehen oder nicht umdrehen können; bei vier Karten ergibt dies $2^4 = 16$ mögliche Antwortmuster pro Selektionsaufgabe. Den Probanden

wird mitgeteilt, daß sich auf der Rückseite jeder Karte eine weitere Aufschrift befindet. Sie sollen die und nur die Karten umdrehen, die eine Verletzung einer bestimmten Wenn-dann-Regel — z.B. „Wenn auf der einen Seite der Karte *p* steht, dann steht auf der anderen Seite *q*" — potentiell offenbaren können. Im Beispiel wäre die Wahl der *p*- und der *nicht-q*-Karte logisch korrekt. Statt der Aufschriften „*p*"und „*q*"können — je nach Anwendungskontext der Selektionsaufgabe — auch andere Symbole, Zahlen oder Aussagen Verwendung finden. Das logisch korrekte Antwortmuster wird bekanntlich nur bei bestimmten inhaltlichen Einbettungen der Aufgabe häufig gegeben. Wählt man wie im Beispiel abstrakte Einbettungen, so dominieren logisch inkorrekte Selektionen, typischerweise reine *p*-Wahlen oder kombinierte *p*- und *q*-Wahlen.

Evans (1977) hat als erster die Möglichkeiten einer stochastischen Modellbildung zur Selektionsaufgabe erkannt. Sein Modell basiert auf der Annahme stochastisch unabhängiger Wahlen aller vier Karten. Für die *p*- und die *nicht-p*-Karten wurden dies nicht explizit geprüft, da — so Evans (1997) — die entsprechenden Selektionswahrscheinlichkeiten nahe bei eins bzw. null liegen. Für die verbleibende 2×2-Kontingenztafel der *q*- und *nicht-q*-Wahlen ließ sich das Modell der stochastischen Unabhängigkeit empirisch aufrechterhalten. Allerdings waren die Stichprobenumfänge in allen von ihm reanalysierten Datensätzen sehr klein (größtenteils kleiner als $N = 40$), so daß die Teststärke des Unabhängigkeitstests unzureichend war. Selbst für mittlere Effekte ($w = .3$) und $\alpha = .05$ lag sie durchweg unter $(1–\beta) = .50$ (Cohen, 1988).

Der Vorteil der Unabhängigkeitsannahme ist, daß man sich auf die Modellierung der Randwahrscheinlichkeiten für die vier Kartenwahlen „*p*", „*nicht-p*", „*q*" und „*nicht-q*" beschränken kann. Die Wahrscheinlichkeiten bestimmter Kartenkombinationen ergeben sich per Voraussetzung durch entsprechende Produktbildung. Evans (1977) betrachtete allerdings nicht nur die Regel „Wenn *p*, dann *q*". Zusätzlich berücksichtigte er die Regeln „Wenn *p*, dann *nicht-q*", „Wenn *nicht-p*, dann *q*" und „Wenn *nicht-p*, dann *nicht-q*". Da zu jeder der vier Regeln vier freie Kartenselektionswahrscheinlichkeiten vorliegen, ergeben sich insgesamt $4 \cdot 4 = 16$ funktional unabhängige Kategoriewahrscheinlichkeiten. Hierzu schlug Evans ein verbundenes binäres MVB-Modell vor, das lediglich zwei Logikparameter und zwei Antworttendenzparameter umfaßt. Die Logikparameter (α-Parameter) sind Wahrscheinlichkeiten für das Erkennen einer aus logischer Sicht umzudrehenden Karte; sie führen per definitionem zu korrekten Selektionen. Ist eine Karte aus logischer Sicht nicht umzudrehen oder wird eine aus logischer Sicht umzudrehende Karte nicht erkannt, so kommen Antworttendenzen ins Spiel: Mit einer bestimmten Ratewahrscheinlichkeit *R* wird die betreffende Karte aufgrund nichtlogischer Evidenz dennoch umgedreht.

Die Logikparameter sollen von der konkret zu prüfenden Wenn-dann-Regel unabhängig sein; sie können sich jedoch zwischen Prämissen- und Konklusionskarten unterscheiden (Parameter α_a bzw. α_c). Erwartet wird $\alpha_a > \alpha_c$. Bei der Standardregel „Wenn *p*, dann *q*" sollte also beispielsweise der Logikparameter für die umzudrehende „Prämissenkarte" *p* größer sein als der für die umzudrehende „Konklusionskarte" *nicht-q* .

Die Antworttendenzwahrscheinlichkeiten für eine Karte hängen nur davon ab, ob die betreffende Karte affirmativ formuliert ist (*p*-Karte, *q*-Karte) oder eine Negation zum Ausdruck bringt (*nicht-p*-Karte, *nicht-q*-Karte). Für Affirmationen wird eine „Ratewahrscheinlichkeit" $R+$, für Negationen eine „Ratewahrscheinlichkeit" $R-$ angenommen. Im Einklang mit dem häufig beobachteten „matching bias" wird dabei $R+ > R-$ postuliert.

Evans (1977) hat sein Modell relativ informell analysiert. Insbesondere berichtete er keine Anpassungsstatistiken, sondern nur Gegenüberstellungen von beobachteten und erwarteten Selektionswahrscheinlichkeiten für einen Datensatz von Evans und Lynch. Die Parameterschätzungen für die Daten von Evans und Lynch entsprachen jedoch den Erwartungen: Der Schätzer für α_a lag bei etwa .50, der für α_c bei .15; $R+$ wurde auf ca. .70 geschätzt, $R-$ auf .18. Insgesamt spricht das Ergebnis für eine leichtere logische Verarbeitbarkeit der Prämisse im Vergleich zur Konklusion. Der „matching bias" wird als Determinante des Rateverhaltens der Vpn bestätigt.

Krauth (1982) hat das Modell von Evans (1977) explizit als Finite-State-Modell mit zwei latenten Zuständen „Einsicht" (Erkennen der logischen Zusammenhänge) und „Matching" (Antworten auf der Basis von Antworttendenzen) reformuliert. Ferner hat er es auf ein Modell mit drei Zuständen „Einsicht", „Verifikationstendenz" und „Matching" verallgemeinert. Dieses verallgemeinerte Modell hat nur einen freien Parameter mehr als das Modell von Evans (1977), führte aber zu deutlich besseren Anpassungsresultaten an mehrere von Krauth (1982) reanalysierte und neu erhobene Datensätze. Insbesondere trat das Problem negativer Schätzwerte für Zustandswahrscheinlichkeiten nicht auf, die bei Evans´ Modell gelegentlich zu beobachten waren. Mittels Pearson-χ^2-Anpassungstests zeigte Krauth (1982), daß sein Modell auf alle betrachteten Datensätze sehr gut paßt. Die Ergänzung des Evans-Modells um den Zustand „Verifikationstendenz" scheint also sinnvoll zu sein.

Gleißner und Erdfelder (1994) haben einen anderen Zugang zur Selektionsaufgabe gewählt, der nicht auf der Annahme stochastisch unabhängiger Kartenwahlen aufbaut. Kerngedanke ihres Modells ist die Idee, daß sich das Erkennen logisch korrekter Wahlen in der *gemeinsamen* Selektion der beiden korrekten Karten (typischerweise *p* und *nicht-q*) manifestiert. Die Wahrscheinlichkeit für eine solchermaßen verstandene analytische Bearbeitung der Selektionsaufgabe sei *a*.

Mit der komplementären Wahrscheinlichkeit 1–*a* kommt es zu einer heuristischen Bearbeitung, die durch stochastische unabhängige Kartenwahlen gekennzeichnet ist. Das Unabhängigkeitsmodell, welches den Ansätzen von Evans (1977) und Krauth (1982) zugrunde liegt, ergibt sich also als Spezialfall des Gleißner-Erdfelder-Modells für $a = 0$. In dem Maße, in dem die Vpn eine analytische Bearbeitung praktizieren, muß sich das Unabhängigkeitsmodell demnach als falsch erweisen. Diese Prognose wurde von Gleißner und Erdfelder (1994) empirisch bestätigt.

Gleißner und Erdfelder (1994) schlugen ferner vor, nicht die komplette 2^4-Kontingenztafel aller möglichen Antwortmuster zu modellieren. Da typischerweise mehrere Zellen dieser Tafel schwach oder gar nicht besetzt sind, stößt ein solcher Modellierungsansatz auf Schwierigkeiten bei der Datenanalyse. Er liefert auch mehr Heuristikparameter als gewöhnlich von Interesse sind. Gleißner und Erdfelder (1994) faßten deshalb die 16 Zellen der Tafel so zu fünf Beobachtungskategorien zusammen, daß ein vollparametrisiertes Vier-Parameter-Modell mit einem Logikparameter (*a*) und drei Parametern für heuristische Kartenwahlen (*p* für die heuristische Wahl der *p*-Karte, $q_{\neg}$ für die heuristische Wahl der *nicht-q*-Karte und *r* für die heuristische Wahl der *nicht-p*- oder der *q*-Karte) identifizierbar wird.

Zugleich berichteten Gleißner und Erdfelder (1994) ein Validierungsexperiment, in dem versucht wurde, den Logikparameter durch formale Hilfen bei der Aufgabenbearbeitung (drei Faktorstufen) und die heuristischen Parameter durch unterschiedliche inhaltliche Einbettungen der Selektionsaufgabe (ebenfalls drei Faktorstufen) zu beeinflussen. Eine Untersuchung an 125 Studierenden zeigte erwartungsgemäß selektive Effekte der beiden unabhängigen Variablen auf die Zielparameter auf. Das entsprechend restringierte verbundene Modell paßte gut auf die Daten aller $3 \cdot 3 = 9$ Versuchsbedingungen ($G^2_{(24)} = 29.27$, $p > .10$). Die Parameterschätzungen für *a* wuchsen monoton mit der Anzahl formaler Hilfestellungen. Die heuristischen Parameter *p* und $q_{\neg}$ konnten durch unterschiedliche inhaltliche Einbettungen der Aufgabe wahlweise erhöht oder gesenkt werden. Wählt man einen sozialen Standardvertrag als zu prüfende Regel („Wenn Du einen Vorteil erhältst, mußt Du auch die Kosten dafür tragen"), so läßt das den *a*-Parameter im Vergleich zur oben geschilderten abstrakten Selektionsaufgaben untangiert, erhöht jedoch die heuristischen Parameter *p* und $q_{\neg}$. Gibt man den Vpn dagegen einen vertauschten („switched") sozialen Vertrag als zu prüfende Regel („Wenn Du die Kosten trägst, mußt Du auch einen Vorteil dafür erhalten"), so reduziert dies *p*, und erhöht gleichzeitig *r*. Der Parameter *a* ist erneut nicht betroffen.

Das beobachtete Ergebnismuster entspricht exakt den Vorhersagen der Theorie sozialer Verträge (Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992), sieht man einmal davon ab, daß diese Theorie logisch-analytische Bearbeitungsstrategien ausklammert. Wenn man die Theorie sozialer Verträge so interpretiert, daß sie analytische Bearbeitungen neben inhaltsspezifischen Bearbeitungen zuläßt, ist das

Modell von Gleißner und Erdfelder (1994) mit dieser Theorie voll kompatibel. Allerdings macht das multinomiale Modell keine Aussagen darüber, durch welche experimentellen Manipulationen die heuristischen Parameter potentiell beeinflußbar sind. Soziale Verträge unterschiedlichen Typs bilden sicher eine Möglichkeit der Beeinflussung, aber nicht die einzige.

7.3.1.2 Propositionales Schlußfolgern

Rips (1983) hat eine der sogenannten regelbasierten Theorien propositionalen Schlußfolgerns vorgeschlagen („A Natural Deductive System", ANDS). Er nahm an, daß Menschen bei Schlußfolgerungsprozessen hauptsächlich auf 11 inhaltsfreie, abstrakte Regeln zugreifen, welche im wesentlichen eine Teilmenge aussagenlogisch korrekter Schlußfolgerungsregeln darstellen (beispielsweise der modus ponens). Mittels dieser Regeln wird versucht, Schlußfolgerung entweder ausgehend von den Prämissen („forward inference") oder ausgehend von der Konklusion zu beweisen („backward inference"). Scheitert der Ableitungsversuch, so wird die Schlußfolgerung als unkorrekt bewertet oder einfach geraten. Bewertungsfehler können aus der Nichtabrufbarkeit einer für einen Beweis wichtigen Schlußfolgerungsregel aus dem Gedächtnis resultieren.

Zur Prüfung seiner Theorie konstruierte Rips (1983) u.a. ein binäres MVB-Modell zu insgesamt 32 Schlußfolgerungsaufgaben. 36 Vpn mußten die 32 Implikationsbehauptungen als „notwendigerweise wahr" oder „nicht notwendigerweise wahr" klassifizieren. In seinem Modell wird die Wahrscheinlichkeit einer korrekten Antwort auf jedes der 32 Probleme als Funktion der Verfügbarkeitswahrscheinlichkeiten der für einen Beweis erforderlichen Schlußfolgerungsregeln — summiert über alle mit ANDS möglichen Beweiswege — ausgedrückt. Ist für jeden möglichen Beweisweg mindestens eine der erforderlichen Schlußfolgerungsregel nicht verfügbar, so wird angenommen, daß eine Vpn mit einer bestimmten Wahrscheinlichkeit korrekt rät. Die Verfügbarkeitswahrscheinlichkeiten der Regeln (p_j, j = 1, ..., 11) sollen dabei unabhängig von der zu beurteilenden Aufgabe sein. Darüber hinaus wurde angenommen, daß die Verfügbarkeit jeder Regel stochastisch unabhängig von der Verfügbarkeit anderer Schlußfolgerungsregeln ist. Trotz dieser strengen und theoretisch teilweise schwer motivierbaren Annahmen gelang mittels einer Kleinst-Quadrat-Anpassung der Modellgleichungen an die empirischen Lösungsraten eine befriedigende Datenanpassung. Ähnlich wie bei Evans (1977) wurden jedoch statistisch eindeutig interpretierbare Modellanpassungsstatistiken nicht präsentiert.[89] Auch eine Validierung der Modellparameter

[89]Rips (1983, S. 62) verwendete die Korrelation r = .933 zwischen beobachteten und erwarteten Lösungsraten als Maß der Anpassungsgüte und testete diese Korrelation gegen die Nullhypothese ρ = .000. Dies ist problematisch, da auch sehr hohe Korrelationen keine Aussage über die Modellgültig-

steht noch aus. Allerdings entsprachen die Schätzungen der Verfügbarkeitswahrscheinlichkeiten weitgehend der erwarteten Schwierigkeitsrangordnung der Schlußfolgerungsregeln (Rips, 1983, S. 62, Tabelle 5).

Klauer und Oberauer (1995) schlugen ein MVB-Modell zu einem anderen Untersuchungsparadigma vor. Sie überprüften mit Hilfe ihres Modells die Mental-Model-Theorie propositionalen Schlußfolgerns von Johnson-Laird und Byrne (1991). Probanden wurden mit einer Verknüpfung von zwei Elementaraussagen konfrontiert. Hierbei handelte es sich entweder um eine exklusive Disjunktion („Entweder *p* oder *q*"), eine inklusiven Disjunktion („*p* oder *q*"), eine Implikationsbehauptung („Wenn p dann *q*") oder einer Äquivalenzbehauptung („p genau dann wenn q"). Zusätzlich erhielten die Vpn den Wahrheitswert einer der beiden beteiligten Elementaraussagen. Die Vpn mußten dann angeben, ob diese Prämissen für die andere Elementaraussage (a) Wahrheit, (b) Falschheit oder (c) gar nichts implizieren. Diese drei Antwortoptionen ergeben zwei freie Kategoriewahrscheinlichkeiten pro Aufgabe. Gekreuzt mit vier Verknüpfungen (i = 1, ..., 4) und vier möglichen Wahrheitswertangaben pro Verknüpfung (j = 1: *p* ist wahr, j = 2: *p* ist falsch, j = 3: *q* ist wahr, j = 4: *q* ist falsch) ergeben sich $3 \cdot 4 \cdot 4 = 48$ unabhängige Kategoriewahrscheinlichkeiten insgesamt. Dies ist eine sehr reichhaltige Basis für die Modellbildung.

Zielsetzung des Modells von Klauer und Oberauer ist Separierung der Einflüsse von (a) Interpretationen der Verknüpfungsregel, (b) der Leichtigkeit des logischen Schlußfolgerungsprozesses und (c) Antworttendenzen auf das Antwortverhalten der Probanden. Für jede Aufgabe wird angenommen, daß Probanden mit einer verknüpfungsspezifischen Wahrscheinlichkeit c_i die logische korrekte Interpretation der Verknüpfung wählen und mit der komplementären Wahrscheinlichkeit $1-c_i$ einem Interpretationsmißverständnis unterliegen (Verwechslung von inklusiver und exklusiver Disjunktion sowie Verwechslung von Implikations- und Äquivalenzbehauptungen). Für die gewählte Interpretation wird sodann mit der Wahrscheinlichkeit $r_{i,j}$ die logisch richtige Antwort auf Verknüpfung i kombiniert mit Wahrheitswertangabe j erkannt. Mit der komplementären Wahrscheinlichkeit $(1 - r_{i,j})$ wird geraten. Für jede Verknüpfung i und jede der drei Antwortoptionen wird eine eigene Ratewahrscheinlichkeit angenommen (s-Parameter).

Insgesamt 35 Studierende beurteilten jeweils 64 Aufgaben des o.g. Typs. Die Anpassung des verbundenen Modells an die insgesamt 2240 Datenpunkte liefert die befriedigende Anpassungsstatistik $G^2_{(4)} = 9.25$. Die Parameterschätzungen zeigten erwartungsgemäß, daß die Implikation in nahezu der Hälfte aller Fälle im

---

keit zulassen. Beispielsweise bereitet es keine Schwierigkeiten, positive Werte für eine Variable X so zu wählen, daß X und X^2 eine Korrelation größer .93 aufweisen (vgl. Abschnitt 4.3.6).

Sinne einer Äquivalenzbehauptung fehlinterpretiert wird. Die Schätzungen der Logikparameter $r_{i,j}$ zeigten über die 16 Bedingungen hinweg ein Muster, das den Vorhersagen der Theorie mentaler Modelle propositionalen Schlußfolgerns (Johnson-Laird & Byrne, 1991) sehr gut entspricht. Insbesondere fielen die Logikparameter hypothesengemäß um so niedriger aus, je mehr mentale Modelle der Prämissen für eine logisch korrekte Beantwortung der Aufgabe erforderlich sind.

Die Arbeit und Klauer und Oberauer (1995) ist ein instruktives Beispiel für eine sehr weitreichende hypothesentestende Anwendung von MVB-Modellen. Die Autoren beschränkten sich nicht auf die Evaluation einiger Vorhersagen der Theorie. Sie wählten vielmehr einen Modellrahmen, der alle wesentlichen Elemente der zu prüfenden Theorie — Interpretation der logischen Verknüpfung, Leichtigkeit der logischen Evaluation und Antworttendenzen — beinhaltet. Das positive Gesamtergebnis ist demzufolge ein besonders überzeugendes Argument für die Adäquatheit der Mental-Model-Theorie propositionalen Schlußfolgerns.

7.3.1.3 Syllogistisches Schlußfolgern

Klauer, Musch und Naumer (1999) haben jüngst ein multinomiales Modell zum Glaubwürdigkeitseffekt (belief bias) beim syllogistischen Schlußfolgern vorgelegt. Hierunter versteht man die Tendenz, Syllogismen mit inhaltlich glaubwürdigen Konklusionen eher für gültig zu halten als Syllogismen mit inhaltlich unplausiblen Konklusionen. So wird z.B. folgender Syllogismus nahezu immer akzeptiert: „Alle Menschen sind Säugetiere und alle Säugetiere sind Lebewesen (Prämisse); folglich sind auch alle Menschen Lebewesen (Konklusion)". Seltener akzeptiert wird ein ebenfalls logisch gültiger Syllogismus mit unglaubwürdiger Konklusion, beispielsweise „Alle Lebewesen sind Säugetiere und alle Säugetiere sind Menschen (Prämisse); folglich sind auch alle Lebewesen Menschen (Konklusion)". Höhere Akzeptanzraten für Syllogismen mit plausiblen Konklusionen finden sich nicht nur — wie im Beispiel — bei logisch gültigen, sondern auch bei logisch ungültig Syllogismen. Im letztgenannten Falle ist der Effekt der Glaubwürdigkeit sogar stärker.

Klauer et al. betrachteten in ihrem verbundenen multinomialen Modell Akzeptanzraten für logisch gültige (i = 1) versus ungültige Syllogismen (i = 2) mit glaubwürdigen (j = b) versus unglaubwürdigen (j = u) Konklusionen, d.h. insgesamt vier verschiedene Versuchsbedingungen. Sie nahmen an, daß eine Vp mit einer versuchsbedingungsspezifischen Wahrscheinlichkeit $r_{i,j}$ den wirklichen logischen Status entdeckt und dann entsprechend ein korrektes Gültigkeitsurteil (bei i = 1) bzw. ein korrektes Ungültigkeitsurteil (bei i = 2) abgibt. Wird mit der komplementären Wahrscheinlichkeit der logische Status nicht entdeckt, so urteilt die Vp in Abhängigkeit von nichtlogischer Evidenz. Glaubwürdige Konklusionen

werden mit einer größeren Wahrscheinlichkeit a_b akzeptiert, unglaubwürdige mit der kleineren Wahrscheinlichkeit a_u. Beide Parameter können als Antworttendenz- oder Rateparameter aufgefaßt werden, welche u.a. durch die Glaubwürdigkeit der Konklusion beeinflußt werden.

Damit das Sechs-Parameter-Modell (vier Logikparameter $r_{i,j}$ und zwei Antworttendenzparameter a_b bzw. a_u) bei nur vier freien Kategoriewahrscheinlichkeiten identifizierbar oder gar testbar wird, muß man den Modellparametern natürlich Restriktionen auferlegen. Glücklicherweise lassen sich aus den konkurrierenden Theorien zum Glaubwürdigkeitseffekt relativ stringent Restriktionen herleiten, die zu identifizierbaren und z.T. sogar testbaren Submodellen führen (Klauer et al., 1999). Einige Theorien können auf diese Weise anhand einer Reanalyse publizierter Daten ausgeschlossen werden. Die meisten Theorien führen jedoch zu saturierten Modellen mit vier Parametern. Wenn man zwischen diesen Theorien empirisch differenzieren will, ist die Datenstruktur des Designs mit vier Versuchsbedingungen nicht reichhaltig genug.

Um dieses Problem zu lösen, wählten Klauer und Mitarbeiter den vierten Weg der Behandlung von Identifizierbarkeitsproblemen: den Weg der Erweiterung des empirischen Designs (vgl. Abschnitt 4.6). Der Grundgedanke ihrer Erweiterung ist, daß die Logikparameter durch eine Manipulation von Antworttendenzen nicht beeinflußt werden sollten. Präsentiert man Vpn also beispielsweise drei Aufgabenmengen mit bekanntermaßen unterschiedlichen Anteilen gültiger versus ungültiger Syllogismen (z.B. 1:5, 1:1 und 5:1), so kann man die Logikparameter für die drei Aufgabenmengen gleichsetzen. Unterschiedliche Akzeptanzraten sollten sich ausschließlich auf die Antworttendenzparameter auswirken. Aus diesen Annahmen resultiert ein Zehn-Parameter-Modell (4 Logikparameter plus $3 \cdot 2 = 6$ Antworttendenzparameter) für $3 \cdot 4 = 12$ Versuchsbedingungen. Dieses Modell ist identifizierbar und mit $df = 12–10 = 2$ auch testbar. Klauer et al. gingen jedoch noch einen Schritt weiter und forderten, daß sich die Antworttendenzparameter multiplikativ zerlegen lassen in einen Effekt des Anteils gültiger Syllogismen in der Aufgabenmenge (α_i, $i = 1, 2, 3$) und einen Effekt der Glaubwürdigkeits- versus Unglaubwürdigkeitsmanipulation (β_j, $j \in \{b, u\}$). Damit diese Parameter identifizierbar werden, muß wie im BTL-Modell ein beliebiger Parameter des multiplikativen Modells fixiert werden (vgl. Abschnitt 4.4 und 4.5). Klauer et al. wählten für die glaubwürdigen Konklusionen $\beta_b = 1$ und forderten für die unglaubwürdigen Konklusionen $0 \leq \beta_u \leq 1$. Der Vorteil dieser Festlegungen ist, daß der Parameter β_u nunmehr als „reines" Maß des Glaubwürdigkeitseffekts auf die Antworttendenzen interpretiert werden kann: Je näher β_u bei null liegt, um so stärker ist dieser Effekt.

Die Zusatzannahmen führen die sechs Antworttendenzparameter des Modells auf vier grundlegende Parameter zurück, nämlich $\alpha_1, \alpha_2, \alpha_3$ und β_u. Darüber hinaus kann die Ordnungsrestriktion $\alpha_1 \leq \alpha_2 \leq \alpha_3$ für den Effekt des Anteils gülti-

ger Syllogismen zugrunde gelegt werden kann ($i = 1$ steht für den geringsten, $i = 3$ für den größten Anteil gültiger Syllogismen). Dieses Basismodell A ist folglich mit $df = 12 - 8 = 4$ testbar. Werden nicht nur Syllogismen mit glaubwürdigen und unglaubwürdigen Konklusionen dargeboten, sondern zusätzlich Syllogismen mit inhaltlich neutralen Konklusionen, kann man das Modell in kanonischer Weise auf ein Basismodell B mit sechs Logikparametern, drei α- und zwei β-Parametern verallgemeinern, welches mit $df = 3 \cdot 6 - 11 = 7$ testbar ist.

Beide Basismodelle A und B paßten sehr gut zu den Daten mehrerer Studien von Klauer et al., darunter auch World-Wide-Web-Studien mit sehr großen Stichprobenumfängen. Die α- und β-Parameter der Modelle unterschieden sich in der erwarteten Richtung, die β-Parameter allerdings nicht immer signifikant. Dies bedeutet, daß Glaubwürdigkeitseffekte sich nicht zwangsläufig auf der Ebene von Antworttendenzen manifestieren müssen. Erwartungsgemäß waren die Logikparameter bei Logikexperten generell höher ausgeprägt als bei Vpn, die angaben, kein Logikvorwissen zu besitzen (im folgenden: Logiknovizen). Die vier Logikparameter ließen sich sowohl bei Experten als auch bei Novizen untereinander nicht gleichsetzen. Bei Logiknovizen wurde oftmals eine Interaktion dahingehend deutlich, daß die Logikparameter besonders groß werden, wenn logische Gültigkeit und Glaubwürdigkeit konform gehen: gültige Syllogismen mit glaubwürdigen Konklusionen und ungültige Syllogismen mit unglaubwürdigen Konklusionen erhalten hohe *r*-Parameterschätzungen (Experimente 1, 2, 5 und 6). Dieses Ergebnismuster fand sich allerdings nicht generell, sondern nur dann, wenn die gültigen Syllogismen durch genau ein repräsentierendes Modell (z.B. ein einziges Venn-Diagramm) erschöpfend beschrieben werden können. Andernfalls waren die *r*-Parameter für (a) gültige Syllogismen mit glaubwürdigen Konklusionen, (b) gültige Syllogismen mit unglaubwürdigen Konklusionen und (c) ungültige Syllogismen mit glaubwürdigen Konklusionen nicht signifikant verschieden voneinander; nur der Logikparameter für ungültige Syllogismen mit unglaubwürdigen Konklusionen war erneut höher (Experiment 3 und 4). Tendenziell ähnliche Ergebnismuster fanden sich bei den Logikexperten. In einigen Experimenten ließen sich deren Logikparameter allerdings multiplikativ zerlegen in einen Haupteffekt des logischen Status (gültige Syllogismen erhielten höhere *r*-Schätzwerte als ungültige) und einen Haupteffekt der inhaltlichen Glaubwürdigkeit (Syllogismen mit glaubwürdigen Konklusionen erhielten höhere *r*-Schätzwerte als Syllogismen mit unglaubwürdigen Konklusionen).

Klauer et al. zeigten, daß das berichtete Ergebnismuster mit keiner der bislang vorgeschlagenen Theorien zum Glaubwürdigkeitseffekt konsistent ist. Sie schlugen als Alternative eine modifizierte Theorie mentaler Modelle vor, die auf der Kernannahme basiert, daß Syllogismen mit glaubwürdigen Konklusionen positiv getestet werden, Syllogismen mit unglaubwürdigen Konklusionen dagegen

negativ. Bei positiven Tests wird versucht, ein kohärentes mentales Modell der Prämissen und der Konklusion ausgehend von der Konklusion zu konstruieren, bei negativen Tests dagegen ausgehend von der negierten Konklusion (Klayman & Ha, 1987). Gelingt ein positiver Test, so belegt das die Gültigkeit des Syllogismus, wenn — wie in den Experimenten 1, 2, 5 und 6 — nur ein repräsentierendes Modell zu den Prämissen existiert. Gelingt ein negativer Test, so widerlegt das die Gültigkeit des Syllogismus in jedem Fall. Fehlschläge in der Modellkonstruktion liefern dagegen grundsätzlich mehrdeutige Evidenz. Wie Klauer et al. zeigten, korrespondiert das beobachtete Muster der *r*-Parameterschätzungen für Logiknovizen den Prognosen zu allen Experimenten nahezu perfekt. Weitere Vorhersagen der Theorie wurden ebenfalls bestätigt (Experiment 5-7).

Die Arbeit von Klauer und Mitarbeitern ist ein eindrucksvolles Beispiel für eine hypothesentestende Anwendung von MVB-Modellen, die belastende Evidenz für alle zum Zeitpunkt der Modellanwendung vorliegenden Theorieansätze erbringt. Sie macht damit den Weg frei für neue Theorien, welche auf Befunden aufbauen, die ohne den verwendeten MVB-Modellansatz unentdeckt geblieben wären. MVB-Modellierung kann also aus substanzwissenschaftlicher Perspektive nicht nur ein Entscheidungsinstrument, sondern auch ein Innovationsinstrument sein, ohne das bestimmte neue Wege der Theoriebildung gar nicht beschritten werden können

7.3.2 Urteilen und Entscheiden

Ein kniffliges Problem der Urteils- und Entscheidungsforschung ist die Diagnose von verwendeten Entscheidungs- oder Urteilsheuristiken auf der Grundlage des beobachteten Entscheidungsverhaltens. Typischerweise kann ein bestimmtes Verhalten immer durch mehrere Heuristiken annähernd gleich gut erklärt werden (Rieskamp & Hoffrage, 1999). Auch die Introspektionsmethode ist aus den bereits in Kapitel 2 erörterten Gründen von fraglichem Wert.

Bröder (1999) hat mehrere Vorschläge unterbreitet, wie dieses Problem einer Lösung zugeführt werden kann. Einer dieser Vorschläge beruht auf einem MVB-Modell, das auf vier aufeinander folgende dichotome Entscheidungen anzuwenden ist (z.B. „Betrachten Sie die Aktienpaare VW vs. Daimler, Deutsche Bank vs. Commerzbank, SAP vs. Siemens, Karstadt vs. Metro. Welche Aktie eines jeden Paares präferieren Sie?”). Die Entscheidungen sind auf der Grundlage binärer 0-1-Merkmalsvariablen zu treffen, auf denen jede der Entscheidungsalternativen durch eine bestimmte Ausprägung gekennzeichnet ist [z.B. Dividende größer als 2% (ja / nein), Gewinnsteigerung im letzten Jahr (ja / nein) etc.]. Alle zu vergleichenden Entscheidungsheuristiken müssen als Regeln formulierbar sein, aus denen eindeutig hervorgeht, welche Entscheidung bei welchem Paar von Merk-

malsmustern zu treffen ist. Der Grundgedanke des Modells von Bröder (1999) ist dem Modell von Levine (1966) verwandt: Die Vpn wählen (nicht notwendigerweise bewußt) eine bestimmte Entscheidungsheuristik H_i mit Wahrscheinlichkeit $p(H_i)$. Die zugrundeliegende Entscheidungsregel wird dann über alle vier Entscheidungen konsistent beibehalten. Die Modellparameter sind also die Wahlwahrscheinlichkeiten $p(H_i)$. Sofern das Modell empirisch scheitert, sind prinzipiell zwei Klassen von Modifikationen denkbar: Die Anzahl der modellierten Entscheidungsheuristiken wird erhöht oder die Wahlen werden — ähnlich wie im Latent-Class-Modell zum Levine-Experiment (vgl. Abschnitt 2.5) — nicht deterministisch, sondern probabilistisch an die gewählten Entscheidungsheuristiken geknüpft.

7.4 Sprachpsychologische Anwendungen

Massaro (z.B. 1987) hat ein „fuzzy logic model of perception" (FLMP) vorgeschlagen, um den Einfluß zweier unabhängig voneinander manipulierter physikalischer Attribute gesprochener oder geschriebener Sprache auf die Wahrnehmung eines Lautes oder eines Buchstabens zu beschreiben. Im Falle gesprochener Sprache kann das erste Attribut beispielsweise akustischer Natur sein (Faktor 1: Grad des Anstiegs des zweiten und dritten Formantenansatzes im Sonagramm) und das zweite Attribut optischer Natur (Faktor 2: Grad des initialen Lippenschlusses des Sprechers). Der Rezipient bekommt alle möglichen Kombinationen der Faktorstufen beider Faktoren präsentiert und muß jeden Reiz daraufhin beurteilen, ob er eher der Lautkombination [ba] oder der Lautkombination [da] entspricht.

Bekanntlich wird die Lautwahrnehmung nicht allein durch die auditorische Komponente, sondern auch durch die visuelle Komponente wesentlich mitbestimmt (McGurk & MacDonald, 1976). Neuere Untersuchungen, die auf der Positronenemissionstomographie (PET-Methode) basieren, zeigen sogar, daß bei Personen, die von Geburt an taub sind, die sekundären auditorischen Regionen des Cortex durch visuelle Zeichensprache in ähnlicher Weise aktiviert werden können wie durch gesprochene Sprache bei Hörenden (Nishimura et al., 1999). Doch wie wirken akustische und visuelle Informationen bei Hörenden zusammen? Nach dem FLMP von Massaro und Mitarbeitern wird jedes physikalische (d.h. akustische oder optische) Attribut von Sprachinformation zunächst *unabhängig von anderen Attributen* daraufhin analysiert, wie gut es mit dem Prototyp einer bestimmten sprachlichen Einheit übereinstimmt (Schritt 1: Merkmalsevaluation). Diese „Übereinstimmungsgrade" werden als „fuzzy truth values" interpretiert, die zwischen 0 (keine Übereinstimmung) und 1 (perfekte Übereinstimmung) variieren können. So haben etwa anfänglich komplett geschlossene Lippen den Wahrheitswert 1 und anfänglich weit geöffnete Lippen den Wahrheitswert 0 für die Lautkombination [ba] im Kontrast zu [da]. Jeder Stufe *i* des Faktors „Lippenschluß"

korrespondiert also ein Parameter $\alpha_i \in [0, 1]$ und analog jeder Stufe *j* des Faktors „Anstieg des zweiten und dritten Formantenansatzes" einem Parameter $\beta_j \in [0, 1]$; beide Parameter indizieren jeweils die Nähe des betreffenden physikalischen Attributs zum Prototyp von [ba] bzw. die Entfernung vom [da]-Prototypen. Wenn Vpn sprachliche Kategorisierungen auf der Grundlage von Kombinationen optischer und akustischer Informationen vornehmen sollen, muß im zweiten Schritt noch das Produkt der entsprechenden Parameter gebildet werden (Schritt 2: Merkmalsintegration). Das Produkt $\alpha_i \cdot \beta_j$ entspricht dem „fuzzy truth value" des Prototyps von [ba], wenn die *Konjunktion* beider physikalischer Attribute gegeben ist. Analog ist $(1-\alpha_i)\cdot(1-\beta_j)$ der „fuzzy truth value" des Prototyps von [da] bei *Konjunktion* der physikalischen Attribute *i* und *j*. Von der Größe dieser Produktwerte — genauer: vom Verhältnis $\alpha_i \cdot \beta_j$ / $(\alpha_i \cdot \beta_j + (1-\alpha_i)\cdot(1-\beta_j))$ — hängt schließlich die relative Häufigkeit der [ba]-Klassifikationen ab (Schritt 3: Musterklassifikation).

Crowther, Batchelder und Hu (1995) haben die Interpretation der Modellparameter in Termini von „fuzzy truth values" aus meßtheoretischer Sicht kritisiert, da das FLMP für zwei Faktoren und Kategorien ohne zusätzliche Restriktionen nicht identifizierbar ist. Die Modellparameter bzw. ihre Schätzer sind folglich nicht eindeutig durch die Vp-Urteile festgelegt. Anders ausgedrückt: Jede Aussage über den „fuzzy truth value" eines Merkmals wäre empirisch nicht bedeutsam, die sie durch erlaubte Reskalierungen der Parameterwerte immer in eine falsche Aussage überführt werden kann.

Als Alternative haben Crowther et al. (1995) eine Interpretation der Modellparameter in Termini von latenten Wahrscheinlichkeiten vorgeschlagen. Beispielsweise wäre in dieser Interpretation α_i die Wahrscheinlichkeit, daß beim *i*-ten visuellen Reiz auf das Vorliegen des Merkmals „Lippenschluß" erkannt wird. „Erkannt" meint hier natürlich einen automatischen Prozeß, der nicht notwendig einem bewußten Erkennen oder Wahrnehmen des Merkmals entsprechen muß. Wie Crowther et al. (1995) zeigten, entsprechen die Modellgleichungen des FLMP in dieser Interpretation immer einfachen parametrisierten multinomialen Modellen, unabhängig davon, für wieviel Faktoren und Antwortkategorien das Modell formuliert wird. Für zwei Faktoren und zwei Antwortkategorien ist das probabilistisch reinterpretierte FLMP mit Parametern α_i und β_j beispielsweise dem Rasch-Modell mit reellwertigen Parametern ξ_i und ε_j formal datenäquivalent, wenn man letzteres auf die Wahrscheinlichkeit einer bestimmten Kategorisierung (z.B. [ba]) bei Faktorstufenkombination (*i, j*) anwendet (vgl. Abschnitt 5.4.2.3):

$$p([ba]-\text{Urteil}) = \frac{e^{(\xi_i - \varepsilon_j)}}{1+e^{(\xi_i - \varepsilon_j)}} = \frac{\alpha_i \cdot \beta_j}{\alpha_i \cdot \beta_j + (1-\alpha_i)\cdot(1-\beta_j)}\ , \qquad (7.4\text{-}1)$$

$$\text{wobei } \alpha_i = (1+e^{-\xi_i})^{-1} \text{ und } \beta_j = (1+e^{\varepsilon_j})^{-1}\ .$$

Für zwei Faktoren und vier Antwortkategorien entspricht das FLMP einem einfachen binären MVB-Modell mit den Modellgleichungen:

$$\begin{aligned} p_{1|i,j} &= \alpha_i \cdot \beta_j \\ p_{2|i,j} &= \alpha_i \cdot (1-\beta_j) \\ p_{3|i,j} &= (1-\alpha_i) \cdot \beta_j \\ p_{4|i,j} &= (1-\alpha_i) \cdot (1-\beta_j) \end{aligned} \qquad (7.4\text{-}2)$$

Hierbei ist $p_{k\,|\,i,j}$ die Wahrscheinlichkeit, daß bei Präsentation der Faktorstufenkombination (i, j) die Antwortkategorie $k \in \{1, 2, 3, 4\}$ gewählt wird (Batchelder & Crowther, 1997; Crowther et al., 1995). Wie man sieht, nimmt das Modell an, daß die vier Antwortkategorien ebenfalls in einem zweidimensionalen Raum angeordnet werden können: Die Wahl zwischen den Antwortkategorien 1 versus 2 und 3 versus 4 kann nur durch den zweiten Merkmalsentdeckungsprozeß, die zwischen 1 versus 3 und 2 versus 4 nur durch den ersten Prozeß beeinflußt werden. Die Beeinflussung erfolgt unabhängig voneinander, d.h. der Effekt eines Merkmalsentdeckungsprozesses ist unabhängig vom Erfolg versus Mißerfolg des anderen Prozesses (Prozeßunabhängigkeit). Zusätzlich wird angenommen, daß jeder Merkmalsentdeckungsprozeß selektiv durch genau einen der experimentellen Faktoren beeinflußt werden kann (Kontextunabhängigkeit). Verallgemeinerungen des Modells auf höherdimensionale Antworträume (mit 8, 16 usw.) Antwortkategorien liegen auf der Hand.

Batchelder und Crowther (1997) haben die strengen Unabhängigkeitsannahmen, die in das FLMP gemäß Gleichungssystem (7.4-2) eingehen, liberalisiert und auf diese Weise eine sehr allgemeine Familie von binären MVB-Modellen für zweifaktorielle Designs mit vier Antwortkategorien entwickelt. Sie zeigten empirisch, daß das FMLP die Klassifikation von Lauten auf der Grundlage der beiden akustischen Attribute „Vokaldauer" (Faktor 1) und „Frequenz des ersten Formantenansatzes" (Faktor 1) nicht adäquat beschreiben kann, während ein verallgemeinertes Modell, das an der Prozeßunabhängigkeitsannahme festhält, jedoch die Kontextunabhängigkeitsannahme aufgibt, eine adäquate Datenanpassung liefert. Dieses allgemeinere Modell hat nahezu die gleichen Modellgleichungen wie das FLMP gemäß Gleichungssystem (7.4-2); der Unterschied besteht lediglich darin, daß die Prozeßparameter nun zwei Indizes ($\alpha_{i,j}$ und $\beta_{i,j}$) tragen. Somit kann jeder Prozeß durch jeden der beiden Faktoren beeinflußt werden; auf die Annahme einer selektiven Beeinflussung der Prozesse durch einzelne Faktoren wird verzichtet.

Während das ursprüngliche FLMP somit auf die Analyse separabler Dimensionen im Sinne Garners (1974) beschränkt ist, erlaubt das verallgemeinerte Modell von Batchelder und Crowther (1997) auch Analysen integraler Dimensionen (Garner, 1974), deren Ausprägungen nicht unabhängig voneinander, sondern kon-

textabhängig verarbeitet werden. Bei vier Antwortkategorien pro Stimulus ist dieses Modell identifizierbar und testbar. Es erlaubt somit den direkten empirischen Vergleich mit einem Modell, das separable Dimensionen postuliert.

7.5 Wahrnehmungspsychologische Anwendungen

In das FLMP von Massaro (1987) und dessen Verallgemeinerung durch Batchelder und Crowther (1997) gehen keine Annahmen ein, die spezifisch für die Sprachwahrnehmung sind. Vielmehr handelt es sich um Informations-Integrations-Modelle im Sinne von Norman Anderson (1982b), die prinzipiell auf beliebige multifaktoriell beeinflußte Wahrnehmungsleistungen anwendbar sind; vorauszusetzen ist nur, daß die Wahrnehmungsprozesse in Kategorisierungen einmünden. Damit ist diese Modellklasse für die Wahrnehmungspsychologie generell von Nutzen, nicht nur für die Sprachwahrnehmung. Für die Bedeutung der verallgemeinerten FLMP-Versionen von Batchelder und Crowther (1997), die Kontextabhängigkeiten zulassen, spricht die Tatsache, daß gegen Informations-Integrations-Modelle der visuellen Wahrnehmung, die auf der Annahme separabler Dimensionen aufbauen, kritische Einwände erhoben wurden (Gigerenzer & Richter, 1990). Die Modellklasse von Batchelder und Crowther (1997) eröffnet die Chance, auch in diesem Bereich eine Klärung durch statistische Modellvergleiche herbeizuführen.

7.5.1 Reizerkennung und Raten

Auch andere MVB-Modelle können sinnvoll in die wahrnehmungspsychologische Forschung eingebracht werden. Beispielsweise hat Broadbent (1967) mehrere Modelle für Reizerkennungsversuche vorgeschlagen und auf ihre Eignung zur Erklärung des Worthäufigkeitseffekts (word-frequency effect) hin überprüft. Der Worthäufigkeitseffekt besteht darin, daß häufige Wörter einer Sprache unter erschwerten Wahrnehmungsbedingungen (z.B. tachistoskopische Darbietung für nur wenige Millisekunden) häufiger richtig erkannt werden als seltene Wörter. Eine einfache Erklärungsmöglichkeit hierfür bietet z.B. das „pure guessing model“, welches im Kern eine Verallgemeinerung des Ein-Hochschwellen-Modells (vgl. Abschnitt 3.5.1) auf n-Alternativen-Multiple-Choice-Tests darstellt. Wenn ein Reiz S_i, $i = 1, ..., n$, dargeboten wird, wird er nach diesem Modell entweder mit Wahrscheinlichkeit r_i erkannt — was zu einer korrekten Antwort R_i führt — oder mit der komplementären Wahrscheinlichkeit nicht erkannt, was zum Erraten der Antwort R_j, $j = 1, ..., n$, mit Wahrscheinlichkeit b_j führt. Selbst wenn die Erkennenswahrscheinlichkeiten für häufige und seltene Wörter gleich sind, kann ein Worthäufigkeitseffekt also allein aufgrund größerer Ratewahrscheinlichkeiten für häufige Wörter resultieren.

Broadbent (1967) hat Gründe genannt, die gegen die Adäquatheit des geschilderten Modells sprechen und ein alternatives Modell (Broadbents Modell 4) vorgeschlagen. Anpassungen der Modelle an Daten hat er allerdings nicht vorgenommen; er hat lediglich Befundmuster daraufhin angeschaut, ob sie im Lichte der Modelle plausibel sind. Später zeigte Nakatani (1970), daß Broadbents Modell 4 dem „pure guessing model" datenäquivalent ist, so daß die Modelle anhand einfacher Reiz-Urteils-Datenmatrizen gar nicht gegeneinander testbar sind. Dies macht deutlich, wie wichtig es sein kann, zwischen inhaltlich unterschiedlich motivierten Modellen nicht ohne einen analytischen Vergleich der Modellgleichungen (vgl. Nakatani, 1970) oder einen empirischen Vergleich der Modellanpassungsstatistiken eine Entscheidung zu treffen. Verzichtet man darauf, so ist immer die Gefahr ein Fehlbewertung der Modellgüte gegeben.

7.5.2 Parallele versus serielle Reizverarbeitung

Die Frage, ob verschiedene kognitive Prozesse seriell oder parallel ablaufen, ist seit langem Gegenstand stochastischer Modellierungsansätze (Luce, 1986; Townsend & Ashby, 1983). Die bisherige Forschung hat gezeigt, daß eine Entscheidung zwischen seriellen und parallelen Verarbeitungsmodellen wohl nur im Rahmen komplexer experimenteller Designs und auf dem Hintergrund mehr oder minder strenger Annahmen zu den Mechanismen serieller und paralleler Prozesse möglich ist. Ohne solche Zusatzannahmen sind serielle und parallele Modelle anhand von Antwortzeitverteilungen empirisch nicht gegeneinander testbar (Vorberg, 1977, zitiert nach Luce, 1986, S. 460).

Brown (1998) hat einen MVB-Ansatz der Modellierung serieller und paralleler Prozesse bei der Objektwahrnehmung vorgestellt, der konzeptuelle Gemeinsamkeiten mit dem Parallel-Seriell-Testparadigma (PST) von Townsend und Ashby (Meyer, 1994; Townsend & Ashby, 1983, Kap. 13) aufweist, sich im Unterschied diesem Ansatz allerdings nicht auf die Modellierung von mittleren Antwortzeiten, sondern auf die Modellierung von Antwortwahrscheinlichkeiten bezieht. Vpn bekommen kurzfristig und kontrastschwach zwei Objekte — z.B. Buchstaben — rechts und links von einem Fixationspunkt dargeboten, wovon entweder keiner, der linke, der rechte oder beide Buchstaben in eine zuvor definierte Klasse (z.B. Kleinbuchstaben oder „d"-Buchstaben mit zwei Strichen) fallen. Dies sind die vier Versuchsbedingungen des experimentellen Paradigmas. Die Vp sollen in jeder Versuchsbedingung beurteilen, ob von den schwer erkennbaren Buchstaben keiner, der linke, der rechte oder beide Buchstaben zur Zielklasse gehören. Insgesamt liegen somit $4 \cdot (4-1) = 12$ freie Kategoriewahrscheinlichkeiten vor, was eine günstige Ausgangsbasis für die MVB-Modellierung paralleler und serieller Prozesse darstellt.

Der Grundgedanke der von Brown (1998) analysierten parallelen Modelle ist, daß zwei „Objektdetektoren" gleichzeitig und unabhängig voneinander arbeiten, wenn auch nicht notwendigerweise gleichermaßen effizient. Die Detektoren erkennen mit den Wahrscheinlichkeiten G bzw. H Zielobjekte in den beiden Positionen korrekt, wobei G und H natürlich von der Darbietungsbedingungen abhängen. Liegt auf einer oder beiden Seiten kein Zielobjekt vor, kann es mit den Wahrscheinlichkeiten L bzw. R zu einem Verwechslungsfehler und damit zu einem falschen Erkennungsurteil kommen. Auch L und R hängen natürlich von den Darbietungsbedingungen ab.

Im Gegensatz dazu wird beim einfachsten seriellen Modell angenommen, daß die Detektoren sequentiell nacheinander arbeiten, und zwar mit Wahrscheinlichkeit P zuerst der linke und mit der komplementären Wahrscheinlichkeit $(1–P)$ zuerst der rechte. Die Objekterkennungs- und Verwechslungswahrscheinlichkeiten G_i und H_i bzw. L_i und R_i hängen hierbei natürlich davon ab, ob der betreffende Detektor als erster arbeitet (höhere Erfolgswahrscheinlichkeit) oder erst danach zum Zuge kommt (geringere Erfolgswahrscheinlichkeit). Im Kern bedeutet das, daß — anders als beim parallelen Modell — Erfolg beim rechten und beim linken Objekt nicht mehr unabhängig voneinander sein muß.

Die beiden geschilderten Modelle sowie zwei leicht modifizierte parallele und serielle Modelle wurden von Brown (1998) anhand des AIC-Kriteriums evaluiert (vgl. Abschnitt 4.3.6). Die beiden parallelen Modelle und ein serielles Modell lieferten adäquate Datenanpassungsergebnisse, d.h. ihre AIC-Statistiken gemäß Gleichung (4.3.6-2) waren negativ. Obwohl somit eines der seriellen Modelle durch die Daten gut gestützt wird, präferierte Brown (1998) dennoch eindeutig die parallelen Modelle. Er begründete dies wie folgt: Da Datensätze, die einem parallelen Modell folgen, im Regelfall auch gut durch das parameterreichere serielle Modell angepaßt werden können, während die Umkehrung nicht gilt, spricht insgesamt mehr für die Annahme, daß die Daten einem parallelen Verarbeitungsmodell folgen.

Validierungsstudien, in denen die räumliche Distanz der beiden Buchstaben und ihre Darbietungsdauer variiert wurde, favorisieren darüber hinaus insbesondere das „gestaltparallele Modell" von Brown (1998), das zusätzlich zu den beiden parallel arbeitenden Objektdetektoren des einfachen parallelen Modells (siehe oben) einen „Gestaltdetektor" für die simultane Entdeckung zweier räumlich versetzter Zielobjekte vorsieht.

Die Arbeit von Brown fällt ebenfalls in die Kategorie theorievergleichender Anwendungen von MVB-Modellen. Sie macht ein weiteres Problem deutlich, das insbesondere dann greift, wenn die Modelle zu verschiedenen psychologischen Theorien entweder in einer Obermodell-Submodell-Relation zueinander stehen (hierarchische Modelle) oder wenn aufgrund unterschiedlicher Parameteranzahl

der Modelle Asymmetrien derart vorliegen, daß ein Modell A Datensätze, die einem Modell B folgen, faßt immer gut anpassen kann, während die Umkehrung nicht gilt. Die reinen Modellanpassungsstatistiken sind in solchen Fällen offenbar nicht hilfreich, denn sie begünstigen immer das parameterreichere Modell. Auch χ^2-Differenzentests (bei hierarchischen Modellen) oder das AIC- bzw. BIC-Kriterium sind nur bedingt von Nutzen, weil das allgemeinere Modell (mit mehr Parametern) natürlich nicht automatisch dadurch invalidiert wird, daß es keine deutlich bessere Datenanpassung als das parameterärmere Modell liefert: Es ist möglich, daß das allgemeinere Modell zwar zutrifft, die Parameterwerte jedoch für das parameterärmere Modell „günstig" liegen, so daß auch mit diesem eigentlich unzutreffenden Modell eine gute Datenanpassung gelingt und die χ^2-Differenz zum allgemeineren Modell gering ausfällt. Browns (1998) Entscheidung gegen das serielle Modell ist also nicht wirklich empirisch begründet, auch wenn seine Argumentation diesen Eindruck erweckt. Sie reflektiert vielmehr eine Voreingenommenheit gegen parameterreiche Modelle, insofern von ihnen implizit nicht nur ihre Gültigkeit verlangt wird, sondern darüber hinaus die Gültigkeit in Verbindung mit Parameterwerten, welche schlecht durch parameterärmere Konkurrenzmodelle approximierbar sind.

Der einzig gangbare Ausweg aus diesem Dilemma ist ein systematisches Validierungsprogramm zum parameterreicheren Modell, in dem möglichst alle Parameter selektiv beeinflußt werden sollten. Gelingt dies, so werden sich zwangsläufig Datenstrukturen ergeben, die das parameterärmere Modell nicht mehr anpassen kann. Man entscheidet sich dann für das allgemeinere Modell. Gelingt es nicht, so ist das allgemeinere Modell mangels erfolgreicher Validierung nicht brauchbar. Für das von Brown (1998) betrachtete Paradigma fehlen solche Studien noch.

7.5.3 Illusorische Konjunktionen bei der Objekterkennung

Wenn Objekte, die durch bestimmte Merkmalskombinationen auf zwei oder mehr perzeptuellen Dimensionen (z.B. Form und Farbe) gekennzeichnet sind, extrem kurzfristig dargeboten werden, kommt es oftmals zu sogenannten „illusorischen Konjunktionen". Beispielsweise kann eine Vp angeben, ein blaues X gesehen zu haben, obwohl nur rote X-Buchstaben und blaue C-Buchstaben tachistoskopisch dargeboten wurden. Treisman und Schmidts (1982) Merkmalsintegrationsmodell erklärt dies durch einen zweistufigen Prozeß. Im ersten, sehr schnell ablaufenden Stadium werden die im Stimulus vorhandenen Farben und Formen automatisch und parallel zueinander identifiziert. Im darauf folgenden zweiten Stadium werden die Merkmale sodann zu „wahrgenommenen Objekten" integriert, was kontrollierte Aufmerksamkeitszuwendung und relativ viel Zeit erfordert. Wird der

Verarbeitungsprozeß im Verlaufe des zweiten Stadiums unterbrochen, so kann es zu illusorischen Merkmalskonjunktionen kommen. Diese sollten in dem Sinne „zufällig" sein, daß sie nicht von physikalischen Eigenschaften der verknüpften Merkmale — etwa von ihrer Ähnlichkeit oder ihrer räumlichen Distanz im Stimulusdisplay — abhängen.

Ashby, Prinzmetal, Ivry und Maddox (1996) stellten dem Merkmalsintegrationsmodell ein Lokalisationsunsicherheitsmodell gegenüber, das die räumliche Nähe von Merkmalen als wesentliche Determinante von illusorischen Konjunktionen ausweist. Nach diesem Modell werden ebenfalls für alle Objekte eines visuellen Stimulusdisplays Forminformationen und Farbinformationen parallel extrahiert. Im Einklang mit neuroanatomischen und neurophysiologischen Befunden werden Form- and Farbextraktionsprozesse als unabhängig voneinander angenommen. Anders als im Merkmalsintegrationsmodell führt eine frühe Unterbrechung der Reizverarbeitung nicht zu zufälligen Merkmalskonjunktionen; vielmehr werden immer die Farb- und Formmerkmale verknüpft, welche die geringste wahrgenommene Distanz zueinander aufweisen. Die perzeptuelle Lokalisation eines Merkmals in einem zweidimensionalen Stimulusdisplay ist in der Frühphase der Reizverarbeitung mit Unsicherheit behaftet. Sie folgt einer bivariaten Normalverteilung mit Korrelation null und gleicher Streuung σ in horizontaler und vertikaler Richtung. Die objektive Lokalisation in der Horizontalen und in der Vertikalen legt dabei die beiden Mittelwerte fest. Somit ist die subjektive Lokalisation eines Merkmals nur durch einen freien Parameter gekennzeichnet, nämlich durch die Streuung σ um die objektive Lokalisation. Diese Streuung sollte z.B. durch längere Darbietungszeiten, durch foveale anstelle parafovealer Darbietung und durch Aufmerksamkeitsfokussierung auf das Zielobjekt reduzierbar sein.

Beide Modelle kann man mit einfachen Ratemodellen kontrastieren. Gelingt für ein Zielobjekt sowohl der Form- als auch der Farbextraktionsprozeß, so wird nach einem Ratemodell auf jeden Fall das korrekte Objekt identifiziert, unabhängig davon, ob auch Form- und Farbmerkmale anderer Objekte im Display identifiziert wurden. Ein zusätzlicher Merkmalsintegrationsprozeß ist also nicht notwendig; Merkmalsverwechslungen verschiedener Objekte werden prinzipiell ausgeschlossen. Sollte ein Form- oder Farbextraktionsprozeß dagegen in der verfügbaren Zeit scheitern, wird das entsprechende fehlende Attribut einfach geraten, auch hier wieder für Form und Farbe unabhängig voneinander. Durch falsches Raten können illusorische Konjunktionsurteile entstehen. „Echte" illusorische Konjunktionen als Folge der Verwechslung bereits identifizierter Farben oder Formen verschiedener Objekte gibt es in den Ratemodellen also nicht.

Ashby et al. (1996) formalisierten (a) ein einfaches Ratemodell, welches gleiche Ratewahrscheinlichkeiten für alle Antwortoptionen annimmt, (b) ein sophistiziertes Ratemodell, welches Ratewahrscheinlichkeiten auf null setzt, wenn die

entsprechende Antwortoption aufgrund der Merkmalsidentifikation anderer Objekte ausgeschlossen werden kann, (c) Treisman und Schmidts Merkmalsintegrationstheorie und (d) ihre eigene Lokalisationsunsicherheitstheorie als vier verschiedene multinomiale Modelle. Die Modelle (a) und (b) umfassen neben Merkmalsidentifikationswahrscheinlichkeiten für Form und Farbe nur a priori fixierte (also nicht freie) Ratewahrscheinlichkeiten. Bei den Modellen (c) und (d) kommt noch jeweils ein Parameter für die korrekte Merkmalskombination hinzu, und zwar α im Merkmalsintegrationsmodell und σ im Lokalisationsunsicherheitsmodell. Während die Modelle (a) bis (c) einfache binäre MVB-Modelle sind, ist das Modell (d) aufgrund des Parameters σ (bzw. aufgrund der involvierten Normalverteilungsannahme für die subjektive Lokalisation) *kein* MVB-Modell. Da es jedoch zur Klasse der parametrisierten multinomialen Modelle zählt, kann man es beispielsweise mit den in Abschnitt 4.7 besprochenen Programmen EXCEL (vgl. Dodson et al., 1998) oder PRAXIS (Gegenfurtner, 1992) relativ leicht analysieren.

In dem von Ashby et al. (1996) verwendeten experimentellen Paradigma wurden Vpn kurzfristig jeweils mit zwei verschiedenfarbigen Buchstaben konfrontiert, von denen genau einer zur zuvor definierten Zielmenge zählte (entweder ein X oder ein T) und einer zur Alternativmenge gehörte (entweder ein C oder ein S). Die Vpn mußten jeweils angeben, ob der präsentierte Buchstabe der Zielmenge ein X oder ein T war und welche Farbe (rot, gelb oder blau) er hatte. Pro Versuchsbedingung können dann sechs Beobachtungskategorien unterschieden werden:

1) korrekte Antworten (Buchstaben- und Farbinformation korrekt);
2) illusorische Konjunktionen (korrekter Buchstabe kombiniert mit der Farbe des interferierenden Buchstabens oder korrekte Farbe kombiniert mit der Bezeichnung des interferierenden Buchstabens);
3) andere Farbfehler (korrekter Buchstabe, nichtvorkommende Farbe);
4) andere Buchstabenfehler (nichtvorkommender Buchstabe, korrekte Farbe);
5) kombinierte Buchstaben-Farbfehler (nichtvorkommender Buchstabe, nichtvorkommende Farbe);
6) Buchstabenfehler mit illusorischer Farbkonjunktion (nichtvorkommender Buchstabe, Farbe des interferierenden Buchstabens).

Insgesamt ist dies eine sehr reichhaltige empirische Basis zur Separierung verschiedener hypothetischer Wahrnehmungsprozesse bei der Objekterkennung und es bereitet keinerlei Probleme, die o.g. Modellklassen anhand dieser Datenbasis gegeneinander zu kontrastieren.

Die vier Modelle wurden an die Daten einzelner Vpn nach der Maximum-Likelihood-Methoden angepaßt (Experiment 1). Da das sophistizierte Ratemodell ein Submodell sowohl des Merkmalsintegrationsmodells als auch des Lokalisationsunsicherheitsmodells ist — aus ersterem geht es durch die Restriktion $\alpha = 1$,

aus letzterem durch die Restriktion $\sigma = 0$ hervor —, wurde Akaikes AIC-Kriterium für einen parameterzahlbereinigten Modellvergleich herangezogen[90] (vgl. Abschnitt 4.3.6). Die Ergebnisse favorisieren die beiden Modelle mit „echten" illusorischen Konjunktionen im Vergleich zu den reinen Ratemodellen. Das Lokalisationsunsicherheitsmodell (durchschnittliches AIC = 4.118) schnitt etwas besser als das Merkmalsintegrationsmodell ab (durchschnittliches AIC = 4.146), was jedoch in Anbetracht der großen interindividuellen Unterschiede zwischen den AIC-Werten nicht als zwingendes Argument für das Lokalisationsunsicherheitsmodell gewertet werden kann. Auch der Anstieg der korrekten Mermalskombinationen mit der Distanz der interferierenden Objekte fiel nicht so konsistent und so deutlich aus wie vom Lokalisationsunsicherheitsmodell vorhergesagt (vgl. Ashby et al., 1996, Abbildung 6, S. 178). Ähnliche Ergebnismuster ergeben sich bei Verallgemeinerung der Modelle auf ein Paradigma, in dem die Vpn zusätzlich die Farbe des zweiten Buchstabens berichten müssen (Experiment 2), und bei Anwendung der Modelle auf Datensätze anderer Autoren (a.a.0, S. 183ff).

Offenbar liegen hier ähnlich wie im letzten Abschnitt empirisch schwer diskriminierbare multinomiale Modelle vor. Systematische Validierungsstudien zu den Merkmalsidentifikationsparametern fehlen bislang, so daß es für ein endgültiges Urteil über das beste Modell noch zu früh ist. Besonders wichtig und potentiell ergiebig scheinen Validierungsstudien zu sein, welche die Erkennbarkeit der interferierenden Farb- und Formmerkmale manipulieren. Bezüglich der zugehörigen Merkmalsidentifikationsparameter N_C (Farbe des interferierenden Buchstabens) und N_L (Bezeichnung des interferierenden Buchstabens) machen die Modelle (c) und (d) nämlich recht unterschiedliche Vorhersagen (vgl. Ashby et al., 1996, Tabellen 5 & 9).

Einmal mehr wird deutlich, daß theorievergleichende Anwendungen von multinomialen Modellen ohne systematische Validierungsstudien zu den zentralen Modellparametern kaum überzeugende Antworten auf die Frage nach der empirischen Adäquatheit verschiedener Theorien geben können.

In einer weiteren Untersuchung haben Prinzmetal, Henderson und Ivry (1995) die Darbietungszeit und die Aufmerksamkeitszuwendung orthogonal zueinander variiert und bei parafovealer Darbietung der Items keinen Effekt auf den Parameter α (Wahrscheinlichkeit einer korrekten Merkmalskombination im Merkmalsintegrationsmodell) gefunden. Derartige Effekte sind nach der Theorie von Treisman und Schmidt (1982) unbedingt zu erwarten, da die Performanz in der Merkmalsintegrationsphase der Objekterkennung wesentlich von einer längeren Darbietungszeit und von verfügbaren Aufmerksamkeitsressourcen profitieren

[90] In der Formel für AIC ist bei Ashby et al. (1996, S. 176, Gleichung (5)) der Faktor 2 für die negative Log-Likelihood vergessen worden (vgl. dazu Gleichung (4.3.6-1) in der vorliegenden Arbeit).

sollte. Der Negativbefund in dieser Studie schafft insofern mehr Klarheit als die Befunde von Ashby et al. (1996) als er die Treisman-Schmidt-Theorie schwer belastet.

7.6 Sozialpsychologische Anwendungen

Große Teile der Forschung zur sozialen Kognition basieren auf Untersuchungsparadigmen, die denen der Gedächtnis-, Denk- und Wahrnehmungspsychologie sehr ähnlich sind oder sogar exakt gleichen. Das Dekomponierungsproblem stellt sich hierbei typischerweise in der gleiche Schärfe wie in der allgemeinpsychologischen Forschung (vgl. Abschnitt 2.6). Deshalb kann es kaum überraschen, daß die MVB-Modellierung zunehmend auch in die Social-Cognition-Forschung Einzug hält.

7.6.1 Soziale Kategorisierung

Die bereits in Abschnitt 4.5 geschilderte Arbeit von Klauer und Wegener (1998) zum „Who-Said-What?"-Paradigma dürfte eine der ersten Anwendung eines MVB-Modells in der Sozialpsychologie überhaupt sein. Im Kern handelt es sich um Zwei-Hochschwellen-Quellengedächtnismodell für multiple Quellen (typischerweise Personen), die jeweils einer von zwei verschiedenen sozialen Gruppen zugeordnet werden können (Männer versus Frauen, ältere versus jüngere Menschen usw.). Die Quellengedächtnisparameter werden in einen Parameter für die soziale Gruppe der Personen und weitere Parameter für die Einzelpersonen aufgeteilt. Implizite soziale Kategorisierungsprozesse bei der Enkodierung der Originalinformation können über den Gruppenparameter erfaßt werden. Insgesamt ist das Modell dem unabhängig entwickelten Modell spezifischen und partiellen Quellengedächtnisses von Dodson et al. (1998) sehr ähnlich, welches in Abschnitt 7.1.2 vorgestellt wurde. Allerdings basiert letzteres auf einer Ein- statt einer Zwei-Hochschwellen-Annahme. Außerdem sind im Klauer-Wegener-Modell Rateparameter für die soziale Kategorie vorgesehen, die im Dodson-Modell fehlen. Rateparameter für die soziale Kategorie können im Kontext der Überprüfung von Theorien sozialer Kategorisierung eine wichtige Rolle spielen.

Das Klauer-Wegener-Modell wurde — wie bereits geschildert — hinsichtlich aller zentralen Parameter erfolgreich validiert (Klauer & Wegener, 1998). Weiterhin wurde das Modell zur Prüfung der Hypothese eingesetzt, daß soziale Kategorisierungsprozesse Aufmerksamkeitsressourcen beanspruchen und deshalb unter erschwerten Verarbeitungsbedingungen (z.B. Zeitdruck bei der Reizverarbeitung) nur eingeschränkt wirksam sein sollten. Im Gegensatz zur traditionellen varianzanalytischen Auswertung der Daten bestätigte der Gruppengedächtnisparameter im Klauer-Wegener-Modell diese Hypothese nicht, sondern stützte eher Annahme, daß soziale Kategorisierungsprozesse auch ohne intensive Aufmerksam-

keitsprozesse automatisch ablaufen. Dieses Ergebnis harmoniert sehr gut mit Befunden von Dodson et al. (1998), die ebenfalls keinen Effekt von Aufmerksamkeitsmanipulationen auf den Gruppengedächtnisparameter ihres Modells fanden (vgl. Abschnitt 7.1.2).

Klauer, Wegener und Ehrenberg (1999) haben das Modell von Klauer und Wegener (1998) eingesetzt, um den Einfluß der Subgruppengröße (Männer vs. Frauen in Experiment 1 und alte versus junge Personen in Experiment 2) auf Gedächtnisparameter und Ratetendenzen zu untersuchen. In beiden Experimenten beobachteten sie, daß sowohl das Itemgedächtnis als auch das Quellengedächtnis für die Person, von der eine zu beurteilende Aussage stammt, um so besser wurden, je kleiner die Subgruppe der Person war. Gleichzeitig verschlechterte sich das Gedächtnis für die Gruppenzugehörigkeit der Person mit kleiner werdender Subgruppengröße. Dieser Effekt wurde allerdings in beiden Experimenten nicht signifikant. Die Rateparameter für die soziale Kategorie wuchsen — wie zu erwarten — mit der Gruppengröße an.

In Experiment 2 wurden darüber hinaus stereotyp-konsistente und stereotyp-inkonsistente Aussagen für jede Altersgruppe gegeneinander kontrastiert. Diese Manipulation hatte keinen Effekt auf die Gedächtnisparameter — ein für verschiedene Schema-Theorien des Gedächtnisses kritischer Befund (vgl. Abschnitt 7.1.6). Allerdings waren stereotyp-konsistente und -inkonsistente Informationen ähnlich wie bei Bayen et al. (2000) in allen Quellen gleich häufig vertreten, so daß möglicherweise Rollenstereotype für Personen der verschiedenen Gruppen nicht maximal stark aktiviert wurden. Erwartungsgemäß fiel der Effekt der Stereotyp-Manipulation auf die Rateparameter aus: Konservative Aussagen wurden eher älteren Personen, progressive Aussagen eher jüngeren Personen zugeschrieben, und zwar vor allem bei gleich großen Altersgruppen. Dieses Ergebnis ist plausibel, da Alternativen zum stereotyp-basierten Raten bei gleich großen Gruppen kaum zur Verfügung stehen.

Insgesamt lassen sich die Ergebnisse am besten durch eine Distinktivitätshypothese erklären, wonach Subgruppenmitgliedern und ihren Aussagen um so mehr Aufmerksamkeit gewidmet wird, je kleiner die Subgruppen sind. Davon sollte — wie von den Autoren beobachtet — sowohl das Itemgedächtnis als auch das Quellengedächtnis profitieren. Lediglich der tendenziell eher entgegengesetzte Trend zum Gedächtnis für die soziale Gruppe des Sprechers harmoniert nicht so gut mit der Distinktivitätshypothese.

Wegener (1999) benutzte das Klauer-Wegener-Modell zur Überprüfung der Theorie von Oakes (1987), wonach die Salienz einer sozialen Kategorie multiplikativ von ihrer Zugänglichkeit (accessibility) einerseits und ihrer Passung (fit) andererseits abhängt. Letzteres meint das Ausmaß, in dem Personen einer sozialen Kategorie untereinander ähnlich und von Personen anderer sozialer Kategorien ver-

schieden sind. In fünf Experimenten wurde die Zugänglichkeit einer Kategorie über unterschiedliche Etikettierungen der Gruppenzugehörigkeit von Personen manipuliert, die bestimmte Einstellungen äußerten (z.B. „Kindererziehung ist auf keinen Fall allein Sache der Frau" oder „Kindererziehung ist vor allem das Privileg der Frau"). Durch eine Voruntersuchung war sichergestellt worden, daß die verschiedenen Gruppen aus Vp-Sicht unterschiedlich zugänglich waren. Beispiele für soziale Kategorien, die aus der Perspektive von Studierenden der Universität Bonn hohe Zugänglichkeit aufweisen, sind „Homosexuelle" versus „Heterosexuelle". Soziale Kategorien mit geringer Zugänglichkeit sind demgegenüber — wiederum aus Bonner Perspektive betrachtet — „Personen aus Münster" versus „Personen aus Aachen". Die Passung wurde durch unterschiedliche Grade der Kovariation zwischen Gruppenzugehörigkeit und Einstellungen manipuliert: Wenn beispielsweise Mitglieder der Gruppe A durchweg progressive und Mitglieder der Gruppe B durchweg konservative Einstellungen äußerten, war die Passung perfekt. Gab es dagegen keinen Zusammenhang zwischen der Gruppenzugehörigkeit und der Wahrscheinlichkeit einer progressiven versus konservativen Einstellungsäußerung, so war die Passung minimal.

In den ersten drei Experimenten zeigte sich, daß das Quellengedächtnis sowohl für die sich äußernde Person als auch für deren soziale Gruppe mit zunehmender Passung anstieg. Außerdem konnte ein inhaltlich plausibler Quellenratebias festgestellt werden: Die Neigung, im Zustand der Gedächtnisunsicherheit eine Aussage einer bestimmten sozialen Gruppe zuzuschreiben, wächst mit dem Prozentsatz der Gruppenmitglieder, die inhaltlich mit dieser Aussage konform gehen. Eine progressive Aussage wird also um so eher einem Mitglied von Gruppe A zugeschrieben, je mehr Mitglieder der Gruppe A sich zuvor progressiv geäußert haben. Auch die Zugänglichkeit beeinflußte das Kategoriegedächtnis und die Ratebereitschaft positiv. Insgesamt haben somit die beiden zentralen Konzepte der Theorie von Oakes (1987) durch die Experimente von Wegener deutliche empirische Stützung erfahren: Sowohl die Zugänglichkeit als auch die Passung beeinflussen einerseits die Erinnerung an die Gruppenzugehörigkeit einer Personen und andererseits das Rateverhalten bei Quellenunsicherheit.

Keine Bestätigung fand Wegener dagegen für die Annahme von Oakes (1987), daß Zugänglichkeit und Passung multiplikativ zusammenwirken. Die darin implizit enthaltene Prämisse, daß ohne eine minimale wahrgenommene Passung soziale Kategorisierung unmöglich ist, konnte er in seinen Experimenten 4 und 5 zurückweisen. Hierbei wurden Personen aus leicht zugänglichen sozialen Kategorien präsentiert — Männer und Frauen —, denen Stimuli ohne jede Passung — nämlich sinnlose Silben — zugeordnet waren. Trotzdem war der Schätzer für das Kategoriegedächtnis (*d*-Parameter) in beiden Experimenten signifikant größer als null. Als Fazit kann festgehalten werden, daß die Zugänglichkeit-mal-Pas-

sung-Theorie von Oakes (1987) offensichtlich relevante Einflußgrößen auf die Salienz sozialer Kategorien benennt, ihr Zusammenspiel im Hinblick auf Gedächtnis- und Urteilsprozesse aber noch einer weiteren Klärung bedarf. Ein einfaches multiplikatives Modell reicht nicht aus.

7.6.2 Illusorische Korrelationen

Eine weitere sozialpsychologische Anwendung von MVB-Modellen stammt von Klauer und Meiser (in press). Sie nimmt auf eine inzwischen klassische Arbeit von Hamilton und Gifford (1976) Bezug. Hamilton und Gifford (1976) haben bei ihren Vpn „illusorische Korrelationen" zwischen sozialer Gruppenzugehörigkeit und positiven versus negativen Eigenschaften dadurch induziert, daß über eine nicht näher beschriebene Gruppe A (die Majorität) viele Informationen vermittelt wurden, über die andere Gruppe B (die Minorität) dagegen wenig Informationen. Hamilton und Gifford zeigten, daß Gruppe A anschließend in verschiedener Hinsicht vorteilhafter beurteilt wurde als Gruppe B, obwohl der Anteil positiver Eigenschaften in beiden Gruppen gleichermaßen bei etwa .70 lag.

Klauer und Meiser untersuchten gedächtnisbasierte Erklärungen des Illusory-Correlation-Phänomens mit Hilfe des 2HTSM-Modells von Bayen at al. (1996). Im Unterschied zum Standardparadigma von Hamilton und Gifford mußten ihre Vpn nach Darbietung von Originalinformationen zu Gruppe A und Gruppe B, die hinsichtlich ihrer Valenz nicht mit der Gruppe korreliert waren, alte und neue Items, die positives oder negatives Verhalten zum Ausdruck brachten, in einem Quellengedächtnistest beurteilen. Es wurde ein starker Quellenratebias (a) zugunsten von Gruppe A festgestellt, d.h. positive Items wurden im Zweifel eher Gruppe A, negative Items eher Gruppe B attribuiert. Auf der anderen Seite unterschieden sich die Gruppen sowohl in bezug auf das Itemgedächtnis als auch in bezug auf das Quellengedächtnis nicht. Die Informationen zu beiden Gruppen wurden also gleich gut erinnert, was u.a. der Distinktivitätshypothese von Hamilton und Gifford (1976) direkt widerspricht. Nach dieser Hypothese müßten die seltenen negativen Verhaltensweisen in Gruppe B besonders tief verarbeitet und deshalb langfristig besser behalten werden.

Besser harmonieren die Befunde mit der Erklärung, daß die subjektive Basisrate für positives Verhalten im Verlauf der Akquisitionsphase durch einen bayesianischen Urteilsprozeß modifiziert wird, der bei neutralen Einschätzungen beider Gruppen beginnt und dann bei der Majoritätsgruppe aufgrund der größeren Stichprobe schneller gegen den hohen Wert .70 konvergiert als bei der kleineren Stichprobe. Die subjektive Basisrate positiven Verhaltens wäre somit am Ende der Akquisitionsphase bei Gruppe A größer als bei Gruppe B, was den Ratebias im a-Parameter verständlich erscheinen läßt (Klauer & Meiser, in press).

Die Befunde von Klauer und Meiser harmonieren im übrigen gut mit denen von Bayen et al., auf die in Abschnitt 7.1.6 bereits eingegangen wurde. Bayen et al. beobachteten, daß Quellenratetendenzen durch Schemaeffekte massiv beeinflußt werden, Item- und Quellengedächtnisparameter dagegen nicht. Klauer und Meiser haben unter Verwendung desselben Meßmodells das gleiche Ergebnismuster in bezug auf Stereotyp-Effekte gefunden. Man kann zur Zeit nur darüber spekulieren, ob dieses Ergebnismuster für alle Fälle typisch ist, in denen keine ausgeprägte schema- oder stereotypbasierte Enkodierung, jedoch ein ausgeprägter schemabasierter Abruf stattfindet.

Meiser und Hewstone (1999) haben den Effekt gekreuzter sozialer Kategorien im Paradigma illusorischer Korrelationen untersucht, wobei sie das bereits in Abschnitt 7.1.2 geschilderte Quellengedächtnismodell für mehrdimensionale Quelleninformationen verwendet haben. In der Versuchsbedingung mit gekreuzten Kategorien wurden positive und negative Verhaltensweisen von Personen aus Gruppe A oder B präsentiert, die entweder aus Stadt X oder Stadt Y kamen. Der Anteil positiver Verhaltensweisen war in allen vier Kombinationskategorien (A-X, A-Y, B-X, B-Y) jeweils .66, d.h. positive versus negative Verhaltensweisen waren weder mit der Gruppe (A versus B) noch mit der Stadt (X versus Y) korreliert. Mehr Informationen wurden über Mitglieder der Gruppe A und Bewohner der Stadt X gegeben, am wenigsten Informationen folglich über Mitglieder der Gruppe B aus Stadt Y. Wie schon Klauer und Meiser fanden auch Meiser und Hewstone keine Hinweise auf Unterschiede im Item- oder Quellengedächtnis zwischen den Majoritäts- und den Minoritätsgruppen. Bei Meiser und Hewstone war das Quellengedächtnis sogar unabhängig von der sozialen Gruppe und unabhängig vom Iteminhalt (positiv versus negativ) nahezu null. Dennoch resultierten starke illusorische Korrelationseffekte bezüglich der Gruppenzugehörigkeit, und zwar sowohl in den anschließend erhobenen Eigenschaftsratings der Gruppen A und B als auch in direkten Häufigkeitsschätzungen. Dies widerspricht klar der Distinktivitätshypothese von Hamilton und Gifford (1976).

Ursächlich für den illusorischen Korrelationseffekt scheint ein Ratebias zugunsten der Majoritätsgruppe im Zustand der Gedächtnisunsicherheit zu sein, der ja auch schon von Klauer und Meiser berichtet wurde. Allerdings wurde der Ratebias von Meiser und Hewstone nur bezüglich der Gruppen A und B, nicht aber für die Städte X und Y klar repliziert. Positive Verhaltensweisen wurden im Zustand der Rekognitionsunsicherheit vorwiegend der Majoritätsgruppe A zugeschrieben, wobei diese Tendenz erwartungsgemäß bei gekreuzten Kategorien etwas schwächer ausgeprägt war als in einer Versuchsbedingung, in der die Stadtzugehörigkeit nicht erwähnt wurde (ungekreuzte Kategorien). Stereotypisierungseffekte im Ratebias kommen also um so stärker zur Geltung, je homogener die gegeneinander kontrastierten Gruppen sind. Auf der anderen Seite werden Stereotypisierungsef-

fekte in den Ratetendenzen durch gekreuzte soziale Kategorien auch nicht komplett verhindert, sondern nur reduziert. Wahrgenommene Intra-Gruppen-Ähnlichkeit, die durch gekreuzte Kategorien verhindert wird, ist also keine conditio sine qua non der Stereotypbildung bzw. des damit verbundenen illusorischen Korrelationseffekts, sondern nur ein begünstigender Einflußfaktor.

Überraschenderweise konnten ähnlich starke Antworttendenzeffekte für die zweite soziale Dimension — Stadt X versus Stadt Y — nicht eindeutig nachgewiesen werden, unabhängig davon, ob diese Dimension nun isoliert oder gekreuzt mit Gruppe A versus B Verwendung fand. Dies ist ein Hinweis darauf, daß sich nicht jede soziale Variable gleichermaßen zur Erzeugung von Stereotypisierungseffekten eignet. Möglicherweise muß hier auch zwischen verschiedenen Vpn-Gruppen unterschieden werden, da ähnliche Variablen in vergleichbaren Untersuchungen mit anderen Vpn-Gruppen durchaus signifikante illusorische Korrelationseffekte produziert haben (vgl. die Übersicht bei Meiser & Hewstone, 1999).

7.7 Entwicklungspsychologische Anwendungen

Anwendungen von MVB-Modellen in der Entwicklungspsychologie bieten sich für alle Bereiche der kognitiven Entwicklung an. Prinzipiell gilt dies für alle zuvor behandelten MVB-Modelle. Ergänzt man die berichteten experimentellen Designs um die (quer- oder längsschnittliche variierte) Altersvariable, so werden aus den berichteten Untersuchungen entwicklungspsychologische Untersuchungen. Praktisch umgesetzt worden ist diese Idee vorwiegend im Bereich der Gedächtnispsychologie. Insbesondere Speicher-Abruf-Modelle sind häufig bei Altersgruppenvergleichen eingesetzt worden. Das Enkodierung-Abruf-Modell von Batchelder und Riefer (1980; 1986) hat vor allem im Bereich der gerontologischen Forschung Aufmerksamkeit gefunden. Im Regelfall wurde ein Altersdefizit hinsichtlich Cluster-Enkodierung und Cluster-Abruf beobachtet; das Abrufdefizit erwies sich jedoch als gravierender als das Enkodierungsdefizit (Erdfelder & Bayen, 1991; Riefer & Batchelder, 1991a). Genau das umgekehrte Bild ergab sich in bezug auf die Bedeutung von Speicherung und Abruf im Gedächtnis von Kindern: Obwohl tendenziell sowohl Speicherung als auch Abrufparameter Verbesserungen mit dem Alter zeigten, erwiesen sich die Speicherungsparameter im allgemeinen als sensitiver in bezug auf den Entwicklungsprozeß (vgl. Abschnitt 7.1.4). Möglich ist, daß die unterschiedlichen Ergebnismuster mit den unterschiedlichen Modellen und Paradigmen zusammenhängen, die bei Kindern und älteren Erwachsenen Verwendung fanden; vergleichende Untersuchungen, die dieselben Modelle und Paradigmen verwenden, stehen noch aus.

Neben gedächtnispsychologisch motivierten Modellen bieten sich natürlich auch Modelle an, welche die Intelligenzentwicklung bei Kindern modellieren.

Hierzu gehören beispielsweise Mischverteilungsmodelle zur Frage der kontinuierlichen versus diskreten Entwicklung (vgl. Abschnitt 6.4). Auch die zahlreichen Aufgabentypen, welche von Piaget und Mitarbeitern zur Untersuchung der Intelligenzentwicklung im Kindesalter vorgeschlagen wurden, haben vielfach Anlaß zu einer modellbasierten Betrachtung gegeben.

Dies soll hier nur an einem Beispiel illustriert werden, nämlich der Arbeit von Hodkin (1987) zu Klasseninklusionsproblemen. Betrachtet man einfach den Anteil der Kinder, die ein Klasseninklusionsproblem korrekt gelöst haben, so riskiert man — wie Hodkin (1987) gezeigt hat — eine erhebliche Überschätzung der Wahrscheinlichkeit, daß ein Kind der betreffenden Altersgruppe die Klassinklusionslogik beherrscht. Einige Kinder haben möglicherweise einfach geraten und dadurch per Zufall korrekte Antworten produziert. Um unverfälschte Schätzungen für die Wahrscheinlichkeit der Beherrschung von Klasseninklusionsproblemen in einer Altersgruppe zu erhalten, verwendete sie ein Paradigma mit zwei 2AFC-Aufgaben zur Klasseninklusion, die nacheinander zu beantworten sind. Eine Frage ist eine Standardinklusionsaufgabe. Sieht ein Kind beispielsweise drei blaue und zwei rote Dreiecke, so könnte man fragen: „Sind dort mehr blaue Dinge oder mehr Dreiecke?" Eine zweite Frage vergleicht die kleinere Subkategorie mit der Oberkategorie: „Sind dort mehr rote Dinge oder mehr Dreiecke?" Hodkin nahm an, daß Kinder, die die Inklusionslogik beherrschen, beide Aufgaben korrekt lösen. Kinder, die die Strategie des Subklassenvergleichs anwenden, geben in der ersten Frage die falsche und in der zweiten Frage die richtige Antwort. Sie mißverstehen die Standardinklusionsaufgabe also als Frage noch der größten Subklasse. Daneben gibt es eine Gruppe von Kindern, die keine der beiden Strategien anwenden und stattdessen einfach raten, und zwar bei den beiden Aufgaben unabhängig voneinander.

Das Modell von Hodkin (1987) läßt sich als einfaches Zwei-Ebenen-MVB-Modell zur 2×2-Kontigenztafel richtiger versus falscher Antworten in den beiden Aufgaben darstellen. Auf der Grundlage dieses Modells lassen sich die Wahrscheinlichkeit der drei Strategien (korrekte Klasseninklusioninterpretation, Subklassenvergleichsinterpretation, Raten) schätzen (Thomas, 1995).

Modifikationen und Verallgemeinerungen dieses Modells wurden von Rabinowitz, Howe und Lawrence (1989) sowie Thomas (1995) vorgeschlagen. In dem Modell von Rabinowitz et al. (1989) sind neben den drei genannten Strategiewahrscheinlichkeiten zusätzlich zwei Gedächtnisparameter für die korrekte Speicherung der Aufgabeninformation enthalten. Identifizierbar und testbar wird das Modell durch Erweiterung auf Fragen zum direkten Subklassenvergleich („Sind dort mehr blaue Dreiecke oder mehr rote Dreiecke?") und durch Unterscheidung zweier verschiedener Fehlerkategorien pro Aufgabe. Howe und Rabinowitz (1996) haben das Modell von Rabinowitz et al. (1989) in einer Querschnittsstudie mit 10- bis 67-jährigen Personen eingesetzt. Entwicklungsprozesse schlagen sich nach ih-

ren Ergebnissen in erster Linie in einem gegenseitigen Austausch zwischen den Strategien „Subklassenvergleichsinterpretation" versus „Klasseninklusionsinterpretation" nieder: Bis zur Altersgruppe der 23-Jährigen nahmen die Schätzungen der Wahrscheinlichkeit *u* für eine korrekte Klasseninklusionsstrategie zu, und zwar auf Kosten der Wahrscheinlichkeit *s* für eine fälschliche Subklassenvergleichsstrategie (Howe & Rabinowitz, 1996, S. 27, Abbildung 3). In der Gruppe der 67-Jährigen war ein geringfügiger Abfall der *u*-Parameterschätzungen auf das Niveau der 13- bis 15-Jährigen zu beobachten. Die Gedächtnisparameterschätzungen lagen unabhängig von der untersuchten Altersgruppe generell auf sehr hohen Niveau. Hierbei ist allerdings zu berücksichtigen, daß Retentionsintervalle länger als 1 Sekunde in der Untersuchung unberücksichtigt blieben.

7.8 Psychometrische Anwendungen

Da standardisierte Fragebögen, Interviews und Tests für jede Frage oder Aufgabe typischerweise diskrete Antwortalternativen vorsehen (Ja-Nein-Antworten, Multiple-Choice-Antworten oder Ratings), bieten sich multinomiale Modelle für die Datenanalyse geradezu an. Einige psychometrische Modelle, die in den letzten 30 Jahren zur Analyse von Fragebögen oder standardisierten Interviews vorgeschlagen wurden, fallen darüber hinaus in die Kategorie der binären MVB-Modelle, wobei nur teilweise auf diese Modellklasse explizit Bezug genommen wird. Wie schon im letzten Kapitel dargestellt, kann die Einbettung der Modelle in den MVB-Kontext wegen der Verfügbarkeit benutzerfreundlicher Computerprogramme und vor allem wegen dann offenkundig werdender Verallgemeinerungs- und Modifikationsmöglichkeiten viele Vorteile mit sich bringen. Obwohl nicht primär kognitionspsychologisch relevant, sollen deshalb einige psychometrisch motivierte MVB-Modelle in diesem Kapitel nicht unerwähnt bleiben.

7.8.1 Randomisierte Antworttechnik

Warner (1965) hat eine randomisierte Antworttechnik („randomized response technique") vorgeschlagen, um aus Selbstbeurteilungsdaten möglichst unverfälschte Schätzungen von Parametern zu erhalten, die sozial negativ bewertete Einstellungen oder Verhaltensmuster betreffen. Die tatsächlichen Probandenantworten auf solche Fragen sind offensichtlich problematisch. Fragt man Probanden etwa danach, ob sie schon einmal illegale Drogen konsumiert haben, so muß man damit rechnen, daß einige Probanden, auf die dies zutrifft, allein deshalb wahrheitswidrig mit „nein" antworten, weil negative Konsequenzen im Falle einer positiven Antwort befürchtet werden.

Um dieses Problem zu umgehen, kann man beispielsweise so vorgehen, daß die Probanden instruiert werden, vor ihrer Antwort eine Münze zu werfen. Wenn

„Kopf" oben liegt, sollen die Probanden wahrheitsgemäß antworten. Wenn dagegen „Zahl" oben liegt, sollen sie in jedem Fall mit „ja" antworten. Dies stellt sicher, daß bei einer „Ja"-Antwort ein Schluß auf die tatsächlichen Einstellungen oder Verhaltensmuster der betreffenden Person nicht möglich ist, so daß objektiv kein Grund mehr für eine Verfälschung der Antwort gegeben ist. Dennoch kann man den Anteil π der Gesamtpopulation, auf den das sozial sanktionierte Attribut zutrifft, ganz einfach schätzen, denn für die Wahrscheinlichkeit einer „Ja"-Antwort, $p(\text{„ja"})$, gilt das binäre MVB-Modell $p(\text{„ja"}) = \pi + (1-\pi) \cdot 0.5$, sofern eine faire Münze geworfen wurde und alle Probanden sich an die Instruktion halten (Dawes & Moore, 1980).

Die Voraussetzung, daß sich alle Probanden an die Instruktion halten, kann man mit einer einfallsreichen Modifikation der Prozedur durch Clark und Desharnais (1998) prüfen. Sie schlagen vor, die Probanden per Zufall zwei experimentellen Bedingungen zuzuweisen, die sich lediglich hinsichtlich des verwendeten „Zufallsgenerators" unterscheiden: Eine Gruppe benutzt einen Zufallsgenerator, der mit (bekannter) Wahrscheinlichkeit q_1 eine zwingende „Ja"-Antwort verlangt, während die andere Gruppe mit (bekannter) Wahrscheinlichkeit q_2 ($q_2 \neq q_1$) in jedem Fall eine „Ja"-Antwort geben muß. Nimmt man an, daß ein unbekannter Anteil γ der Population der Instruktion *nicht* folgt und immer mit „nein" antwortet (z.B. um die Möglichkeit des Zutreffens der sozial sanktionierten Eigenschaft prinzipiell auszuschließen), so ergibt sich ein verbundenes binäres MVB-Modell mit zwei Modellgleichungen[91] $p_i(\text{„ja"}) = (1-\gamma) \cdot (\pi + (1-\pi) \cdot q_i)$ für die Gruppen $i = 1$ und $i = 2$. Das Modell ist saturiert und identifizierbar; die Hypothese $\gamma = 0$ ist über die in Abschnitt 4.3.2 geschilderten Parametertests prüfbar.

Auch in das Modell von Clark und Desharnais (1998) gehen natürlich wieder problematische Annahmen ein, beispielsweise die Annahme, daß alle nicht instruktionsgemäß antwortenden Probanden unabhängig vom Ergebnis des Zufallsgenerators immer „Nein"-Antworten geben. Dies ließe sich in einem Mehrgruppenmodell mit mehr als zwei Zufallsgeneratorwahrscheinlichkeiten q_i prüfen. Schon in einem Drei-Gruppen-Design mit unterschiedlichen vorgegebenen Wahrscheinlichkeiten q_1, q_2 und q_3 wäre das Modell von Clark und Desharnais (1998) nicht mehr saturiert und könnte folglich einem Modellgeltungstest unterzogen werden.

Verallgemeinerungen des Ansatzes auf mehrkategoriale Antwortoptionen (multiple choice tests) sind leicht möglich. Soll ein Proband etwa von $J > 2$ Freizeitbeschäftigungen diejenige angeben, der er bzw. sie am liebsten nachgeht, so

[91] Die Modellgleichungen werden hier in einer anderen Parametrisierung als bei Clark und Desharnais (1998) wiedergegeben. π ist bei ihnen der *Anteil der Gesamtpopulation*, der ehrlich antwortet *und* auf den das kritische Merkmal zutrifft. Hier ist π der *Anteil unter den instruktionsgemäß antwortenden Probanden*, auf den das kritische Merkmal zutrifft.

muß ebenfalls mit Verzerrungen zugunsten sozial erwünschter Antworten gerechnet werden. Man kann dem durch eine Modifikation der o.g. randomisierten Antworttechnik begegnen, bei der Probanden Zufallskarten ziehen, auf denen in einem bestimmten Prozentsatz der Fälle die Frage steht, die durch Angabe der Antwortnummer wahrheitsgemäß zu beantworten ist, während in den restlichen Fällen auf der Karte einfach eine Zufallszahl zwischen 1 und *J* steht, die laut vorgelesen werden soll (Levy, 1980). Die Gültigkeit der Annahme, daß die Probanden die Instruktion befolgen, kann auch hier durch Erweiterung auf ein Design mit mehreren Gruppen überprüft werden, die unterschiedliche Verteilungen der Zufallszahlen auf die *J* Antwortkategorien aufweisen.

7.8.2 Konsensusanalyse

Die Konsensusanalyse („cultural consensus analysis") ist ein Analyseverfahren für Datenmatrizen, welche aus der Beantwortung von *m* Fragen durch *n* Informanden (oder Probanden) resultieren. Die Antworten werden als kategorial vorausgesetzt; im einfachsten Fall sind sie dichotom (trifft zu / trifft nicht zu). Die korrekten Antworten auf die Fragen gelten als unbekannt: Es geht gerade darum, die korrekten Antworten auf die Fragen aus dem Konsensus der Informanden zu schätzen. Da die Informanden in ihrer Expertise bezüglich der gestellten Fragen variieren, ist es allerdings nicht einfach damit getan, die über die Informanden hinweg häufigste Antwort auf eine Frage als die mutmaßlich richtige auszuzeichnen. Antworten von Informanden mit hoher Expertise sollten stärker gewichtet werden als Antworten von Nichtexperten. Ähnlich wie die korrekte Antwort pro Frage wird auch der Grad der Expertise pro Informand als a priori unbekannt betrachtet. Beides sind zu schätzende Parameter im Konsensus-Modell.

Die Konsensusanalyse ist für viele Wissenschaften ausgesprochen nützlich. In der Ethnologie stellt sich beispielsweise die Frage nach den wirklichen Verhältnissen in einer weitgehend unbekannten Kultur (definiert über Antworten auf *m* Fragen). Diese Frage muß anhand der Antworten von *n* Forschern mehr oder minder ausgeprägter Expertise beantwortet werden. In der forensischen Psychologie stellt sich analog die Frage nach dem wirklichen Tathergang einer Straftat, die auf der Basis der Antworten von *n* Zeugen unbekannter Expertise auf *m* Fragen zu beantworten ist. Die Kette der Beispiele ließ sich leicht fortsetzen.

Zur Erläuterung der Modellannahmen soll das Ausgangsmodell von Batchelder und Romney (1986; 1988; Romney, Batchelder & Weller, 1987; Romney, Weller & Batchelder, 1986) betrachtet werden (sog. „General Condorcet Model" für dichotome Items). Diesem Modell liegt die Annahme zugrunde, daß die Fragen homogen bzw. „gleich schwer" sind. Dies bedeutet, daß jedes Item *k* mit einer für alle Items gleichen Apriori-Wahrscheinlichkeit π zutreffend ($X_k = 1$) und mit der

komplementären Wahrscheinlichkeit $1-\pi$ nicht zutreffend ist ($X_k = 0$). Wenn ein Item in Wahrheit zutreffend ist, antwortet der Informand i mit der informandenspezifischen „Trefferrate" P_{1i} korrekterweise „zutreffend" ($Y_{ik} = 1$) und mit der komplementären Wahrscheinlichkeit „nicht zutreffend" ($Y_{ik} = 0$). Ist ein Item k in Wahrheit nicht zutreffend, so antwortet der Proband mit einer informandenspezifischen „Rate falschen Alarms" P_{0i} fälschlicherweise „zutreffend" ($Y_{ik} = 1$) und sonst „nicht zutreffend" ($Y_{ik} = 0$). Alle Informanden $i = 1, ..., n$ antworten lokal (für gegebenen Wahrheitswert des Items) stochastisch unabhängig voneinander.

Das „General Condorcet Model" ist unschwer als Latent-Class-Modell für eine transponierte Datenmatrix erkennbar, bei der Items und Vpn miteinander vertauscht sind (Batchelder & Romney, 1988, S. 75, Beobachtung 5; vgl. auch Abschnitt 6.3). Wenn man explizite Maße für die Expertise bzw. „Kompetenz" (Parameter D_i) und den Zustimmungsbias eines Informanden (Parameter g_i) ableiten will, empfiehlt es sich, P_{1i} und P_{0i} im Sinne eines Ein-Hochschwellen-Modells (Batchelder et al., 1997, S. 46) oder — besser noch — im Sinne eines Zwei-Hochschwellen-Modells (Batchelder & Romney, 1988, S. 73) zu reparametrisieren:

$$\begin{aligned} p(Y_{ki} = 1 \,|\, X_k = 1) &= P_{1i} = D_i + (1 - D_i) \cdot g_i \\ p(Y_{ki} = 1 \,|\, X_k = 0) &= P_{0i} = (1 - D_i) \cdot g_i \end{aligned} \qquad (7.8.2\text{-}1)$$

Signalentdeckungstheoretische Modelle anstelle der Gleichungen (7.8.2-1) sind ebenfalls denkbar (Batchelder et al., 1997, S. 44f), führen allerdings aus der Klasse der MVB-Modelle heraus. Schätzungen des wahren Antwortschlüssel für die m Fragen kann man aus den „recruitment probabilities" zu gegebenem Antwortmuster $\boldsymbol{Y}_k = \boldsymbol{y}_k$ aller n Informanden auf eine Frage k ableiten (vgl. Gleichung 2.3-1):

$$\begin{aligned} & p(X_k = 1 \,|\, Y_{1k} = y_{1k}, \ ..., Y_{nk} = y_{nk}) = \\ & = \frac{\pi \cdot \prod_{i=1}^{n} p(Y_{ik} = y_{ik} \,|\, X_k = 1)}{\pi \cdot \prod_{i=1}^{n} p(Y_{ik} = y_{ik} \,|\, X_k = 1) + (1 - \pi) \cdot \prod_{i=1}^{n} p(Y_{ik} = y_{ik} \,|\, X_k = 0)} \end{aligned} \qquad (7.8.2\text{-}2)$$

Ist $p(X_k = 1 \mid Y_{1k} = y_{1k}, ..., Y_{nk} = y_{nk}) > .5$, so ist die bestmögliche Vorhersage für das k-te Item „zutreffend" und sonst „unzutreffend". Schätzungen der Modellparameter liefern die gewünschten Informationen über Expertise und Antworttendenzen verschiedener Informanden sowie den wahren Sachverhalt, der ihren Antworten mutmaßlich zugrunde liegt.

Das Grundmodell hat verschiedene Verallgemeinerungen auf nichtdichotome Antwortformate and Rangdaten erfahren (Batchelder & Romney, 1989; Romney et al., 1987). Die neuesten Anwendungen liegen im Bereich der Analyse sozialer Netzwerkdaten. Hierbei muß jedes von N Gruppenmitgliedern für alle

möglichen Paare von Gruppenmitgliedern angeben, ob die betreffenden Personen befreundet sind („trifft zu") oder nicht („trifft nicht zu"). Mittels eines modifizierten Konsensus-Modells wird es möglich, aus den sozialen Netzwerkdigraphen verschiedener Gruppenmitglieder (a) einen „wahren" Konsensus-Digraphen der gesamten Gruppe sowie (b) Kompetenzen und Antworttendenzen in bezug auf die Diagnose von Freundschaftsbeziehungen in der Gruppe zu schätzen (Batchelder et al., 1997; Kumbasar, in press). Eine Reanalyse sozialer Netzwerkdaten von Kumbasar, Romney und Batchelder (1994) zeigte erwartungsgemäß, daß Kompetenz und Antworttendenz (d.h. die Bejahungsbereitschaft für Freundschaft für Freundschaftsbeziehungen) mit zunehmender Ego-Distanz abnehmen: Für Beziehungen, in die man selbst involviert ist, sind die Parameterschätzungen für D und g am größten, gefolgt von Beziehungen zwischen eigenen Freunden („ego´s innet") und Beziehungen zwischen Nichtfreunden („ego´s outnet").

Einschränkend muß gesagt werden, daß der Befund von Batchelder et al. (1997) einer der wenigen Validierungsbefunde zum Konsensus-Modell ist. Insbesondere ist noch unklar, ob bzw. unter welchen Bedingungen die Schätzungen des wahren Antwortschlüssels aus den Informandenantworten als valide gelten können. Dieser Sachverhalt steht praktischen Anwendungen des Modells — beispielsweise im Bereich der Psychologie der Zeugenaussage — bislang noch im Wege.

7.8.3 Übereinstimmungsanalyse

Dem Konsensus-Modell konzeptuell verwandt ist das Übereinstimmungsmodell von Klauer und Batchelder (1996). Hierbei steht allerdings die Schätzung des Übereinstimmungsgrades zweier Beurteiler — die sogenannte „Beurteilerobjektivität" — im Vordergrund, nicht die Schätzung der Kompetenz einzelner Beurteiler. Im einfachsten Fall zweier Beurteiler werden N Beurteilungsobjekte von jedem Beurteiler in genau eine von K Kategorien eingruppiert, so daß eine $K{\cdot}K$-Datenmatrix resultiert, deren Hauptdiagonale Übereinstimmungen der Beurteiler reflektiert. Beispielsweise können zwei unabhängig diagnostizierende Psychiater N Patienten jeweils einer von drei nosologischen Kategorien zuordnen (z.B. Persönlichkeitsstörung, Neurose und Psychose, vgl. Klauer & Batchelder, 1996, S. 201).

Das Übereinstimmungsmodell nimmt an, daß ein Beurteiler $r = 1, 2$ mit Wahrscheinlichkeit λ die wahre Kategorie erkennt. Der Parameter λ ist hierbei unabhängig vom Beurteiler und auch unabhängig von der wahren Kategorie eines beurteilten Objekts. Wird mit der komplementären Wahrscheinlichkeit $(1–\lambda)$ die wahre Kategorie nicht erkannt, so rät der betreffende Beurteiler eine Kategorie mit beurteilerspezifischen Ratewahrscheinlichkeiten $\varepsilon_r(k)$, $k = 1, ..., K$. Nimmt man zusätzlich an, daß die Urteile verschiedener Beurteiler stochastisch unabhängig sind,

so lassen sich die Wahrscheinlichkeiten für die $K \cdot K$ Zellen der Übereinstimmungsmatrix als Funktion der Parameter λ („Beurteilerreliabilität") , $\varepsilon_1(k)$, $k = 1, ..., K$, $\varepsilon_2(k)$, $k = 1, ..., K$ (Antworttendenzen) sowie den wahren Anteilen von Objekten in Kategorie k, $\tau(k)$, $k = 1, ..., K$, ausdrücken. Es handelt sich — wie im Falle des Konsensus-Modells aus dem letzten Abschnitt — um ein restringiertes Latent-Class-Modell, bei dem die wahren Kategorien der beurteilten Objekte als latente Klassen fungieren und die Beurteiler die Rolle der Variablen einnehmen. Verallgemeinerung auf Übereinstimmungsdatensätze von drei oder mehr Beurteilern ergeben sich relativ problemlos.

Klauer und Batchelder (1996) zeigten, daß λ^2 eine enge Beziehung zum Kappa-Koeffizienten aufweist, den Cohen als Maß der Beurteilerreliabilität vorgeschlagen hat. Man kann also sagen, daß das Übereinstimmungsmodell in gewisser Weise die Modellannahmen explizit macht, die Cohens Kappa zugrunde liegen. Damit ergibt sich die Möglichkeit, die Modellannahmen zu testen. Nur wenn sie erfüllt sind, hat der Kappa-Koeffizient eine klare Interpretation. Klauer und Batchelder (1996) diskutierten auch mehrere Submodelle, die durch Parameterrestriktionen aus dem Übereinstimmungsmodell hervorgehen. Darüber hinaus wurden verschiedene Modifikationen und Erweiterungen des Übereinstimmungsmodells erörtert.

Ein etwas problematischer Fall ist das dichotome Kategoriensystem ($K = 2$). Für diesen Fall ist das allgemeine Übereinstimmungsmodell nämlich nicht identifizierbar. Lösungsmöglichkeiten hierfür zeigte Klauer (1996b) auf.

Klauer und Migulla (1995) haben das Übereinstimmungsmodell zur Untersuchung der Bedingungen des sogenannten kontrafaktischen Denkens eingesetzt. Hierunter versteht man die mentale Konstruktion irrealer Vorstellungswelten. Zwei Beurteilter sortierten schriftliche Protokolle von Versuchspersonen in eine von vier Kategorien ein, darunter auch die Kategorie „kontrafaktisches Denken". Die Protokolle bezogen sich auf positiv- versus negativ-valente Ereignisse, die entweder erwartet oder unerwartet waren. Die $\tau(k)$-Parameterschätzungen zeigten erwartungsgemäß, daß kontrafaktisches Denken primär durch Ereignisse mit negativer Valenz ausgelöst wird. Im Widerspruch zu einigen Theorien stand jedoch der Befund, daß unerwartete Ereignisse nicht mehr kontrafaktisches Denken auslösen als erwartete.

7.8.4 Finite-State-Theorie für Multiple-Choice-Tests

Hutchinson (1982) hat vermutlich als erster den Begriff „Finite-State-Theorie" benutzt, um damit psychometrische Modelle abzugrenzen, die — ähnlich wie Hochschwellen-Modelle für Rekognitionstests — von der Annahme ausgehen, daß Probanden bezüglich eines Testitems entweder im Zustand des Wissens oder im

Zustand des Nichtwissens sein können. Garcia-Pérez (1987; 1989a; 1989b; 1990) hat gezeigt, daß ein auf dieser Annahme aufbauendes testtheoretisches Modell für *k*-Alternativen-Multiple-Choice-Tests sehr wohl in der Lage sein kann, (a) vollständiges Wissen, (b) partielles Wissen, (c) komplett fehlendes Wissen und (d) Rateverhalten bezüglich eines Items abzubilden. Der Trick, mit dem ihm dies gelingt, ist die Auflösung eines *k*-Alternativen-Multiple-Choice-Items in *k* Aussagen. Betrachten wir als Beispiel das folgende Drei-Alternativen-Multiple-Choice-Item:

> Welche Autoren haben erstmals den Begriff „multinomial processing-tree models" (MVB-Begriff) eingeführt?
>
> (a) Batchelder und Riefer (1986)
> (b) Riefer und Batchelder (1988)
> (c) Batchelder und Riefer (1990).

Hierzu lauten die drei Aussagen:

(a) Batchelder und Riefer (1986) haben den MVB-Begriff eingeführt;
(b) Riefer und Batchelder (1988) haben den MVB-Begriff eingeführt;
(c) Batchelder und Riefer (1990) haben den MVB-Begriff eingeführt.

Es wird angenommen, daß ein Proband die korrekte Antwort zu jeder Aussage mit Wahrscheinlichkeit λ kennt und daß Kenntnisse bezüglich verschiedener Aussagen stochastisch unabhängig sind — eine problematische Annahme, wie mir scheint (siehe dazu weiter unten). Da vorausgesetzt werden kann, daß genau eine der *k* Aussagen pro Item wahr ist, gibt es prinzipiell drei Wege zu einer korrekten Antwort:

(1) *Wissensprinzip:* Der Proband weiß eine Teilmenge der *k* Antworten und die zutreffende Aussage — im Beispiel Aussage (c) — ist darunter.
(2) *Ausschlußprinzip:* Der Proband weiß $k-1$ Antworten und sie lauten durchweg „falsch". Im Beispiel könnte der Proband z.B. wissen, daß die Aussagen (a) und (b) falsch sind. Dann folgt daraus, daß die letzte Antwort (c) wahr sein muß.
(3) *Rateprinzip:* Der Proband kennt weniger als $k-1$ Antworten und die korrekte Antwort ist nicht darunter. Durch Raten unter den noch nicht ausgeschlossenen Alternativen wird die korrekte Antwort zufällig „erwischt".

Das Modell von Garcia-Pérez (1987, 1989b) setzt nicht voraus, daß Probanden zwischen den verbleibenden Alternativen raten, wenn der Fall (3) eintritt. Vielmehr raten sie nur mit Wahrscheinlichkeit γ, und zwar dann jede der verbleibenden Antwortalternativen mit gleicher Wahrscheinlichkeit. Mit der komplementären Wahrscheinlichkeit $1-\gamma$ gibt der Proband gar keine Antwort auf das Item. Insgesamt werden somit drei Beobachtungskategorien pro Item unterschieden: korrekte Antwort, falsche Antwort und keine Antwort. Die verschiedenen Wege zu

den drei Antwortkategorien lassen sich problemlos über ein Verarbeitungsbaumdiagramm veranschaulichen (vgl. z.B. den Anhang bei Garcia-Pérez, 1987).

Garcia-Pérez (1987) zeigte, daß bei Anwendung seines Modell auf Ja-Nein-Tests über die traditionelle Ratekorrektur („Anteil korrekter Antworten minus Anteil falscher Antworten") ein ML-Schätzer für den λ-Parameter abgeleitet werden kann. Das gilt jedoch nicht mehr für *k*-Alternativen-Multiple-Choice-Tests ($k \geq 2$). Wie Simulationsstudien belegen, sind die aus seinem Modell abgeleiteten ML-Schätzer für λ alternativen Maßen der Testperformanz überlegen (Garcia-Pérez, 1989b; Garcia-Pérez & Frary, 1989). Daß das Modell wirklich gültig ist, versuchte Garcia-Pérez (1987) dadurch plausibel zu machen, daß er analoge Finite-State-Modelle für andere Antwortmodalitäten zu Multiple-Choice-Items entwickelte. Hierzu zählen beispielsweise das Ankreuzen aller bekanntermaßen falschen Alternativen („Distraktoren-Erkennung" bzw. Coombs response mode) und die Möglichkeit eines zweiten Antwortversuchs nach einer falschen Erstantwort („second guessing"). Wenn Probanden die gleichen Items unter verschiedenen Antwortmodalitäten bearbeiten, lassen sich aus den Modellgleichungen testbare Relationen zwischen den Kategoriewahrscheinlichkeiten für verschiedene Antwortmodalitäten deduzieren. Diese Beziehungen scheinen sich empirisch auch einzustellen (Garcia-Pérez, 1987).

Garcia-Pérez (1989a) hat das Modell für verschiedene Werte von *k* und verschiedene Antwortmodalitäten explizit formuliert. Ferner hat er abgeleitet, wie viele Testitems für eine Mißerfolg-Erfolg-Entscheidung ($H_0: \lambda \leq \lambda_0 - \varepsilon$ und $H_1: \lambda > \lambda_0 + \varepsilon$) nach vorgegebenem Erfolgskriterium λ_0 und nach vorgegebener Fehlertoleranz ε erforderlich sind, wenn bestimmte α– und β–Fehlerwahrscheinlichkeiten bei der statistischen Entscheidung zugrunde gelegt werden. Für derartige kriteriumsorientierte Entscheidungen erwies sich die Distraktorerkennungsaufgabe nach Coombs als besonders effizient, d.h. sie lieferte eine gewünschte Präzision mit der kleinstmöglichen Itemanzahl.

Garcia-Pérez (1990) verglich sein Finite-State-Modell mit einem Modell, das statt einer dichotomen latenten Fähigkeitsvariablen (Wissen versus Nichtwissen) eine stetig verteilte Fähigkeitsvariable unterstellt. Hinsichtlich der Modellanpassungsgüte erwiesen sich die Modelle als gleichwertig. Der Autor argumentierte jedoch, daß die Finite-State-Theorie flexibler auf verschiedene Antwortmodalitäten leichter adaptierbar ist. Außerdem sei die Finite-State-Theorie als Modell des Antwortprozesses psychologisch plausibler. Einen Vergleich der itemcharakteristischen Kurven (item characteristic curves, ICC) der Finite-State-Theorie mit den ICCs anderer psychometrischer Modelle präsentierten Garcia-Pérez und Frary (1991a). Garcia-Pérez (1994) hat schließlich verschiedene Parameterschätzverfahren und Anpassungsstatistiken aus der Potenz-Divergenz-Familie anhand seines Finite-State-Modells miteinander verglichen. Hierauf wurde bereits eingegangen

(vgl. die Abschnitt 4.2.4, 4.2.5 und 4.3.5). Die ML-Methode schnitt in seiner Simulationsstudie insgesamt recht gut ab.

Garcia-Pérez und Frary (1991b) sowie Garcia-Pérez (1993) haben Erweiterungen des geschilderten Modellansatzes auf Multiple-Choice-Tests mit der Antwortoption „keine der Antwortalternativen ist richtig" („none of the above", NOTA-Alternative) präsentiert. Die Modellgleichungen ergeben sich in recht kanonischer Weise aus den bereits geschilderten Annahmen. Garcia-Pérez (1993) zeigte, daß Multiple-Choice-Tests mit NOTA-Alternativen genauere Schätzungen von λ ermöglichen, d.h. den Standardschätzfehler relativ zu konventionellen Multiple-Choice-Tests ohne NOTA-Alternativen reduzieren.

Insgesamt ist das Forschungsprogramm von Garcia-Pérez ein beeindruckender Beleg für die Brauchbarkeit von MVB-Modellen auch im Bereich psychometrischer Modellbildung. Bestimmte Auswertungsverfahren und Ratekorrekturen lassen sich auf dem Hintergrund dieses Modells rechtfertigen, andere müssen als suboptimal verworfen werden. Besonders nützlich ist die Flexibilität des Modellansatzes, die eine Anwendung auf ganz unterschiedliche Antwortformate für Multiple-Choice-Tests ermöglicht. So wird formal entscheidbar, welches Antwortformat für bestimmte Ziele optimal ist.

Eine Schwäche des Modellansatzes scheint mir jedoch die Unabhängigkeitsannahme bezüglich der Kenntnis der Antworten auf verschiedene Antwortalternativen zu sein. Die korrekte Antwortalternative muß m.E. hiervon ausgenommen werden, denn die Kenntnis der korrekten Antwort impliziert logisch die Kenntnis der Falschheit aller anderen Antwortalternativen. Dieser Sachverhalt ist im Modell von Garcia-Pérez nicht berücksichtigt worden. Modifikationen des Modellansatzes, die den geschilderten Einwand berücksichtigen, sind aber relativ leicht konzipierbar. Konkrete Untersuchungen müssen dann darüber entscheiden, ob der Einwand auch empirisch begründet werden kann.

7.9 Zusammenfassung

In diesem Kapitel wurde ein Überblick über bisherige Anwendungen von MVB-Modellen in der Psychologie präsentiert. Anwendungen findet man in erster Linie in der Gedächtnispsychologie, aber auch in der Denkpsychologie, Wahrnehmungspsychologie, Sprachpsychologie und Sozialpsychologie. Bei der Darstellung einschlägiger Arbeiten wurde weniger Wert auf eine detaillierte Darstellung der verwendeten Modelle gelegt. Vielmehr ging es in erster Linie um Analyse des Stellenwerts der MVB-Modellierung im substanzwissenschaftlichen Kontext. Sieben Anwendungsformen von MVB-Modellen wurden identifiziert:

(1) *Meßmodellanwendungen*: Im Vordergrund steht hierbei die Ableitung von möglichst unkontaminierten Maßen für bestimmte psychologische Prozesse.

(2) *Validierungsanwendungen*: In erster Linie geht es darum, die zentralen Parameter eines Modells zu validieren, d.h. aufzuzeigen, daß diese Parameter selektiv durch bestimmte Maßnahmen beeinflußt werden können.

(3) *Exploratorische Anwendungen*: Eine Fragestellung soll beantwortet werden, zu der keine eine Apriori-Hypothese vorliegt. Als Beispiel mag die Frage dienen: „Ist das episodische Gedächtnisdefizit von Epileptikern mit linkstemporalem Fokus primär speicher- oder primär abrufbedingt?"

(4) *Hypothesentestende Anwendungen*: Ein als psychologisch valide geltendes (oder vorausgesetztes) Modell wird zur Prüfung einer oder mehrerer Hypothesen eingesetzt, die das Eintreten bzw. Ausbleiben eines Parameterunterschieds zwischen zwei oder mehr Bedingungen behaupten. Als Konsequenz können entweder einzelne dieser Hypothesen Bestätigung erfahren oder alternativ auch sämtliche existierenden Hypothesen widerlegt werden. Im letztgenannten Fall schließt sich idealerweise die Konstruktion neuer substanzwissenschaftlicher Hypothesen an die MVB-Modellierung nahtlos an (vgl. etwa Abschnitt 7.3.1.3).

(5) *Theorievergleichende Anwendungen*: Mehrere MVB-Modelle, die als Formalisierungen unterschiedlicher Theorien aufgefaßt werden, werden anhand der gleichen Daten auf der Basis von Anpassungsstatistiken wie G^2, AIC oder BIC verglichen, um die „beste Theorie" herauszufinden.

(6) *Ableitungskritische Anwendungen*: Geprüft wird, ob etablierte empirische Zusammenhangsbehauptungen („Gesetze") sich wirklich aus bestimmten theoretischen Annahmen nach geeigneter Formalisierung deduzieren lassen, oder ob sie lediglich empirische Generalisierungen mit geringerem Stellenwert im Theorieevaluationsprozeß sind (vgl. insbesondere Abschnitt 7.1.8).

(7) *Fundierende Anwendung:* MVB-Modelle dienen dem Zweck, bestimmte Indizes oder Koeffizienten, die in der psychologischen Datenanalyse bereits gebräuchlich sind, gewissermaßen post hoc theoretisch zu fundieren. Auf diese Weise wird erkennbar, welche theoretischen Annahmen man zugrunde legen muß, damit die betreffenden Koeffizienten sinnvoll interpretierbar sind. Sofern das MVB-Modell testbar ist, kann auch die Adäquatheit der Annahmen geprüft werden. Ein Beispiel hierfür ist das Übereinstimmungsmodell von Klauer und Batchelder (1996), welches Cohens Kappa-Maß der Beurteilerübereinstimmung fundiert (vgl. Abschnitt 7. 8. 3).

Sechs Hauptprobleme bisheriger Anwendung von MVB-Modellen sind offensichtlich. Erstens stehen einer sehr großen Zahl von MVB-Modellen nur ganz wenige Validierungsuntersuchungen zu diesen Modellen gegenüber. In vielen Fällen wird die Modellbewährung über eine insignifikante Modellanpassungsstatistik oder eine negative AIC- oder BIC-Statistik offenbar als hinreichende Validierung

eines Modells angesehen, was jedoch allein wegen des Problems datenäquivalenter Modelle nicht überzeugen kann (vgl. Abschnitt 4.4).

Zweitens ergeben sich unübersehbare Probleme speziell bei theorievergleichenden Anwendungen von MVB-Modellen. Zunächst einmal wird generell wenig Mühe darauf verwendet, die Eignung eines bestimmten Paradigmas zur vergleichenden Evaluation verschiedener Theorien zu begründen. Damit die zumeist verbal formulierten Theorien als MVB-Modelle zu einem bestimmten Paradigma darstellbar sind, müssen im Regelfall technische Zusatzannahmen ergänzt werden. Wenn nun ein Modell viele problematische Zusatzannahmen beinhaltet, ein anderes dagegen nur wenige, dann ist das bessere Abschneiden des letztgenannten Modells allein aufgrund der problematischen Zusatzannahmen des erstgenannten Modells zu erwarten. In einem anderen Untersuchungsparadigma hätte sich möglicherweise genau das umgekehrte Ergebnis ergeben. Mit anderen Worten: Das Ergebnis der vergleichenden Theorieevaluation ist möglicherweise ein Artefakt des gewählten Untersuchungsparadigmas.

Darüber hinaus ergeben sich besondere Probleme beim modellbasierten Theorienvergleich, wenn eines der Modelle parameterreich, ein anderes dagegen parameterarm ist. Datenstrukturen, die mit dem letztgenannten Modell konsistent sind, können somit fast immer auch durch das erstgenannte erklärt werden, während die Umkehrung nicht gilt. Hier hilft nur ein umfassendes Validierungsprogramm zu allen Parametern des allgemeineren Modells weiter (vgl. Abschnitt 7.5.2).

Drittens wurden theoretische Annahmen bei der MVB-Modellierung nicht immer adäquat umgesetzt. Beispielsweise bleiben theoretisch geforderte Ordnungsrestriktionen zu den Modellparametern bei den statistischen Analysen gelegentlich unberücksichtigt, was zu einer Überschätzung der Modellanpassungsgüte führen kann (vgl. z.B. die Untersuchung von Kail et al., 1984, in Abschnitt 7.1.4).

Viertens wurden gelegentlich Parameterschätzwerte in Korrelations- oder Varianzanalysen weiteranalysiert, die aus nachgewiesenermaßen inadäquaten Modellen abgeleitet wurden (vgl. die Studie von Henkel et al., 1998; siehe Abschnitt 7.1.2). Für derartige Parameterschätzwerte existieren keine sinnvollen psychologischen Interpretationen. Sie können somit zu Fehlinterpretationen der Befunde führen.

Fünftens sind die empirischen Befunde, die mit verschiedenen Modellen erzielt wurden, nicht immer konsistent. So sprechen beispielsweise einige Speicher-Abruf-Modelle für eine größere Bedeutung des Speicherungsprozesses bei der Gedächtnisentwicklung des Kindes, andere eher für eine größere Bedeutung des Abrufprozesses. Dies muß natürlich kein „Modellartefakt" sein, sondern kann auch auf den unterschiedlichen Lernmaterialien und Testparadigmen in den Untersuchungen beruhen. Klärende Untersuchungen hierzu fehlen jedoch bislang.

Sechstens — und das dürfte eines der Hauptprobleme bisheriger MVB-Modellierungen sein — ist die mangelnde Vernetzung der verschiedenen Modelle untereinander unverkennbar. Vor allem die Flut gedächtnispsychologisch motivierter Modelle macht deutlich, daß man sich bislang wenig um Kohärenz und Synergie bemüht hat. Im Gegenteil: MVB-Modelle zum menschlichen Gedächtnis machen einen extrem atomistischen, paradigmenspezifischen und desintegrierten Eindruck. Obwohl alle Modelle auf nur wenige Gedächtnissysteme und Gedächtnisprozesse Bezug nehmen, sind bislang so gut wie keine Versuche unternommen worden, die Parameter verschiedener Modelle formal oder empirisch zueinander in Beziehung zu setzen. Diesbezüglich gibt es sicher erheblichen Nachholbedarf. Ansätze zu einer integrativen Betrachtung verschiedener MVB-Modelle und der dazugehörigen empirischen Paradigmen werden im folgenden Kapitel präsentiert.

8. Modelle zu unterschiedlichen Quellengedächtnistests

Im vorliegenden Kapitel sollen Möglichkeiten paradigmenübergreifender multinomialer Modellierung am Beispiel von Quellengedächtnistests untersucht werden. Ganz offensichtlich sind unterschiedliche Paradigmen denkbar, die auf eine Erfassung der Item- und Quellengedächtnisfähigkeit von Vpn abzielen. So kann man etwa im Anschluß an ein positives Rekognitionsurteil („Das Item X ist alt") einen Zwei-Alternativen-Forced-Choice-Quellentest darbieten („Stammt das Items X aus Quelle A oder aus Quelle B?"), stattdessen lediglich einen Ja-Nein-Quellentest zur Anwendung bringen („Stammt das Item X aus Quelle A: Ja / Nein?") oder auf einen gewöhnlichen Drei-Alternativen-Forced-Choice Test zurückgreifen, was in der Praxis wohl am häufigsten geschieht („Bitte kreuzen Sie eine der folgenden drei Alternativen an: Item X ist neu / Item X ist alt und stammt aus Quelle A / Item X ist alt und stammt aus Quelle B"). Alternativ kann man auch daran denken, unter zwei verschiedenen Instruktionen — beispielsweise der Inklusions- und der Exklusionsinstruktion aus Jacobys (1991) Prozeßdissoziationsprozedur — Rekognitionsurteile zu Items zweier verschiedener Quellen zu erheben. Beide Urteile zusammengenommen geben Aufschluß über die Quellenzuordnung eines Items (vgl. Buchner, Erdfelder et al., 1997; vgl. auch Abschnitt 7.1.3).

Obwohl die verschiedenen Aufgaben unterschiedliche Urteilsstrategien bei den Vpn anregen mögen, sollten sich im Regelfall doch die *gleichen* Item- und Quellengedächtniseinflüsse in den Antworten niederschlagen, unabhängig von der verwendeten Aufgabe. Es sollte also möglich sein, eine Familie von MVB-Modellen zu den verschiedenen Paradigmen zu konstruieren, die über die Gedächtnisparameter miteinander verknüpft sind. Sofern vergleichbare Vpn und Items unter identischen Versuchsbedingungen mit den verschiedenen Quellengedächtnistests untersucht werden, müßten sich folglich über die aufgabenspezifischen Modelle hinweg auch identische Item- und Quellengedächtnisparameter ergeben. Sind stattdessen signifikante Unterschiede korrespondierender Gedächtnisparameterschätzungen zu beobachten, so ist die psychologische Validität der gesamten Modellfamilie potentiell in Frage gestellt. Nur wenn sich überzeugende psychologische Argumente dafür anführen lassen, warum bestimmte Quellengedächtnisaufgaben das Item- oder Quellengedächtnis fördern bzw. behindern, ließen sich entsprechende Unterschiede korrespondierender Gedächtnisparameter tolerieren. Für die oben exemplarisch aufgeführten Aufgaben sind solche Argumente jedoch kaum denkbar.

In den folgenden Abschnitten wird die Paradigmenunabhängigkeit von Item- und Quellengedächtnismaßen in dreierlei Weise untersucht. Zunächst wird gefragt, ob gewöhnliche Drei-Alternativen-Quellendiskriminationstests und Ja-Nein-Quellentests Daten liefern, die mit der Hypothese aufgabenunabhängiger

Item- und Quellengedächtnisparameter verträglich sind. Zweitens wird untersucht, ob die Item- und Quellengedächtnisparameter für eine bestimmte Quelle A unabhängig davon sind, von welcher anderen Quelle B die Quelle A im Quellengedächtnistest zu diskriminieren ist. Diese beiden Fragen beschäftigen uns in Abschnitt 8.1. Abschnitt 8.2 untersucht demgegenüber die Frage, ob Aussagen zur Validität verschiedener Quellendiskriminationsmeßmodelle über Paradigmen hinweg Gültigkeit haben. Eine zusammenfassende Diskussion schließt das Kapitel ab (Abschnitt 8.3).

8.1 Drei-Alternativen-Tests und Ja-Nein-Tests

Marsh und Hicks (1998) haben jüngst den Einfluß des Aufgabenformats auf korrekte versus falsche Antworten bei verschiedenen Quellengedächtnistests untersucht. Sie fanden erhebliche Einflüsse sowohl des Aufgabenformats als auch des Quellenkontexts auf den Prozentsatz korrekter Quellenattributionen. Da sie auf eine modellbasierte Analyse der Daten verzichtet haben, lassen ihre Ergebnisse jedoch offen, ob die beobachteten Effekte Gedächtnisprozesse betreffen oder lediglich auf unterschiedliche Antwort- bzw. Ratetendenzen zurückzuführen sind.

Die Ergebnisse der Experimente 1 und 2 von Marsh und Hicks (1998) sind so gut dokumentiert, daß eine modellbasierte Reanalyse der Daten zwecks Beantwortung der Frage zur Bedeutung von Antworttendenzen relativ leicht möglich ist (vgl. die Tabellen 1 und 2 bei Marsh & Hicks, 1998). In ihrem Experiment 1 mußten insgesamt 72 Vpn 40 Items lesen und 40 Items aus Anagrammen generieren, bevor nach einem kurzen Retentionsintervall ein Quellengedächtnistest mit insgesamt 120 Items folgte (die 80 alten Items plus 40 neue Items). Die Zuordnung eines Items zu den drei Bedingungen „lesen", „generieren" und „neues Item" geschah pro Vp zufällig. 33 der 72 Vpn erhielten einen gewöhnlichen Drei-Alternativen-Test (3AFC-Test) als Quellengedächtnistest („Wurde das Item gelesen, generiert oder ist es neu?") während 39 Vpn einfache Ja-Nein-Tests erhielten, und zwar für jeweils 20 Items jeder Bedingung in der Form „Wurde das Item gelesen?" (ja/nein), für die anderen 20 Items jeder Gruppe in der Form „Wurde das Item generiert?" (ja/nein). Experiment 2 war eine konzeptuelle Replikation von Experiment 1, bei der die Anagramm-Bedingung lediglich durch eine akustische Darbietungsbedingung (ohne Itemgenerierung) ersetzt wurde. Hier erhielten 30 Vpn den Drei-Alternativen-Test und 35 Vpn den Ja-Nein-Test.

Die relativen Antworthäufigkeiten für alle Versuchsbedingungen und beide Experimente sind hier in Tabelle 8.1 wiedergegeben. Wie deutlich zu erkennen ist, führen Ja-Nein-Frageformate generell zu größeren Zustimmungsraten als es aufgrund der 3AFC-Antworthäufigkeiten zu erwarten ist. In der Anagramm-Bedingung von Experiment 1 ist dieser Effekt z.B. so stark, daß das Quellenurteil

„gesehen" in nur 19% der Fälle beim 3AFC-Test, jedoch in 34% der Fälle im Ja-Nein-Test gegeben wird. Marsh und Hicks (1998) sahen zwei Möglichkeiten der Interpretation: Einerseits kann eine Ja-Nein-Frage den Gedächtnissuchprozeß bezüglich der erfragten Quelle intensivieren und so das Quellengedächtnis verbessern. Andererseits kann die Frage nach einer speziellen Quelle aber auch präsumtiv wirken und so die Ja-Sage-Bereitschaft im Sinne einer Antworttendenz erhöhen.

Tabelle 8.1:
Relative Antworthäufigkeiten in Drei-Alternativen-Forced-Choice-Tests (3AFC) und Ja-Nein-Tests zur Erfassung des Quellengedächtnisses (entnommen aus Marsh & Hicks, 1998, Tabelle 1 und Tabelle 2). Die Zeilen entsprechen den Itembedingungen, die Spalten den Vp-Urteilen.

Experiment 1						Experiment 2					
Test:	3-AFC			Ja-Nein		Test:	3-AFC			Ja-Nein	
Urteil:	*Seen*	*Gener.*	*New*	*Seen*	*Gener.*	Urteil:	*Seen*	*Heard*	*New*	*Seen*	*Heard*
Seen	.53	.12	.36	.63	.19	Seen	.60	.17	.23	.64	.32
Gener	.19	.71	.10	.34	.83	Heard	.20	.59	.21	.27	.70
New	.20	.07	.73	.33	.08	New	.13	.15	.73	.21	.18

Zugleich ist in Tabelle 8.1 zu erkennen, daß die relativen Antworthäufigkeiten für Items aus Quelle A (gesehene Items) zum Teil davon abhängen, ob die zu diskriminierende Quelle generierte Items (Experiment 1) oder gehörte Items (Experiment 2) umfaßt: Gesehene Items werden in Experiment 2 häufiger korrekt erkannt und seltener fälschlich als neu bezeichnet. Weist dies darauf hin, daß das Item- oder Quellengedächtnis für gesehene Items durch die Kontextquelle beeinflußt werden kann?

Zur Beantwortung der aufgeworfenen Fragen wurde ein verbundenes MVB-Modell konstruiert, das simultan auf einen 3AFC-Quellendiskriminationstest und zwei Ja-Nein-Quellendiskriminationstests zu Quelle A und Quelle B anzuwenden ist. Der Teil des Modells, der sich auf den 3AFC-Test bezieht, ist mit dem bereits geschilderten 2HTSM-Modell des Quellengedächtnisses identisch. Dieses Modell hat sich — wie bereits erwähnt — in Validierungsstudien besser als andere Modelle zu 3AFC-Tests bewährt (Bayen et al., 1996; Erdfelder et al., 1995; vgl. auch die Abschnitte 4.4, 4.5 und 7.1.2 der vorliegenden Arbeit). Der Grundgedanke dieses Modells ist, daß das Itemgedächtnis dem Quellengedächtnis vorgeordnet ist: Nur wenn ein altes Item aus Quelle i (i = 1, 2) rekogniziert wird (mit Wahrscheinlichkeit D_i), kann mit der bedingten Wahrscheinlichkeit d_i zusätzlich auch die Quelle erinnert werden, worauf ein korrektes Quellenurteil folgt. Wird ein Item zwar rekogniziert, jedoch mit Wahrscheinlichkeit $1 - d_i$ die Quelle nicht erinnert, so wird sie geraten, und zwar mit Wahrscheinlichkeit a die erste und mit Wahrscheinlichkeit $1-a$ die zweite Quelle. Wird ein Item mit Wahrscheinlichkeit $1- D_i$ nicht

rekogniziert und folglich auch die Quelle nicht erkannt, so wird sowohl der Alt-Neu-Status (mit Wahrscheinlichkeit b bzw. $1-b$) als auch die Quelle (mit Wahrscheinlichkeit g bzw. $1-g$) geraten. Neue Items werden mit Wahrscheinlichkeit D_3 als neu erkannt (worauf ein „Neu"-Urteil folgt) oder mit der komplementären Wahrscheinlichkeit $1-D_3$ nicht als neu erkannt. Im letztgenannten Falle folgen dann die gleichen Rateprozesse wie bei nichtrekognizierten alten Items.

Wenn eine paradigmenunabhängige Erfassung des Item- und Quellengedächtnisses möglich ist, sollte man mit einer Übertragung dieser Annahmen auf Ja-Nein-Tests Gedächtnisparameter erhalten, die unter vergleichbaren experimentellen Bedingungen denen des 2HTSM-Modells numerisch entsprechen. Das für die beiden Ja-Nein-Tests konzipierte Meßmodell ist daher dem 2HTSM-Modell weitgehend analog. Der einzige Unterschied besteht darin, daß die Ratewahrscheinlichkeiten der Ja-Nein-Tests mit denen des 3AFC-Tests nicht übereinstimmen müssen. Konkret: Wird ein altes Item rekogniziert und die Quelle korrekt erinnert, so wird modellgemäß auf die Frage nach dieser Quelle mit „ja" und auf die Frage nach der anderen Quelle mit „nein" geantwortet — wie es das 2HTSM-Modell nahelegt. Dagegen wird auf einen Ja-Nein-Test zur ersten Quelle (A) mit Wahrscheinlichkeit a_a „ja" geraten, wenn das Item zwar rekogniziert wurde, jedoch die Quelle nicht erinnert wird. Diese Wahrscheinlichkeit wird im allgemeinen nicht mit dem analogen Parameter a des 2HTSM-Modells übereinstimmen. Ist darüber hinaus nicht nur die Quelle unbekannt, sondern wird ein Item auch nicht rekogniziert, so erfolgen nicht zwei sukzessive Rateprozesse wie im 2HTSM-Modell (zum Alt-Neu-Status und ggf. zur Quelle), sondern lediglich ein Rateprozeß, bei dem mit Wahrscheinlichkeit g_a das Urteil „ja" und mit der komplementären Wahrscheinlichkeit das Urteil „nein" resultiert. Hierbei ist im allgemeinen $g_\mathrm{a} < a_\mathrm{a}$ zu erwarten, da die Apriori-Wahrscheinlichkeit, ein Item aus der ersten Quelle (A) vorliegen zu haben, bei Rekognitionsunsicherheit niedriger ist als bei Quellenunsicherheit im Zustand der Rekognitionssicherheit: Bei Rekognitionsunsicherheit kommen nämlich drei Ratemöglichkeiten in Frage (Quelle A, Quelle B oder neu), bei Quellenunsicherheit nur zwei (Quelle A oder Quelle B).

Ganz analoge Annahmen gelten für den Ja-Nein-Test zur zweiten Quelle (B). Hierfür seien die beiden „Ja"-Rateparameter im Zustand der Quellenunsicherheit und Zustand der Rekognitionsunsicherheit mit a_b bzw. g_b bezeichnet. Sie müssen mit denen des ersten Ja-Nein-Tests nicht zwangsläufig übereinstimmen, obwohl eine solche Übereinstimmung — wie noch erläutert wird — unter bestimmten Bedingungen plausibel ist. Übereinstimmung wird dagegen (a) zwischen den Itemgedächtnisparametern des 3AFC-Tests und denen der beiden Ja-Nein-Tests ($D_i = D_{\mathrm{a}i} = D_{\mathrm{b}i}$) und (b) zwischen den zugehörigen Quellengedächtnisparametern ($d_i = d_{\mathrm{a}i} = d_{\mathrm{b}i}$) zwingend gefordert. Hierin drückt sich gerade die Annahme paradigmenunabhängiger Messung des Item- und des Quellengedächtnisses aus.

Eine exakte Beschreibung des Gesamtmodells für alle drei Tests ist in Anhang 10 in Form einer EQN-Datei angegeben, die mit den Computerprogrammen MBT, GPT oder AppleTree analysiert werden kann. Ebenfalls dort wiedergegeben ist die Datei der Antworthäufigkeiten zu Experiment 1 und 2 von Marsh und Hicks (1998), die aus der obigen Tabelle 8.1 durch Multiplikation der relativen Häufigkeiten mit der entsprechenden Vpn-Anzahl und der Itemanzahl pro Vp geschätzt werden kann. Geringfügige Rundungsfehler in den relativen Häufigkeiten führen zu Fehlern in den Schätzungen, die jedoch bei den gegebenen Stichprobenumfängen vernachlässigbar sind und auf Parameterschätzungen sowie Modellanpassungsstatistiken keinen bemerkenswerten Einfluß haben.

Tabelle 8.2:
ML-Parameterschätzungen, 90%-Konfidenzintervalle (in Klammern) und Modellanpassungsstatistiken (G^2) zum gemeinsamen MVB-Modell für 3AFC- und Ja-Nein-Quellengedächtnistests, bezogen auf die Daten gemäß Tabelle 8.1. Die Ergebnisse basieren auf dem Submodell mit der Restriktion $a_a = a_b$.

Experiment 1			Experiment 2		
Itemged. „seen"	D_1	.50 (.40 - .60)	Itemged. „seen"	D_1	.58 (.51 - .64)
Itemged. „gen."	D_2	.86 (.83 - .89)	Itemged. „heard"	D_2	.62 (.56 - .69)
Itemged. „new"	D_3	.00 (-.46-.46)	Itemged. „new"	D_3	.37 (.27 - .48)
Quelleng. „seen"	d_1	.66 (.62 - .71)	Quelleng. „seen"	d_1	.78 (.65 - .91)
Quelleng. „gen."	d_2	.61 (.55 - .67)	Quellg. „heard"	d_2	.53 (.45 - .61)
Itembias (3AFC)	b	.27 (.14 - .40)	Itembias (3AFC)	b	.45 (.36- .53)
Quellb 1 (3AFC)	a	.49 (.42 - .56)	Quellb 1 (3AFC)	a	.42 (.30 - .54)
Quellb 2 (3AFC)	g	.74 (.71 - 78)	Quellb 2 (3AFC)	g	.47 (.42 - .51)
Bias 1 (seen?)	$a_{a\ (*)}$	.87 (.80 - .94)	Bias 1 (seen?)	$a_{a\ (*)}$	.48 (.35 - .61)
Bias 2 (seen?)	ga	.32 (.18 - .47)	Bias 2 (seen?)	ga	.33 (.27 - .39)
Bias 1 (gen.?)	$a_{b(*)}$	.87 (.80 - .94)	Bias 1 (heard?)	$a_{b(*)}$	.48 (.35 - .61)
Bias 2 (gen?)	g_b	.08 (.04 - .12)	Bias 2 (heard?)	g_b	.61 (.51 - .71)
Modellfit	$G^2_{(1)}$	1.08	Modellfit	$G^2_{(1)}$	0.89

Das Gesamtmodell wurde mit der ML-Methode für jedes der beiden Experimente getrennt an die Daten angepaßt. Mit den geschilderten Gleichheitsrestriktionen zu den Item- und Quellengedächtnisparametern umfaßt es fünf Gedächtnisparameter (D_1, D_2, D_3, d_1, d_2) und sieben Rateparameter, drei zum 3AFC-Test (b, a, g) und je zwei zu den beiden Ja-Nein-Tests (a_a, g_a bzw. a_b, g_b). Bei $M = 12$ funktional unabhängigen Kategoriewahrscheinlichkeiten ergibt dies ein vollparametrisiertes Modell, das zwar identifizierbar, aber nicht testbar ist. Testbar wird das Modell jedoch, wenn man die Rateparameter a_a und a_b der Ja-Nein-Tests gleichsetzt. Dies ist im gegebenen Fall eine plausible Restriktion: Wird ein Item rekogniziert, die Quelle jedoch nicht erinnert, so sollte keine Ratepräferenz zugunsten einer der

beiden Quellen bestehen, sofern die Vp weiß, daß beide Quellen mit gleicher Wahrscheinlichkeit im Gedächtnistest repräsentiert sind.

Die Ergebnisse der ML-Parameterschätzungen und die $G^2_{(1)}$-Anpassungsstatistiken zu diesem restringierten Modell sind in Tabelle 8.2 aufgeführt. Offensichtlich paßt das Modell für beide Experimente gut auf die Daten. Auch die Schätzwerte zum *a*-Rateparameter des 3AFC-Tests, der unrestringiert gelassen wurde, fügen sich gut in das oben skizzierte Bild ein: Für beide Experimente liegt die *a*-Parameterschätzung nahe bei .50, d.h. es gibt keine Ratepräferenz für ein Quellenurteil im Falle von Rekognitionsunsicherheit.

Eine scheinbar naheliegende zusätzliche Restriktion des Modells — die Gleichsetzung der 3AFC-Quellenratewahrscheinlichkeiten im Zustand der Quellenunsicherheit (*g*) und im Zustand der Rekognitionssicherheit (*a*) — läßt sich nicht generell aufrechterhalten. Relativ zum soeben geschilderten Modell ergibt sich ein χ^2-Differenzenwert von immerhin $\Delta G^2_{(1)} = 15.70$ für die Daten des ersten Experiments und von $\Delta G^2_{(1)} = 0.30$ für Experiment 2. Der erste Wert ist signifikant auf dem beim gegebenen Stichprobenumfang vernünftigerweise anzusetzenden Niveau $\alpha = .05$. Noch deutlicher ist die Unhaltbarkeit der Hypothese gleicher Ja-Nein-Ratewahrscheinlichkeiten bei Rekognitionsunsicherheit (d.h. $g_a = g_b$). Anders als im Falle der Gleichsetzungsrestriktion $a_a = a_b$ muß diese Hypothese nicht nur für Experiment 1 klar zurückgewiesen werden ($\Delta G^2_{(1)} = 163.95$), sondern auch für Experiment 2 ($\Delta G^2_{(1)} = 48.81$). Aus den Parameterschätzungen der Tabelle 8.2 ist ersichtlich, daß auf die Frage wurde „Wurde das Item generiert?" bei Rekognitionsunsicherheit auffällig selten „ja" geraten wurde (.08), auf die Frage „Wurde das Item gehört?" dagegen auffällig häufig (.61). Dieses Ergebnismuster ist für die signifikanten ΔG^2-Statistiken in erster Linie verantwortlich.

Muß die Validität des Gesamtmodells aufgrund dieser zunächst überraschenden Ergebnisse in Frage gestellt werden? Diese Frage ist zu verneinen. Strack und Bless (1994) haben besonders eindrucksvoll gezeigt, daß Personen bei Erinnerungsunsicherheit häufig metakognitive Strategien einsetzen, um plausible Urteile abzuleiten. Liegen zwei Klassen von Items vor, von denen eine Klasse sehr gut, die andere eher mäßig gut erinnerbar ist, so ist es vernünftig, bei Rekognitionsunsicherheit zu raten, daß das Item zur mäßig gut erinnerbaren Klasse zählt. Aufgrund ihres metakognitiven Wissens können die Vpn dann nämlich schlußfolgern, daß sie in der Lage sein müßten, das Item zu erinnern, wenn es zur sehr gut erinnerbaren Klasse zählen würde.

Auf diesem Hintergrund ist es durchaus verständlich, daß die Vpn in Experiment 1 bei Rekognitionsunsicherheit sehr selten raten, ein Item ursprünglich generiert zu haben: Da das Item nicht rekogniziert wurde, kann es eigentlich nicht zur Klasse der generierten Items zählen, denn diese sind sehr gut erinnerbar. Dies wird in Tabelle 8.2 durch den sehr hohen Schätzwert von .86 für das Itemgedächt-

nis deutlich. Nicht ganz so gut erklärbar ist, warum in Experiment 2 bei Rekognitionsunsicherheit auf die Frage „Wurde das Item ursprünglich gehört?" auffällig häufig „ja" geraten wird. Das Itemgedächtnis für gehörte Items unterscheidet sich nämlich von dem für gesehene Items nicht bedeutsam. Als Erklärungsmöglichkeit bleibt nur das schlechtere Quellengedächtnis für gehörte Items (vgl. Tabelle 8.2), das die Vpn möglicherweise als schlechteres Itemgedächtnis fehldeuten. Unter der Voraussetzung, daß das Itemgedächtnis für gehörte Items tendenziell schlecht ist, wäre es wieder vernünftig, auf die Frage „Wurde das Item ursprünglich gehört?" häufig „ja" zu raten. Diese Post-Hoc-Erklärung bedarf natürlich noch einer weitergehenden Prüfung. Insbesondere wäre zu klären, ob die metakognitiven Einschätzungen von Vpn mit den obigen Annahmen im Einklang steht.

Die Item- und Quellengedächtnisparameterschätzungen zu beiden Experimenten liefern Ergebnisse, die aufgrund bisheriger Forschungsresultate plausibel sind (vgl. die Abschnitte 7.1.2 und 7.1.6). Für generierte Items ist das Itemgedächtnis offensichtlich besser als für bloß gelesene Items, nicht aber das Quellengedächtnis. Ganz ähnliche Ergebnisse haben Erdfelder und Bredenkamp (1998) in bezug auf skriptbasierte Informationen gefunden. Offenbar führt das Generieren von Information zu einer elaborierten itemspezifischen Verarbeitung, wobei jedoch der Grund für die tiefere Verarbeitung — der Akt des Generierens — nicht als zusätzliche Markierung (tag) in der Gedächtnisspur repräsentiert ist (Erdfelder & Bredenkamp, 1998). Das Rekognitionsgedächtnis profitiert hiervon, nicht jedoch das Quellengedächtnis.

Anders als gelesene und generierte Items (Experiment 1) unterscheiden sich gelesene und gehörte Items (Experiment 2) hinsichtlich des Itemgedächtnisses kaum, wohl aber hinsichtlich des Quellengedächtnisses. Letzteres ist für gelesene Items — vermutlich aufgrund ausgeprägterer perzeptueller Gedächtnisspuren — besser als für bloß gehörte Items. Da gelesene Items in beiden Experimenten verwendet wurden und die Enkodierungsbedingungen vergleichbaren waren, müssen speziell für diese Itemklasse gleiche Item- (D_1) und Quellengedächtnisparameter (d_1) in Experiment 1 und 2 erwartet werden. Die Ergebnisse entsprechen diesen Erwartungen. Deskriptiv sind die entsprechende Schätzwerte in Experiment 2 zwar tendenziell etwas besser, aber die Unterschiede sind nicht signifikant, und zwar weder für das Itemgedächtnis ($\Delta G^2_{(1)} = 0.82$) noch für das Quellengedächtnis ($\Delta G^2_{(1)} = 1.00$). Auch die D_3-Parameterschätzungen für die Entdeckung neuer Items — wiewohl in beiden Experimenten deskriptiv deutlich unterschiedlich — unterscheiden sich zwischen den Experimenten nicht signifikant ($\Delta G^2_{(1)} = 1.35$). Es gibt also keine klaren Hinweise auf die Kontextabhängigkeit des Item- und des Quellengedächtnisses, auch wenn die Rohdaten in Tabelle 8.1 eine solche Vermutung vielleicht nahelegen.

Insgesamt sprechen die Ergebnisse für die Möglichkeit einer paradigmenunabhängigen Messung des Item- und des Quellengedächtnisses sowie für die Unabhängigkeit der betreffenden Maße von der jeweiligen Kontextquelle. Einschränkend muß allerdings betont werden, daß anhand der Daten von Marsh und Hicks (1998) kein strenger Test der Paradigmenunabhängigkeitshypothese vorgenommen werden kann. Dies liegt daran, daß das Gesamtmodell durch die Gleichsetzung der Item- und Quellengedächtnisparameter von 3AFC- und Ja-Nein-Tests überhaupt erst identifizierbar wird. Ein strengerer Test wäre möglich, wenn die Gedächtnisparameter verschiedener Tests separat identifizierbar wären und ihre Gleichheit anschließend durch einen teststärkeren χ^2-Differenzentest geprüft werden würde. Erreichbar wäre dies z.B. durch eine konzeptuelle Replikation der Experimente von Marsh und Hicks (1998) bei der zusätzlich eine Manipulation der Rateparameter (z.B. durch unterschiedliche Auszahlungsmatrizen) vorgenommen wird. Setzt man für ein solches Paradigma die Item- und Quellengedächtnisparameter eines Tests über die Ratetendenz-Manipulationsbedingungen hinweg gleich, so werden sie identifizierbar, ohne daß eine Gleichsetzung mit den korrespondierenden Parametern anderer Test vorgenommen werden muß.

Auch wenn ein strenger Test der Paradigmenunabhängigkeitshypothese und der Kontextunabhängigkeitshypothese für verschiedene Quellendiskriminationstests bislang noch aussteht, ist doch Grund zum Optimismus gegeben. Für diese Hypothesen sprechen nämlich nicht nur die oben referierten Reanalysen, sondern auch die schon erwähnten Ergebnisse von Meiser und Bröder (1999), die das 2HTSM-Modell für zwei Quellen mit einem analogen Modell für vier Quellen aus gekreuzten Quellendimensionen (z.B. Darbietungsort gekreuzt mit Darbietungsfont) verglichen haben (vgl. Abschnitt 7.1.2). Auch in ihrer Untersuchung zeigte sich, daß die Item- und Quellengedächtnisparameter für eine bestimmte Quelle unabhängig davon sind, ob ein 3AFC-Test für zwei Quellen oder ein kombinierter Tests für vier Quellen (zum Alt-Neu-Status, zur ersten Quellendimension und zur zweiten Quellendimension) Verwendung findet.

8.2 Inklusions- und Exklusionstests

Buchner, Erdfelder et al. (1997) haben den üblichen 3AFC-Quellendiskriminationstest mit den in Abschnitt 7.1.3 geschilderten Inklusions- und Exklusionstests aus Jacobys (1991) Prozeßdissoziationsprozedur (PDP) verglichen. Vordergründig zielt die PDP — wie bereits diskutiert — nicht auf eine Separierung von Item- und Quellengedächtnis ab, sondern stattdessen auf eine Separierung von bewußter Erinnerung (recollection) und Vertrautheit (familiarity). Dennoch ist für viele Anwendungskontexte die Annahme nicht unplausibel, daß Antworten zur Inklusions- und zur Exklusionsbedingung aus latenten (d.h. von der Vp nicht explizit ge-

äußerten) Quellenurteilen abgeleitet werden. Das übliche Zwei-Listen-Rekognitionsparadigma der PDP weist starke Ähnlichkeiten mit einem Zwei-Quellen-Diskriminationsparadigma auf: Zunächst bekommen Vpn eine Itemliste 1 dargeboten (z.B. visuell) und anschließend eine Itemliste 2 (z.B. akustisch). Danach sollen die Vpn Rekognitionsurteile zu Items beider Listen und zusätzlichen Distraktoritems abgeben. Die Inklusionsbedingung entspricht dabei einem üblichen Ja-Nein-Rekognitionstest („Ist das Item alt, d.h. war es in Liste 1 *oder* in Liste 2?"), die Exklusionsbedingung dagegen einem Ja-Nein-Quellentest zur Itemliste 2 („Ist das Item alt *und nicht* aus Liste 1, d.h. stammt das Item aus Liste 2?"). Man kann sich den Antwortprozeß leicht so vorstellen, daß die Vpn mental zunächst einen 3AFC-Test durchführen. Kommen sie unter Zugriff auf das Itemgedächtnis, das Quellengedächtnis und/oder etwaige Ratestrategien zum Ergebnis, daß das Item aus Liste 1 stammt, so leiten sie unter der Inklusionsbedingung daraus die Antwort „ja" und unter der Exklusionsbedingung die Antwort „nein" ab. Kommen sie zu dem Ergebnis, daß das Item aus Liste 2 stammt, wird unter beiden Testbedingungen „ja" geantwortet. „Neu"-Einschätzungen haben ganz analog „Nein"-Antworten in beiden Testbedingungen der PDP zur Folge.

So gesehen wären die Vp-Antworten in der Rekognitionsvariante der PDP nichts anderes als transformierte Quellenurteile. Diese Idee findet sich in ähnlicher Weise auch bei Dodson und Johnson (1996), Mulligan und Hirshman (1997) sowie Roediger und McDermott (1994).[92] Buchner, Erdfelder et al. (1997) sind jedoch insofern noch einen Schritt weiter gegangen, als sie unter identischen experimentellen Bedingungen für drei randomisierte Vpn-Gruppen Inklusionsurteile, Exklusionsurteile und 3AFC-Quellenurteile erhoben haben. Ihr erstaunliches Ergebnis war, daß die Antworthäufigkeiten für die beiden PDP-Testbedingungen aus denen zum 3AFC-Quellendiskriminationstest nahezu perfekt vorhersagbar waren, wenn die oben dargestellten Transformationsregeln zugrunde gelegt wurden (vgl. auch Tabelle 8.3). Dies bedeutet, daß die Paradigmenunabhängigkeitshypothese in diesem Fall nicht auf die Gedächtnisparameter beschränkt werden muß. Anders als im Falle der Daten von Marsh und Hicks (1998), die im letzten Abschnitt besprochen wurden, gilt die Äquivalenz zwischen 3AFC-Quellentests und Inklusions- vs. Exklusionstests offenbar nicht nur für die Gedächtnisparameter, sondern auch für die Urteils- bzw. Rateparameter. Letztlich gilt diese Äquivalenzbehauptung auch unabhängig vom verwendeten Meßmodell, denn wenn die einander korrespondierenden Kategoriewahrscheinlichkeiten in zwei Untersuchungsparadigmen identisch sind, dann müssen es auch die Modellparameter beliebiger identifizierbarer Meßmodelle sein, die auf die korrespondierenden Datensätze ange-

[92] Roediger und McDermott (1994) attribuieren diese Idee wiederum auf F.I.M. Craik (persönliche Mitteilung).

wendet werden. Diese strenge Form der Äquivalenz zwischen zwei verschiedenen Untersuchungsparadigmen soll im folgenden „Paradigmenäquivalenz" genannt werden.

Über die Gründe für die Diskrepanz zwischen den Ergebnissen von Buchner, Erdfelder et al. (1997), die Paradigmenäquivalenz fanden, und Marsh und Hicks (1998), die sie nicht fanden, kann man gegenwärtig nur spekulieren. Insbesondere ist überraschend, daß ein Ja-Nein-Quellentest zur zweiten Quelle offenbar eine Erhöhung der Zustimmungsrate relativ zum 3AFC-Quellentest zur Folge hat (Marsh & Hicks, 1998, Exp. 1 und 2), ein Exklusionstest, der ja praktisch nur eine Paraphrasierung des Ja-Nein-Quellentests zur Liste 2 ist, dagegen nicht (Buchner, Erdfelder et al., 1997). Drei Erklärungen kommen sind in erster Linie in Erwägung zu ziehen:

(1) Die sprachliche Formulierung der Testfragen ist entscheidend. Der Präsuppositionseffekt, welcher Ratetendenzen beeinflußt, greift vor allem bei expliziten Ja-Nein-Tests, nicht bei beliebigen Paraphrasierungen dieser Aufgabenstellung (z.B. bei Exklusionstests). Generell ist der Anteil der „Ja"- bzw. „Alt"-Antworten nicht nur vom *Instruktionsinhalt* abhängig, sondern auch von der *Instruktionswortwahl* (bei gleichem Inhalt).

(2) Die Manipulation der Testfragen innerhalb versus zwischen den Vpn ist entscheidend. Buchner, Erdfelder et al. (1997) haben — wie schon erwähnt — in der Inklusions- und in der Exklusionsbedingung unterschiedliche (randomisierte) Probandengruppen verwendet, während die Vpn von Marsh und Hicks (1998) auf einen Teil der Items Ja-Nein-Tests zur ersten Quelle, auf einen anderen Teil Ja-Nein-Tests zur zweiten Quelle erhielten. Möglicherweise erschwert der Wechsel der Testinstruktion die Aufgabe für die Vpn erheblich bzw. erzeugt Verwirrung bei den Vpn, so daß Urteile auf die gerade nicht gestellte Frage gegeben werden. Diese Kritik ist nicht neu; sie wurde speziell in bezug auf Inklusions- und Exklusionstests von verschiedenen Autoren geäußert (Buchner et al., 1995; Curran & Hintzman, 1995; Graf & Komatsu, 1994; Horton & Vaughan, 1999; Richardson-Klavehn, Gardiner & Java, 1995).

(3) Der Unterschied zwischen den Ergebnismustern ist nur ein scheinbarer. Die Tatsache, daß Buchner, Erdfelder et al. (1997) keinen Effekt der Testinstruktion gefunden haben, ist lediglich auf zu geringe Teststärke in ihrem Experiment zurückzuführen.

Wie die folgende Tabelle 8.3 zeigt, ist die dritte Erklärung nicht überzeugend. Der obere Teil der Tabelle 8.3 enthält in den Spalten 4 und 5 die relativen Häufigkeiten von „Ja"-Antworten zu den Inklusions- und den Exklusionstests von Buchner, Erdfelder et al. (1977) sowie in den Spalten 6 und 7 die aus 3AFC-Quellendiskriminationstests vorhergesagten Anteile von „Ja"-Antworten, wobei die o.g. Transformationsregeln verwendet wurden. Nur für zwei von drei Itemklassen

(gelesene und generierte Items aus Liste 1) ist — wie nach den Ergebnissen von Marsh und Hicks (1998) zu erwarten — der Anteil von „Ja"-Antworten im Exklusionstest größer als der aus dem 3AFC-Test vorhergesagte Anteil (d.h. der Anteil der „Liste 2"-Urteile im 3AFC-Test). Die Differenzen (.01 bzw. .03) sind allerdings deutlich geringer als die zwischen Ja-Nein-Quellentests und 3AFC-Tests bei Marsh und Hicks (1998). Auch die von Buchner, Erdfelder et al. (1997) durchgeführte Teststärkekontrolle zeigt, daß die Teststärke schon für Effekte in der Größenordnung *w* = .1 sehr gut war. Es bleibt also nur der Schluß, daß Unterschiede zwischen 3AFC-Quellenurteilen und Inklusions-Exklusions-Urteilen entweder nicht vorhanden oder aber vernachlässigbar gering waren.

Tabelle 8.3:
Relative Zustimmungsraten bei Inklusions- und Exklusionstests (beob.) und vorhergesagte Zustimmungsraten (vorherg.) auf der Basis von 3AFC-Quellendiskriminationstests. Die Daten wurden errechnet aus Tabelle 2 bei Buchner, Erdfelder et al. sowie aus Tabelle 2 und Tabelle 3 bei Steffens et al.

Autoren	Experiment	Itemklasse	Inklusion (beob.)	Exklusion (beob.)	Inklusion (vorherg.)	Exklusion (vorherg.)
Buchner,	1	gelesen	.52	.20	.53	.17
Erdfelder		generiert	.79	.14	.85	.13
et al.		neu	.16	.06	.19	.07
Steffens	1	gelesen	.58	.24	.53	.20
et al.		generiert	.84	.20	.83	.14
		neu	.21	.10	.19	.08
Steffens	2	alt	.58	.31	.60	.28
et al.		neu	.34	.27	.37	.27

In der Tabelle ebenfalls aufgeführt sind die relativen Antworthäufigkeiten aus einer neueren Arbeit von Steffens et al. (in press). Diese Daten ermöglichen ansatzweise eine Überprüfung der ersten o.g. Erklärungshypothese. Experiment 1 von Steffens et al. war nämlich eine nahezu perfekte Replikation des Experiments von Buchner, Erdfelder et al. (1997); der einzige wesentliche Unterschied bestand in der Formulierung der Testinstruktion in der Exklusionsbedingung. Während die von Buchner, Erdfelder et al. (1997) verwendete Formulierung einer Quellendiskriminationsinstruktion ähnelte („Antworten sie ‚neu' auf Items aus Liste 1 und Distraktoritems; antworten Sie ‚alt' auf Items aus Liste 2"), entsprach die von Steffens et al. verwendete Formulierung exakt der von Jacoby (1991) propagierten Instruktion: „Antworten Sie ‚alt' nur wenn das Item zuvor gehört wurde. Beachten Sie, daß in der Testliste auch zuvor gelesene oder generierte Worte enthalten sind; diese sind ‚neu' zu nennen". Wie man aus dem mittleren Teil der Tabelle erkennt, hat dieser Wechsel der Formulierung praktisch keinerlei Einfluß auf die

Differenzen zwischen beobachteten und vorhergesagten Zustimmungsraten. Dies spricht gegen die Bedeutung der Wortwahl in der Testinstruktion.

Vielleicht noch überzeugender ist jedoch Experiment 2 von Steffens et al., in dem anstelle von Rekognitionsurteilen Berühmtheitsurteile zu Personen abzugeben waren. Die Vpn bekamen zunächst eine Liste von Namen unbekannter Personen vorgelegt. Nach einem Retentionsintervall wurden die Vpn aufgefordert, eine weitere Liste von Namen daraufhin zu beurteilen, ob die betreffende Person berühmt oder nicht berühmt ist. Faktisch bestand diese zweite Liste aus (a) wirklich berühmten Personen, (b) nicht berühmten Personen, deren Namen zuvor dargeboten worden waren, und (c) nicht berühmten Personen, deren Namen komplett neu waren. Vpn der Inklusionsbedingung erhielten zusätzlich die falsche Information, daß alle Namen der zuvor dargebotenen Liste zu berühmten Personen gehören; die Vpn sollten daher ein Berühmtheitsurteil abgeben, wenn sie wußten, daß die betreffende Person berühmt ist *oder* wenn sie erinnerten, daß der Name zuvor in der ersten Liste dargeboten worden war. Vpn der Exklusionsbedingung erhielten demgegenüber die korrekte Information, daß alle zuvor dargebotenen Namen zu unbekannten Personen zählen. Sie sollten daher nur dann ein Berühmtheitsurteil abgeben, wenn sie die berühmte Person wirklich kannten. Erinnerte Namen der ersten Liste sollten sie dagegen als „nicht berühmt" einstufen.

Auch hier kann man sich wieder vorstellen, daß zunächst ein Quellenattributionsprozeß stattfindet: Die Vpn klassifizieren einen Namen zunächst als „vertraut aufgrund vorheriger Darbietung" (Quelle A), „vertraut aufgrund von Berühmtheit" (Quelle B) oder als „neu". Aus diesen Quellenurteilen leiten sie dann im nächsten Schritt die korrespondierenden Inklusions- und Exklusionsantworten ab. Um diese Annahme zu prüfen, gingen Steffens et al. genau wie in ihrem ersten Experiment vor. Sie erhoben für drei randomisierte Vpn-Gruppen 3AFC-Quellenurteile zum Vertrautheitsstatus sowie Inklusions- und Exklusionsurteile zur Berühmtheit der Personen. Die aus den 3AFC-Urteilen vorhergesagten Inklusions-Exklusions-Häufigkeiten sind unten rechts in Tabelle 8.3 zu finden, die korrespondierenden realen Urteilshäufigkeiten links daneben. Auch hier zeigt sich wieder kein bedeutsamer Unterschied zwischen vorhergesagten und beobachteten relativen Antworthäufigkeiten für Namen der ersten Liste (vorletzte Zeile) und neue Namen unbekannter Personen (letzte Zeile der Tabelle). Man kann also nicht nur in bezug auf Rekognitionsurteile, sondern auch in bezug auf Berühmtheitsurteile von einer Paradigmenäquivalenz zwischen Inklusions-Exklusions-Tests einerseits und 3AFC-Quellentests andererseits sprechen, jedenfalls soweit Items der ersten Lernphase und neue (neutrale) Distraktoritems betroffen sind.

Die Stabilität und Robustheit der Befunde von Buchner, Erdfelder et al. sowie Steffens et al. läßt vermuten, daß bei interindividueller Manipulation der In-

klusions- versus Exklusionstestbedingungen tatsächlich Paradigmenäquivalenz zu 3AFC-Quellentests vorliegt, jedenfalls für die von diesen Autoren untersuchten Enkodierungs- und Abrufbedingungen. Daß die Daten von Marsh und Hicks (1998) bei recht ähnlichen Enkodierungsbedingungen zwar für die Hypothese der Paradigmenunabhängigkeit der Item- und Quellengedächtnisparameter sprechen, die stärkere Hypothese der Paradigmenäquivalenz von Ja-Nein Quellentests und 3AFC-Quellentests jedoch nicht bestätigen, hat daher möglicherweise tatsächlich etwas mit dem permanenten Wechsel der Ja-Nein-Tests zwischen Quelle A und Quelle B innerhalb von Vpn zu tun (Erklärungsansatz 2). Ob diese Vermutung zutrifft, kann natürlich nur durch Folgeexperimente geklärt werden.

Im gegebenen Kontext ist nun folgendes interessant: Wenn Inklusions-Exklusionstests unter den geschilderten Bedingungen zu 3AFC-Quellendiskriminationstest paradigmenäquivalent sind, dann sollten Validierungsexperimente für Meßmodelle des Item- und Quellengedächtnisses, die ursprünglich für 3AFC-Quellengedächtnistests konzipiert wurden, bei entsprechender Transformation der Antwortkategorien im Prinzip auch anhand von Inklusions-Exklusions-Daten durchführbar sein. Das Meßmodell, das sich für 3AFC-Tests als besonders valide erweist, sollte sich auch bei Anwendung auf PDP-Daten als besonders valide erweisen und vice versa. Diese Einsicht ermöglicht — wie zu zeigen sein wird — ein Fortführung des Validierungsprogramms zum Ein-Hochschwellen-Modell (1HTSM-Modell), Zwei-Hochschwellen-Modell (2HTSM-Modell) und dem Niedrigschwellen-Modell (LTSM-Modell) der Quellenüberwachung, und zwar ohne Erhebung neuer Daten. Bislang liegen Validierungsdaten hierzu nur von Bayen et al. (1996) vor. Bayen et al. (1996) hatten das Itemgedächtnis über die Ähnlichkeit von alten Items und Distraktoritems und das Quellengedächtnis über die Quellenähnlichkeit manipuliert. Nur die Item- und Quellengedächtnisparameter des 2HTSM-Modells reflektierten beide Manipulationen korrekt in den Item- und Quellengedächtnisparametern, so daß Bayen et al. dieses Modell dem 1HTSM-Modell und dem LTSM-Modell vorzogen (vgl. Abschnitt 4.5).

Mangels geeigneter experimenteller Daten wurde bislang noch nicht geprüft, ob auch Manipulationen von Antworttendenzen selektiv die Rateparameter des 2HTSM-Modells beeinflussen und die beiden Gedächtnisparameter unaffiziert lassen. Ist das 2HTSM-Modell wirklich valider als die beiden anderen Meßmodelle, so sollte das Modell auch diesen Tests am besten meistern. Buchner et al. (1995) haben in der schon mehrfach erwähnten Arbeit Antworttendenzen in Inklusions- und Exklusionstests der PDP manipuliert, um Prozeßdissoziationsmodelle zu validieren (vgl. Abschnitt 7.1.3). Unter Zugrundelegung der Paradigmenäquivalenzhypothese kann man dieselben Daten aber auch dazu verwenden, verschiedene Meßmodelle des Item- und des Quellengedächtnisses miteinander zu vergleichen, so-

fern die 3AFC-Antwortkategorien in Inklusions-Exklusions-Antwortkategorien transformiert werden.

Die Übertragung der Modellgleichungen des 2HTSM-, des 1HTSM- und des LTSM-Quellendiskriminationsmodells auf Inklusions-Exklusions-Antwortkategorien erfolgt so, daß jeder Teilbaum des betreffenden Modells zunächst verdoppelt wird, so daß ein Baum sich auf die Inklusionsbedingung und der anderen Baum auf die Exklusionsbedingung bezieht. Die Parameter beider Teilbäume sind absolut identisch, nur die Antwortkategorien, in die die Äste einmünden, unterscheiden sich. Im „Inklusionsbaum" werden die 3AFC-Antwortkategorien „Quelle A" und „Quelle B" durch das Inklusionsurteil „ja" ersetzt, während die „Neu"-Kategorie der „Nein"-Antwort im Inklusionstest entspricht. Im „Exklusionsbaum" wird entsprechend nur die Antwortkategorie „Quelle B" durch das Urteil „ja" ersetzt, während die 3AFC-Antwortkategorien „Quelle A" und „neu" beide in ein „Nein"-Urteil transformiert werden.

Die Modellbeschreibungsdateien (EQN-Dateien) der auf Inklusions- und Exklusionstests adaptierten 1HTSM-, 2HTSM- und LTSM-Modelle, die für Analysen der Daten von Buchner et al. (1995) mittels MBT, GPT oder AppleTree (vgl. Abschnitt 4.7) geeignet sind, findet man in Anhang 11. Jede dieser Dateien beschreibt 16 Teilbäume des Gesamtmodells, acht für die erste und acht für die zweite Antworttendenzbedingung jedes der drei Experimente von Buchner et al. (1995). Experiment 1 kontrastierte ein Standarddistraktorset mit einem erweiterten Distraktorset, bei dem die Bereitschaft zu „Alt"- bzw. „Ja"-Urteilen in Inklusions- und Exklusionstests herabgesetzt sein sollte. Experiment 2 stellte einer liberalen Auszahlungsbedingung eine konservative gegenüber, wobei letztere die Bereitschaft zu „Ja"-Urteilen reduzieren sollte. Experiment 3 schließlich verglich eine Standardinstruktion mit einer Instruktion, welche die Gefahr falscher „Nein"-Urteile in der Inklusionsbedingung und falscher „Ja"-Urteile in der Exklusionsbedingung besonders betonte. Alle drei Experimente folgten ansonsten dem gleichen Schema: In Phase 1 wurde ein Teil der Items gelesen, ein anderer Teil aus Anagrammen generiert. In Phase 2 wurden weitere Items gehört. Im anschließenden Gedächtnistest waren Items aller drei Gruppen sowie neue Distraktoritems unter Inklusions- oder Exklusionstestbedingungen zu beurteilen.

Auf die Häufigkeiten von „Ja"- und „Nein"-Antworten zu jeder der vier Itemklassen (gelesene, generierte, gehörte und neue Items) wurden gemeinsam für beide Test- und Antworttendenzbedingungen 2HTSM-, 1HTSM- und LTSM-Modelle unter der Restriktion angepaßt, daß nur die Rateparameter der Modelle, nicht aber die Itemgedächtnis- und Quellengedächtnisparameter durch die Antworttendenzmanipulationen beeinflußt werden. Die Datei mit allen Antworthäufigkeiten (BEVRAHD.MDT) ist ebenfalls in Anhang 11 wiedergegeben. Sie entspricht den Häufigkeiten in Tabelle A1 bei Buchner et al. (1995), ist allerdings um die Ant-

worthäufigkeiten zu den akustisch dargebotenen Items aus Phase 2, die Buchner et al. (1995) nicht untersucht haben, erweitert.

Die Annahmen des 2HTSM-Quellendiskriminationsmodells wurden bereits im letzten Abschnitt 8.1 ausführlich geschildert. Im gegebenen Kontext ist das Modell auf drei alte Itemklassen (gelesene und generierte Items aus Phase 1 und gehörte Items aus Phase 2) sowie eine neue Itemklasse (Distraktoren) zu erweitern, wobei sich die Modellgleichungen kanonisch aus den Modellannahmen ergeben. Eine exakte Beschreibung des Modells findet sich in Anhang 11.

Das 1HTSM-Modell nach Batchelder und Riefer (1990) ergibt sich aus dem 2HTSM-Modell durch die Restriktion, daß neue Items nie sicher als neue Items erkannt werden ($D_3 = 0$). Es handelt sich also um ein echtes Submodell des 2HTSM-Modells. Eine genaue Beschreibung des auf Inklusions- und Exklusionstests adaptierten 1HTSM-Modells findet man ebenfalls in Anhang 11.

Tabelle 8.4:
G^2, AIC- und BIC-Anpassungsstatistiken für die auf Inklusions-Exklusions-Daten adaptierten 2HTSM-, 1HTSM- und LTSM-Modelle der Quellendiskrimination bei Anwendung auf Experiment 1 bis 3 von Buchner et al. (1995). Die Anpassungsstatistiken betreffen restringierte Modelle mit über die Antworttendenzbedingungen hinweg gleichgesetzten Itemgedächtnisparametern und Quellengedächtnisparametern. N = 6400 pro Experiment.

	2HTSM-Modell				1HTSM-Modell				LTSM-Modell			
Exp.	G^2	df	AIC	BIC	G^2	df	AIC	BIC	G^2	df	AIC	BIC
1	15.15	3	9.15	-11.1	38.02	4	30.02	2.96	60.71	4	52.71	25.65
2	8.89	3	2.89	-17.4	13.78	4	5.78	-21.3	27.53	4	19.53	-7.53
3	0.74	3	-5.26	-25.6	7.23	4	-0.77	-27.8	39.55	4	31.55	4.49
alle	24.78	9	6.78	-64.0	59.03	12	35.03	-59.3	127.8	12	103.8	9.45

Das LTSM-Modell nach Batchelder et al. (1974) entspricht dem 1HTSM-Modell mit der Ausnahme, daß bei nicht rekognizierten alten Items — d.h. beim Scheitern des Itemgedächtnisses — nicht geraten, sondern grundsätzlich „neu" geurteilt wird. Das Teilmodell für Distraktoritems gleicht dem des 1HTSM-Modells und damit dem des 2HTSM-Modells unter der Restriktion $D_3 = 0$. Auch für das LTSM-Modell liefert Anhang 11 eine exakte Modellbeschreibung in Form einer EQN-Datei.

Von primärem Interesse ist die Anpassungsgüte der restringierten Modelle, die gleiche Item- und Quellengedächtnisparameter für die beiden Antworttendenzbedingungen annehmen. Für die 1HTSM- und LTSM-Modelle wurden folglich Modellvarianten mit drei Itemgedächtnisparametern D_1(gelesene Items), D_1(generierte Items), D_2(gehörte Items), drei Quellengedächtnisparametern d_1(gelesene Items), d_1(generierte Items), d_2(gehörte Items) und drei Rateparametern b, a und g zugrunde gelegt, wobei nur die Rateparameter zwischen den bei-

den Antworttendenzbedingungen variieren durften. Daraus resultieren insgesamt $3 + 3 + 2 \cdot 3 = 12$ freie Parameter, denen pro Experiment $2 \cdot 8 = 16$ freie Kategoriewahrscheinlichkeiten gegenüberstehen, so daß der G^2-Modelltest $df = 16 - 12 = 4$ Freiheitsgrade hat. Beim Test des restringierten 2HTSM-Modells fällt ein Freiheitsgrad fort, da das Modell einen freien Parameter mehr umfaßt, nämlich den Detektionsparameter D_3 für neue Items.

Die G^2-Anpassungsstatistiken der drei Modelle sind für jedes der drei Experimente von Buchner et al. (1995) und auch für die simultane Analyse aller drei Experimente in Tabelle 8.4 zusammengefaßt. Legt man das Signifikanzniveau $\alpha = .01$ zugrunde, was bei Stichprobenumfängen der Größenordnung $N = 6400$ pro Experiment eher zu streng angesetzt ist[93], so kann das restringierte 2HTSM-Modell zumindest für Experiment 2 und 3 akzeptiert werden, das 1HTSM-Modell dagegen nur für Experiment 3. Das LTSM-Modell muß für jedes der drei Experimente eindeutig zurückgewiesen werden. Da die Modelltests unterschiedlich viel Freiheitsgrade aufweisen, wurden zusätzlich informationstheoretische Maße der Anpassungsgüte errechnet (vgl. dazu Abschnitt 4.3.6). Auch deren Ergebnisse sprechen eindeutig gegen das LTSM-Modell und favorisieren das 2HTSM-Modell im Vergleich zum 1HTSM-Modell. Legt man den BIC-Index zugrunde, der im Gegensatz zum AIC-Index konsistent ist, so erweist sich das 2HTSM-Modell als einziges Modell mit durchgängig negativen Indizes, d.h. es kann nach dem BIC-Entscheidungskriterium für alle Experimente und auch für die Kombination der drei Datensätze angenommen werden.

Der Vorteil des 2HTSM-Modells im Vergleich zum 1HTSM-Modell wird vor allem in Experiment 1 von Buchner et al. (1995) deutlich, was — wie Tabelle 8.5 zeigt — mit dem dort besonders hohen Schätzwert für den Distraktorentdeckungsparameter D_3 zusammenhängt. Die Differenz zu den D_3-Schätzwerten für Experiment 2 oder 3 sollte allerdings nicht überinterpretiert werden: Wie man ebenfalls in Tabelle 8.5 erkennt, überlappen sich die 90%-Konfidenzintervalle der D_3-Schätzwerte für alle drei Experimente, so daß die Unterschiede durchaus auf Stichprobenfehlern beruhen können.

Die ML-Parameterschätzwerte zum 2HTSM-Modell in Tabelle 8.5 zeigen ansonsten ein Muster, das trotz einiger Abweichungen mit den in Abschnitt 8.1 geschilderten Reanalysen der Experimente von Marsh und Hicks (1998) recht gut harmoniert. Auch in den Experimenten von Buchner et al. (1995) ist das Itemgedächtnis für generierte Items deutlich besser als für gelesene; die Parameterschätzwerte liegen zwar etwas unter denen in Tabelle 8.2, was jedoch auf das infolge der interpolierten akustischen Items längere Retentionsintervall zurückgeführt wer-

[93] Bereits sehr kleine Effekte der Größe $w = .05$ werden bei $N = 6400$, $\alpha = .01$ und $df = 4$ mit einer Teststärke von $1-\beta = .77$ aufgedeckt (Buchner et al., 1992). Bei $df = 3$ erhöht sich die Teststärke sogar auf $1-\beta = .82$.

den kann. Das kürzere Retentionsintervall für akustisch dargebotene Items im Vergleich zu gelesenen und generierten Items kann auch die z.T. überraschend hohen Itemgedächtnisparameterschätzungen für gehörte Items erklären, die größtenteils über denen in Tabelle 8.2 liegen.

Tabelle 8.5:
ML-Parameterschätzungen und 90%-Konfidenzintervalle für das auf Inklusions-Exklusions-Daten adaptierten 2HTSM-Modell der Quellendiskrimination bei Anwendung auf Experiment 1 bis 3 von Buchner et al. (1995).

	Experiment 1		Experiment 2		Experiment 3	
Parameter	Schätzwert	90%-Konf.	Schätzwert	90%-Konf.	Schätzwert	90%-Konf.
D_1(gelesen)	.19	.04 - .35	.45	.30 - .60	.44	.34 - .53
D_1(gener.)	.66	.58 - .73	.77	.69 - .85	.71	.65 - .76
D_2(gehört)	.57	.50 - .65	.71	.62 - .80	.71	.66 - .77
D_3(neu)	.52	.45 - .60	.39	.22 - .56	.35	.20 - .51
d_1(gelesen)	.53	.05- 1.00	.51	.18 - .85	.51	.36 - .67
d_1(gener.)	.93	.83 - 1.00	.83	.69 - .97	.76	.68 - .84
d_2(gehört)	.84	.53 - 1.00	.46	.00 - 1.00	.48	.00 - .98
b(niedrig)	.47	.39 - .56	.41	.26 - .55	.25	.17 - .34
b(hoch)	.67	.60 - .74	.59	.45 - .73	.53	.43 - .63
a(niedrig)	.00	.00 - 1.00	.41	.00 - .93	.11	.00 - .32
a(hoch)	.46	.00 - 1.00	.26	.00 - .73	.31	.03 - .59
g(niedrig)	.77	.68 - .86	.87	.80 - .93	.51	.35 - .67
g(hoch)	.74	.66 - .82	.58	.49 - .67	.85	.79 - .90

Die vielleicht auffälligste Abweichung betrifft den Quellengedächtnisvorteil generierter Items bei Buchner et al. (1995), der anhand der Daten von Marsh und Hicks (1998) sowie Erdfelder und Bredenkamp (1998) nicht bestätigt werden kann. Zu vermuten ist, daß entweder die größere Schwierigkeit der Generierungsaufgabe bei Buchner et al. oder das Niederschreiben der Items, welches Buchner et al. verlangt haben, die oben genannten Autoren jedoch nicht, für die Unterschiede verantwortlich ist. In beiden Fällen wäre zu konstatieren, daß gelesene und generierte Items nicht unter allen Umständen hinsichtlich des Quellengedächtnisses vergleichbar sein müssen: In dem Maße, in dem die Aufmerksamkeit auf den Akt des Wortgenerierens selbst gerichtet ist — sei es durch eine schwierigere Aufgabenstellung oder durch eine motorische Zusatzaufgabe — scheint nicht nur das Itemgedächtnis, sondern auch das Quellengedächtnis zu profitieren.

Die Rateparameter des 2HTSM-Modells verhalten sich im wesentlichen so, wie es aufgrund der Antworttendenzmanipulationen von Buchner et al. zu erwarten war. Die Manipulationen schlagen hypothesenkonform in erster Linie auf Pa-

rameter *b* (die Bereitschaft, „alt" zu raten) durch, sieht man einmal von dem relativ schwachen Effekt in Experiment 2 ab. Die Quellenrateparameter *a* und *g* sind demgegenüber kaum betroffen. Man kann sie zwischen den Antworttendenzbedingungen gleichsetzen, wobei allerdings Experiment 3 auszunehmen ist. Auch diese Ausnahme war zu erwarten. Reformuliert man die in der Bedingung mit hoher Antworttendenz verwendete Instruktion so, daß sie sich explizit auf Quellenurteile bezieht, dann besagt sie, daß es relativ gefahrlos ist, ein Item als zur Phase 1 gehörig zu raten, wohingegen ein Rateurteil zugunsten von Phase 2 besondere Risiken in sich birgt, weil es in der Exklusionsbedingung ein falsches „Ja"-Urteil mit relativ hoher Wahrscheinlichkeit zur Folge haben kann. Besonders der Schätzwert von *g*, aber auch der Schätzwert von *a* spiegeln diesen Quellenratebias in Experiment 3 wieder.

Analog zu den Ergebnissen in Tabelle 8.2 findet sich auch bei der Reanalyse der Daten von Buchner et al. (1995) ein stärkerer Ratebias zugunsten von Quelle A (hier: gelesene oder generierte Items aus Phase 1), sofern Rekognitionsunsicherheit vorliegt: Die *g*-Parameterschätzungen sind durchweg größer als die *a*-Parameterschätzungen. Man kann das dahingehend interpretieren, daß die Vpn korrekterweise voraussetzen, über ein gutes Itemgedächtnis für soeben gehörte Items zu verfügen. Wenn sie ein Item nicht rekognizieren können und „alt" raten, ist es daher plausibel anzunehmen, daß das Item aus der länger zurückliegenden ersten Phase des Experiments stammt. Umgekehrt wird ein rekogniziertes Item, dessen Quelle nicht bekannt ist, eher den akustisch dargebotenen Items der Phase 2 zugeordnet, weil die Vpn auch das nur mäßig ausgeprägte Quellengedächtnis für gehörte Items metakognitiv richtig einschätzen können.

8.3 Zusammenfassung

Dieses Kapitel hat Möglichkeiten der paradigmenübergreifenden Modellierung kognitiver Prozesse am Beispiel von Quellengedächtnisaufgaben untersucht. Es zeigte sich, daß Item- und Quellengedächtnis anhand verschiedener Untersuchungsparadigmen so erfaßbar sind, daß die psychologischen Schlußfolgerungen nicht vom verwendeten Paradigma abhängen. Konkret: die Wahrscheinlichkeit einer Item- oder Quellenrekognition kann mit herkömmlichen 3AFC-Tests genauso erfaßt werden wie mit Kombinationen von Ja-Nein-Quellentests (vgl. Abschnitt 8.1) oder mit Inklusions-Exklusionstests (vgl. Abschnitt 8.2). Die Grenzen paradigmenspezifischer Modellierungen, die sich im Übersichtskapitel 7 als typischer Anwendungsfall von MVB-Modellen herausgestellt haben, sind also prinzipiell überwindbar.

Durchgängig erwiesen sich Zwei-Hochschwellen-Modelle anderen Konzeptionen des Itemgedächtnisses als überlegen, was sehr gut mit den Ergebnissen zu

einfachen Ja-Nein-Rekognitionstests harmoniert (Snodgrass & Corwin, 1988). Gelegentlich geht die Ähnlichkeit verschiedener Paradigmen so weit, daß man von Paradigmenäquivalenz sprechen kann: Die in Paradigma A gewonnenen Kategoriehäufigkeiten lassen sich aus den in Paradigma B gewonnen Häufigkeiten vorhersagen. Dies wurde in Abschnitt 8.2 für Inklusions-Exklusionstests und 3AFC-Quellendiskriminationstests demonstriert. Aber auch wenn nicht von Paradigmenäquivalenz gesprochenen werden kann — wie im Falle von 3AFC-Tests und Ja-Nein-Tests des Quellengedächtnisses — sind die Gedächtnisparameter möglicherweise dennoch unabhängig vom verwendeten Paradigma erfaßbar, d.h. so, daß die psychologisch einander korrespondierenden Parameter über verschiedenen Paradigmen hinweg gleichgesetzt werden können (vgl. Abschnitt 8.1).

Die Befunde dieses Kapitels sind für die Realismus-Instrumentalismus-Debatte von großer Bedeutung. Aus instrumentalistischer Sicht muß eine paradigmenübergreifende Erfassung kognitiver Prozesse nicht zwingend gefordert werden, da kognitionspsychologische Konstrukte ohnehin nur um „nützliche Fiktionen" zur ökonomischen Beschreibung und Erklärung von Datenstrukturen sind (Herrmann, 1983). Aus realistischer Sicht muß es dagegen eine Tatsachenwelt unabhängig von den verwendeten Untersuchungsparadigmen geben, die prinzipiell auf unterschiedlichen Wegen erforschbar sein sollte. Die in diesem Kapitel referierten Ergebnisse stützen die realistische Position insofern, als sie Querverbindungen zwischen Befunden zu verschiedenen Paradigmen aufzeigen, die aus realistischer Sicht zwingend sind, aus instrumentalistischer Sicht dagegen wie eine erstaunliche Häufung von Zufällen anmuten.

9. Modelle zu Remember-Know-Prozeduren

Neben den verschiedenen Varianten der Prozeßdissoziationsprozedur von Jacoby (1991) wurden in den letzten zehn Jahren vor allem Remember-Know-Prozeduren als Instrumente zur Separierung verschiedener Gedächtnissysteme eingesetzt. Die Technik wurde ursprünglich von Tulving (1985) verwendet, um damit zu erfassen, ob der Abruf einer Information aus dem episodischen oder aus dem semantischen Gedächtnis erfolgt. Anders als Jacoby (1991) propagierte Tulving (1985) kein spezielles Untersuchungsparadigma oder Meßmodell, um die beiden Gedächtnissysteme zu trennen. Er ging vielmehr davon aus, daß die subjektive Erinnerungserfahrung (recollective experience) einer Person direkt Auskunft darüber gibt, welches Gedächtnissystem benutzt wurde. Der Abruf einer Information aus dem episodischen Gedächtnis geht demnach immer mit der bewußten Erinnerung von Aspekten einer vergangenen Episode einher, in der die Zielinformation schon einmal erfahren wurde. Dieses „autonoetische Bewußtsein" unterscheidet sich qualitativ vom „noetischen Bewußtsein", das charakteristisch für den Abruf aus dem semantischen Gedächtnis ist: Man „weiß" einen bestimmten Sachverhalt ganz einfach (z.B. seinen eigenen Namen, die Bedeutung eines Wortes etc.), ohne sich an eine bestimmte Episode erinnern zu können, in der dieser Sachverhalt erlernt wurde.

Folgerichtig postulierte Tulving (1985), daß man durch eine einfache Nachfrage zu Informationen, die eine Person reproduziert oder rekogniziert hat, in Erfahrung bringen kann, welches Gedächtnissystem benutzt wurde. Sinngemäß lautet diese Nachfrage: „Erinnern Sie die Situation der Informationsaufnahme bewußt („remember") oder wissen sie die Information ganz einfach, ohne Aspekte des Lernkontextes erinnern zu können („know")?" Wenn die betreffende Person willens ist, ihre subjektive Erinnerungserfahrung verbal auszudrücken — dies ist die einzige erforderliche Zusatzannahme —, dann spiegelt die Antwort der Vp — so Tulving (1985) — das benutzte Gedächtnissystem.

Am häufigsten ist die Remember-Know-Nachfrage in Verbindung mit Ja-Nein-Rekognitionstests angewendet worden. Sie läuft dann auf eine Erweiterung eines Ja-Nein- bzw. Alt-Neu-Tests hinaus, bei der die „Ja"- bzw. „Alt"-Kategorie in zwei Subkategorien „remember" und „know" aufgesplittet wird. Der Terminus „einfache Remember-Know-Prozedur" (kurz: RK-Prozedur) soll im folgenden für derartige Rekognitionstests mit drei Antwortkategorien „remember" (R), „know" (K) und „new" (N) reserviert werden.

In zahlreichen Experimenten wurde untersucht, wie die Häufigkeiten von R- und K-Antworten durch verschiedene Variablen beeinflußt werden (zusammenfassend: Gardiner & Java, 1993; Rajaram & Roediger, 1997). Im wesentlichen konvergieren die Befunde dahingehend, daß Einflußfaktoren auf das sogenannte

explizite Gedächtnis (z.B. Aufmerksamkeitsteilung) primär auf die R-Häufigkeit wirken, Einflußfaktoren auf das sogenannte implizite Gedächtnis (z.B. die Anzahl der Darbietungen einer Information oder ein einfaches „maintenance rehearsal") dagegen eher auf K-Häufigkeiten. Einige Autoren sind deshalb von Tulvings ursprünglicher Interpretation der Antwortkategorien abgewichen und sehen in den K-Antworten weniger den Ausdruck eines (konzeptuellen) semantischen Gedächtnisses, sondern primär Manifestationen eines perzeptuellen impliziten Gedächtnisses (Gardiner & Java, 1990; Rajaram, 1993).

Abgesehen von den Unklarheiten bezüglich der korrekten Interpretation der Gedächtnissysteme, die durch die RK-Prozedur separiert werden, hat es auch immer wieder methodische Kritik an der Prozedur selbst gegeben. Ein Kritikpunkt ist die Mehrdeutigkeit speziell des „Know"-Begriffes, der sowohl „wissen" (wissen um ein Faktum) als auch „kennen" (mit etwas vertraut sein) bedeuten kann. Strack und Förster (1995) haben u.a. darauf hingewiesen, daß diese Mehrdeutigkeit Verwirrung bei den Vpn erzeugen und artifizielle Effekte zur Folge haben kann: Ob man die K- oder die N-Antwortalternative wählt, wenn die R-Kategorie aufgrund der subjektiven Erinnerungserfahrung ausscheidet, dürfte wesentlich von der Interpretation des „Know"-Begriffes abhängen. Allerdings kann dieses Mehrdeutigkeitsproblem durch bessere und ausgiebigere Testinstruktionen entschärft werden, so daß es keine prinzipiellen Änderungen an der Prozedur erzwingt.

Ein gravierenderer Kritikpunkt wurde ebenfalls von Strack und Förster (1995) formuliert: Die einfache RK-Prozedur beruht auf der unkritischen Annahme, daß die R- und die K-Kategoriehäufigkeiten prozeß- bzw. systemrein in dem Sinne sind, daß sie jeweils durch genau ein Gedächtnissystem affiziert werden und durch sonst nichts. Wie schon in Abschnitt 2.6 besprochen, ist diese Annahme für die Psychologie unrealistisch. Typischerweise sind wir in der Psychologie mit einem Dekomponierungsproblem konfrontiert, derart, daß unterschiedliche Gedächtnis- und Urteilsprozesse auf eine bestimmte Klasse von Vp-Reaktionen Einfluß nehmen. Ein Rückschluß von einer bestimmten Reaktion auf die zugrundeliegende mentale Systemkomponente bzw. auf den verursachenden kognitiven Prozeß ist also im allgemeinen nicht möglich. Sollte dies gerade für die RK-Prozedur nicht gelten? Strack und Förster (1995) haben dies bezweifelt und deshalb Experimente durchgeführt, in denen der Einfluß von Urteilsprozessen, die bei der Auswertung der RK-Prozedur unberücksichtigt bleiben, untersucht wurde. Wie kaum anders zu erwarten zeigten ihre Ergebnisse — das sei an dieser Stelle schon vorweggenommen —, daß die Häufigkeiten von R- und vor allem von K-Antworten massiv mit Antworttendenzen kontaminiert sind.

Auf die Arbeit von Strack und Förster (1995) und verschiedene modellbasierte Ansätze zur Lösung des geschilderten Kontaminierungsproblems komme ich im nächsten Abschnitt 9.1 zurück. Abschnitt 9.1 wird auf den Vorschlag hin-

auslaufen, die RK-Prozedur durch ein geeignetes Meßmodell für die interessierenden kognitiven Prozesse zu ergänzen, welches der berechtigten Kritik von Strack und Förster (1995) standhält. Diese Konsequenz wurde von Gardiner, Richardson-Klavehn und Ramponi (1997) in ihrer Replik auf Strack und Förster (1995) nicht gezogen. Sie schlugen stattdessen eine Erweiterung der einfachen RK-Prozedur um eine „Guess"-Antwortkategorie (G-Antwort) vor. Diese und weitere denkbare Erweiterungen der einfachen RK-Prozedur werden Gegenstand des dann folgenden Abschnitts 9.2 sein. In diesem Abschnitt wird empirisch gezeigt, daß die Erweiterung der RK-Prozedur um zusätzliche Antwortkategorien die Kritik von Strack und Förster nicht gegenstandslos macht. Allerdings können auch für diese erweiterten RK-Prozeduren wieder geeignete Meßmodelle formuliert werden, die das Kontaminierungsproblem lösen. Die wesentlichen Resultate dieses Kapitels werden in Abschnitt 9.3 zusammengefaßt.

9.1 Modelle zur einfachen Remember-Know-Prozedur

Strack und Förster (1995) haben die Validität der einfachen RK-Prozedur in ähnlicher Weise geprüft, wie Buchner et al. (1995) dies für verschiedene Meßmodelle zur Prozeßdissoziationsprozedur getan haben: durch Manipulation von Antworttendenzen. In Strack und Försters erstem Experiment wurden 23 studentischen Vpn zunächst 25 häufige und 25 seltene Wörter der deutschen Sprache für jeweils 1 Sekunde dargeboten. Zusätzlich wurden am Anfang und am Ende der Wortliste jeweils vier weitere Items präsentiert, die der Absorption von Primacy- und Recency-Effekten dienten und im folgenden unberücksichtigt blieben. Den Vpn wurde gesagt, daß sie sich alle Wörter für einen späteren Gedächtnistests einprägen sollten. Nach einem 45-minütigen Retentionsintervall, das mit einer Distraktortätigkeit ausgefüllt war, bekamen die Vpn einen einfachen RK-Test vorgelegt, bei dem die 50 alten Items und zusätzlich 50 neue Items zu beurteilen waren. Auch von diesen Distraktoritems waren die Hälfte häufige, die andere Hälfte seltene Wörter der deutschen Sprache.[94] Einer Zufallsauswahl von 11 Vpn wurde vor dem Rekognitionstest wahrheitsgemäß mitgeteilt, daß die Hälfte der zu beurteilenden Items alt, die andere Hälfte neu sei (liberale Antworttendenzbedingung). Den restlichen 12 Vpn wurde dagegen suggeriert, daß nur 30% der Items alt seien; damit sollte eine geringere Bereitschaft zu „Alt"-Antworten erzeugt werden (konservative Antworttendenzbedingung).

Die Ergebnisse des Experiments sind in Tabelle 9.1 in Form von relativen Antworthäufigkeiten wiedergegeben, die aus den mittleren Urteilshäufigkeiten in

[94]Es ist zu vermuten, daß die Itemsets, die die alten bzw. die neuen Items bildeten, über Vpn-Gruppen hinweg ausbalanciert wurden. Allerdings geht das aus der Beschreibung des Experiments nicht eindeutig hervor.

Tabelle 1 von Strack und Förster (1995, S. 355) geschätzt wurden. Die absoluten Antworthäufigkeiten kann man daraus rekonstruieren, indem man jede relative Häufigkeit mit dem Produkt von Vpn und Items für die betreffende Bedingung multipliziert. In der 50%-Basisratenbedingung sind das 11 Vpn · 25 Items pro Bedingung, also 275. In der 30%-Bedingung ergibt sich bei 12 Vpn und 25 Items pro Gruppe ein Multiplikationsfaktor von 300.

Tabelle 9.1:
Relative Häufigkeiten von R-, K- und N-Antworten auf häufige versus seltene Zielitems und Distraktoritems unter zwei Antworttendenzbedingungen gemäß Experiment 1 von Strack und Förster (1995). Die Antworthäufigkeiten wurden aus den Mittelwerten in Tabelle 1 bei Strack und Förster (1995, S. 355) geschätzt.

	Hochfrequente Wörter				Niedrigfrequente Wörter			
	Zielitems		Distraktoren		Zielitems		Distraktoren	
Antwort	50% alt	30% alt	50% alt	30% alt	50% alt	30% alt	50% alt	30% alt
Remember (R)	.28	.22	.02	.02	.46	.38	.03	.03
Know (K)	.35	.24	.24	.07	.31	.28	.22	.12
New (N)	.37	.53	.74	.91	.23	.34	.75	.84

Zunächst einmal ist ein Effekt der Worthäufigkeitsvariable auf R-Antworten, nicht jedoch auf K-Antworten zu konstatieren. Dieses Ergebnis repliziert Befunde anderer Autoren (vgl. z.B. Gardiner & Java, 1993). Darüber hinaus ist jedoch unschwer erkennbar, daß die K-Antworten und tendenziell auch die R-Antworten durch die Antworttendenzmanipulation in die erwartete Richtung beeinflußt wurden, und zwar für Zielitems und Distraktoritems beider Worthäufigkeitsklassen. Bei der von Strack und Förster (1995) vorgenommenen varianzanalytischen Auswertung wurde allerdings nur der Antworttendenz-Haupteffekt auf die K-Antworten signifikant; der Haupteffekt auf die R-Antworten verfehlte die Signifikanzgrenze zum Niveau $\alpha = .05$ relativ klar. Strack und Förster (1995) interpretierten ihre Ergebnisse folgerichtig dahingehend, daß die K-, nicht aber die R-Antwortkategorie mit Antworttendenzen kontaminiert sei. Folglich könne die Häufigkeit von K-Antworten nicht als prozeßreines Maß für implizite Gedächtnisprozesse aufgefaßt werden: „It is difficult to reconcile the effect of base-rate information (...) on reports of knowing with assumptions about implicit or semantic memory, because neither system allows for inferential processes" (Strack & Förster, 1995, S. 356).

Das zweite Experiment von Strack und Förster (1995) verwendete eine etwas andere Manipulation von Antworttendenzen als ihr erstes Experiment. Eine vorgegebene Anzahl von Items des Rekognitionstests war hier als „alt" anzukreuzen und anschließend in die R- oder die K-Kategorie einzugruppieren. Ein laxes Antwortkriterium wurde dadurch induziert, daß eine relativ große Anzahl von 50

Items als „alt" anzukreuzen war, ein strenges Antwortkriterium entsprechend durch eine geringe Anzahl verlangter „Alt"-Antworten (20). Erwartungsgemäß zeigte sich, daß ein laxes Antwortkriterium den relativen Anteil der K-Antworten erhöht und den relativen Anteil von R-Antworten senkt. Die ebenfalls erhobenen Konfidenzurteile der Vpn reflektierten zudem, daß R-Antworten mit signifikant größeren Konfidenzratings einhergehen als K-Antworten. Beides stützt die Interpretation der K-Kategorie als hybride Antwortkategorie, welche besonders anfällig für Antwort- bzw. Ratetendenzen der Vpn ist.

Beide Experimente von Strack und Förster (1995) belegen somit, daß die Vp-Antworten der einfachen RK-Prozedur mit Antworttendenzen kontaminiert sind und folglich Dekomponierungsprobleme aufwerfen. Dies gilt insbesondere für die K-Kategorie. Gardiner et al. (1997) haben dieses Problem in ihrer Replik auf Strack und Förster (1995) im Prinzip zugestanden, auch wenn sie die Konsequenzen geringer gewichten als Strack und Förster (1995) dies getan haben. Ihr Alternativvorschlag läuft darauf hinaus, zusätzlich zu den Antwortkategorien R, K und N eine vierte Antwortkategorie „guess" (G) anzubieten, die die Antworttendenzmanipulation auffangen soll. „Alt"-Antworten, die subjektiv lediglich geraten sind, sollen die Vpn in diese Kategorie plazieren. Eine Replikation des ersten Experiments von Strack und Förster (1995) zeigte denn auch, daß in der erweiterten RKGN-Prozedur nur noch die Häufigkeiten von G-Antworten durch Antworttendenzmanipulationen signifikant beeinflußt werden, nicht aber die Häufigkeiten von R- und K-Antworten. Auf die Frage, ob mit dieser Prozedur eine generelle Lösung des Kontaminierungsproblems erreicht werden kann, komme ich im nächsten Abschnitt zurück.

Hier ist zunächst die Frage aufzuwerfen, ob das Dekomponierungsproblem bei der einfachen RK-Prozedur auch durch ein geeignet konstruiertes Meßmodell gelöst werden kann. Dies würde eine Erweiterung der Prozedur überflüssig machen und im übrigen auch Reanalysen der zahlreichen bereits publizierten Experimente ermöglichen, so daß entscheidbar wird, ob die Schlußfolgerung, welche die Autoren aus varianzanalytischen Auswertungen von R- und K-Antworten gezogen haben, auch bei modellbasierter Analyse der Daten noch haltbar sind. Würde man dagegen zu dem Ergebnis kommen, daß das Dekomponierungsproblem nur durch Erweiterung um ein G-Antwortkategorie gelöst werden kann, müßte man konsequenterweise schlußfolgern, daß alle Interpretationen, die auf der einfachen RK-Prozedur basieren — insbesondere die, welche sich auf die K-Häufigkeiten beziehen — von fraglicher Validität sind.

9.1.1 Das eindimensionale Kontinuum-Modell

Donaldson (1996) hat als erster ein Meßmodell zur einfachen RK-Prozedur propagiert (vgl. auch Hirshman & Master, 1997; Inoue & Bellezza, 1998). Sein eindimensionales Kontinuum-Modell (von ihm als „detection model" bezeichnet) geht von der Prämisse aus, daß die Annahme zweier qualitativ unterschiedlicher Gedächtnissysteme „hinter" den R- und K-Antwortkategorien falsch ist. In Übereinstimmung mit dem signalentdeckungstheoretischen Modell (vgl. Abschnitt 3.3) nahm Donaldson (1996) an, daß sich die Items hinsichtlich des „Vertrautheitsgrades" oder des „Bewußtheitsgrades" auf einem eindimensionalen Kontinuum unterscheiden. „Remember"- und „Know"-Antworten korrespondieren folglich nicht qualitativ unterschiedlichen Gedächtnissystemen, sondern lediglich quantitativ unterschiedlichen Gedächtnisspurenstärken. Anders ausgedrückt: Die Vpn benutzen die einfache RK-Rekognitionsprozedur exakt so, wie es die Signalentdeckungstheorie für Konfidenzratingskalen annimmt (vgl. Macmillan & Creelman, 1991, Kap. 3). Sowohl zur K-Kategorie als auch zur R-Kategorie gehören Antwortkriterien auf dem Vertrautheitskontinuum, wobei das Kriterium c_k der K-Kategorie unter dem Kriterium c_r der R-Kategorie liegt, d.h. $c_k < c_r$. Diese beiden Kriterien unterteilen das Vertrautheitskontinuum in drei Bereiche: Liegt die Vertrautheit bzw. die Erinnerungsgewißheit bezüglich eines Items unter c_k, so wird „new" geantwortet, liegt sie zwischen c_k und c_r, so wird „know" geantwortet; nur wenn c_r überschritten wird, gibt die Vp eine „Remember"-Antwort.

Legt man die üblichen Verteilungsannahmen zugrunde — d.h. normalverteilte und varianzgleiche Vertrautheitsgrade für alte Items (X_a) und für Distraktoren (X_d) — so lassen sich die Modellgleichungen des Kontinuum-Modells leicht ableiten. Wird die gemeinsame Streuung der beiden Vertrautheitsverteilungen mit 1, der Vertrautheitsmittelwert der Distraktoren mit $-d'/2$ und der von Zielitems mit $d'/2$ angesetzt[95], so ist die Wahrscheinlichkeit $p_{1,r}(d', c_r, c_k)$ für eine korrekte R-Antwort gerade der Anteil der Zielitemverteilung oberhalb von c_r:

$$p_{a,r}(d', c_r, c_k) = Pr(\{X_a \geq c_r\}) = 1 - \Phi(c_r - d'/2) = \Phi(d'/2 - c_r)\,. \qquad (9.1.1\text{-}1)$$

Hierbei bezeichnet $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung. Der Wert $z = (c_r - d'/2) / 1$ ist der z-transformierte Wert des Antwortkriteriums c_r in der Zielitemverteilung, so daß $\Phi(c_r - d'/2)$ den Anteil links vom Antwortkriterium c_r und $\Phi(d'/2 - c_r)$ den Anteil rechts davon angibt.

---

[95] Dies hat gegenüber den üblicherweise angenommenen Mittelwerten (0 für Distraktoren und d' für Zielitems, vgl. Abschnitt 3.3) den Vorteil, daß die numerischen Werte der Antwortkriterien c_r und c_k leichter zu interpretieren sind. Der Wert 0 entspricht unter diesen Umständen einem neutralen Kriterium, negative Werte entsprechen einem laxen Kriterium und positive Werte einem konservativen Antwortkriterium für „Alt"-Antworten. Die Mittelwertsdifferenz d' ist von der Skalenänderung natürlich nicht betroffen.

Die Wahrscheinlichkeit für falsche R-Antworten bei Distraktoritems läßt sich ganz analog ableiten. Bei der *z*-Transformation muß hier allerdings statt des Zielitem-Mittelwerts $d'/2$ der Vertrautheitsmittelwert der Distraktoritems ($-d'/2$) eingesetzt werden:

$$p_{d,r}(d', c_r, c_k) = Pr(\{X_d \geq c_r\}) = 1 - \Phi(c_r - (-d'/2)) = \Phi(-d'/2 - c_r)\,. \quad (9.1.1\text{-}2)$$

Diese beiden Modellgleichungen müssen nun noch durch die korrespondierenden Gleichungen für die Überschreitung des c_k-Antwortkriteriums bei alten Items und bei Distraktoren ergänzt werden. Sie geben die Wahrscheinlichkeiten einer korrekten bzw. einer falschen „Alt"-Antwort (d.h. einer korrekten R- *oder* K-Antwort) an:

$$p_{a,rk}(d', c_r, c_k) = Pr(\{X_a \geq c_k\}) = 1 - \Phi(c_k - d'/2) = \Phi(d'/2 - c_k) \quad (9.1.1\text{-}3)$$

$$p_{d,rk}(d', c_r, c_k) = Pr(\{X_d \geq c_k\}) = 1 - \Phi(c_k - (-d'/2)) = \Phi(-d'/2 - c_k)\,. \quad (9.1.1\text{-}4)$$

Die einfachste Version von Donaldsons Kontinuum-Modell erklärt somit 2·2 = 4 freie Kategoriewahrscheinlichkeiten mit drei freien Parametern d', c_r und c_k. Prinzipiell könnte also für jeden Zielitem-Distraktoritem-Datensatz ein Modellanpassungstest mit $df = 4–3 = 1$ durchgeführt werden. Interessanterweise sind derartige Modelltests bislang aber noch nie publiziert worden. Üblicherweise erfolgte die Überprüfung des Kontinuum-Modells immer so, daß aus individuellen oder aggregierten Trefferraten und den Raten falschen Alarms Schätzwerte für die Sensitivität (d' oder ein nonparametrischer Index wie A') und die Antworttendenz abgeleitet wurden, und zwar für R- und „Alt"-Antworten (R- oder K-Antworten) getrennt. Im nächsten Schritt hat man dann überprüft, ob sich diese Schätzwerte über verschiedene experimentelle Bedingungen hinweg „vernünftig" verhalten, d.h. so, wie man es bei Gültigkeit des Kontinuum-Modells auf dem Hintergrund psychologischer Erwägungen erwarten sollte. Obwohl derartige Ergebnisse natürlich nicht wertlos sind (s.u.), machen sie Modellanpassungstests dennoch nicht überflüssig. Auch ein statistisch inadäquates Modell kann in vielen Anwendungsfällen Schätzwerte generieren, die psychologisch gut interpretierbar sind. Im Anschluß an die Schilderung bisheriger Befunde werde ich deshalb Modellanpassungstests für verschiedene Varianten des Kontinuum-Modells anhand einer Reanalyse bereits publizierter Remember-Know-Experimente wie auch anhand eines neuen Experiments durchführen.

9.1.1.1 Bisherige Befunde zum Kontinuum-Modell

Es liegen eine ganz Reihe von Ergebnissen vor, die auf dem Hintergrund des Kontinuum-Modells gut interpretierbar sind. Donaldson (1996) führte eine Meta-Analyse von 80 RK-Datensätzen aus 28 Experimenten durch und fand im Durchschnitt

nur geringfügige Unterschiede der Sensitivitätsindizes *d´* bzw. *A*, wenn sie (a) ausschließlich anhand der Treffer und falschen Alarme bei R-Antworten oder (b) anhand der Treffer und falschen Alarme aller „Alt"-Antworten (d.h. der kombinierten R- und K-Antworten) geschätzt wurden. Seine Analyse schließt allerdings nicht aus, daß in einem erheblichen Teil der betrachteten Experimente dennoch bedeutsame Unterschiede vorlagen. *d´*-Schätzwertdifferenzen in der Größenordnung 0.5 und z.T. sogar noch größere Differenzen kommen in der Übersichtstabelle 1 bei Donaldson (1996, S. 528f) durchaus vor; ob sie statistisch bedeutsam sind, wurde von Donaldson (1996) nicht evaluiert.

Donaldson, MacKenzie und Underhill (1996) haben *A´*-Werte für drei verschiedene Enkodierungsbedingungen verglichen: (a) kontextfrei gelesene Items, (b) mit sprachlichem Kontext gelesene Items und (c) generierte Items. Sie fanden deskriptiv ebenfalls nur geringfügige Unterschiede der *A´*-Schätzer für R-Antworten und für alle „Alt"-Antworten, während sich die *A´*-Schätzer zwischen den experimentellen Bedingungen in der prognostizierten Richtung unterschieden: Generierte Items lieferten den größten *A´*-Schätzer, kontextfrei gelesene Items den niedrigsten. Auch in dieser Studie wurde allerdings auf eine inferenzstatistische Analyse des Unterschieds der *A´*-Schätzer für reine R-Antworten und alle „Alt"-Antworten verzichtet.

Weitere bestätigende Evidenz für das Kontinuum-Modell berichteten Inoue und Bellezza (1998). Sie fanden varianzanalytisch keine signifikanten Unterschiede zwischen individuellen *A´*-Schätzern relativ zu den Kriterien c_r (nur R-Antworten) und c_k (alle „Alt"-Antworten), und zwar unabhängig davon, ob der Lernkontext und der Testkontext ungleich oder gleich waren. Letzteres ist eine Vorhersage des Kontinuum-Modells, die auf dem Hintergrund eines Modells mit zwei distinkten Gedächtnissystemen unplausibel erscheint: Ein identischer Kontext in der Lern- und der Abrufsituation sollte insbesondere die Sensitivität der R-Antwort relativ zur Sensitivität einer globalen „Alt"-Antwort erhöhen, weil bei gleichem Kontext mehr Details verfügbar sind, die Anlaß für eine bewußte Erinnerung sein können.

Hirshman und Master (1997) haben ihre Variante des Kontinuum-Modells explizit mit der Zusatzannahme verbunden, daß die Varianz der Vertrautheitsverteilung alter Items möglicherweise größer als die der Distraktoritems ist. Dadurch umfaßt ihr Modell einen Parameter mehr, nämlich die Streuung σ für alte Items. Insgesamt stehen den vier freien Kategoriewahrscheinlichkeiten (jeweils zwei für alte Items und Distraktoren) somit auch vier zu schätzende Parameter *d´*, σ, c_r und c_k gegenüber, so daß das Modell anhand eines Datensatzes nicht mehr getestet werden kann. Hirshman und Master (1997, Figure 2) zeigten jedoch, daß alle in der Literatur berichteten einfachen und doppelten Dissoziation zwischen R- und K-Antworten durch geeignete Konstellationen obiger Parameter erklärbar sind. Na-

türlich hängen alle Ergebnismuster wesentlich von der Plazierung der Antwortkriterien ab, was erwarten läßt, daß sich insbesondere die Ergebnisse zur K-Antwortkategorie nicht immer gut replizieren lassen: Effekte bestimmter unabhängiger Variablen auf die K-Antworten, die bei einer Konstellation von Antwortkriterien auftreten, können bei einer veränderten Konstellation der Antwortkriterien verschwinden oder sich sogar umkehren. Genau dies haben Hirshmann und Master (1997; Tabelle 1) bei einer Reanalyse der Literatur bestätigt gefunden.

Neben den Antwortkriterien c_k und c_r determiniert natürlich auch die Diskriminierbarkeit alter und neuer Items die Häufigkeit von R- und K-Antworten im Kontinuum-Modell: Nimmt d' zu, so muß der relative Anteil von R-Antworten bei Zielitems zwangsläufig steigen, während der Anteil der K-Antworten konstant bleiben oder u.U. sogar sinken kann. Ein solches Ergebnis haben Conway und Dewhurst (1995) beim Vergleich verschiedener Enkodierungsbedingungen gefunden: Werden persönlichkeitsbeschreibende Adjektive (gutmütig, reizbar etc.) auf das Zutreffen für die eigene Person beurteilt, dann erhalten sie beim späteren Remember-Know-Test deutlich mehr R-Antworten als die gleichen Adjektive, wenn sie hinsichtlich des Zutreffens auf eine andere Personen beurteilt werden. Die Autoren interpretierten dies als Hinweis auf die große Bedeutung des Selbstbezugs beim bewußten Erinnern, ganz im Einklang mit der Idee autonoetischer Bewußtheit bei Tulving (1985). Hirshman und Lanning (1999) haben den gleichen Befund auf dem Hintergrund des Kontinuum-Modells betrachtet und geschlußfolgert, daß er einfach eine Konsequenz eines höheren d'-Wertes für Selbstbezugitems sein kann. Stimmt diese Idee, dann sollte es möglich sein, den R-Antwortanteil bei Selbstbezugitems durch experimentelle Manipulation des d'-Werts so zu beeinflussen, daß er größer, kleiner oder auch genauso groß wird wie der Anteil von R-Antworten bei Kontrollitems. Genau dies bestätigten Hirshman und Lanning (1999) in vier Experimenten. Beispielsweise ist der Anteil der R-Antworten bei Adjektiven, die auf das Zutreffen für die eigene Person beurteilt wurden, *kleiner* als der Anteil von R-Antworten bei Adjektiven, die (a) auf Gültigkeit für den durchschnittlichen Amerikaner und zusätzlich (b) hinsichtlich der Merkmalsvalenz beurteilt wurden (Hirshman & Lanning, 1999, Experiment 3). Durch das verlangte Doppelurteil wurde d' und damit auch der R-Anteil der Antworten für die zweite Bedingung größer als für die erste Bedingung, obwohl ein Ich-Bezug nicht gegeben war. Offenbar sind auf die eigene Person bezogene Ereignisse also nicht per se mit höheren R-Gedächtnisurteilen assoziiert.

Hirshman und Henzler (1998) haben eine weitere Vorhersage des Kontinuum-Modells geprüft: Wenn die Häufigkeit von R- und K-Antworten durch die Lage der Antwortkriterien c_r bzw. c_k determiniert wird, dann sollte es prinzipiell möglich sein, durch Manipulation von Antworttendenzen *beide* Antwortkriterien und damit sowohl K- als auch R-Antworten signifikant zu beeinflussen. Daß

Strack und Förster (1995) zwar einen signifikanten Effekt der Antworttendenzen auf K-, nicht jedoch auf R-Antworten gefunden haben, kann daher prinzipiell als Evidenz gegen das Kontinuum-Modell gewertet werden. Auszuschließen wäre dann allerdings, daß das Ausbleiben eines signifikanten Effektes auf R-Antworten nicht allein auf der mangelnden Durchschlagskraft der von Strack und Förster verwendeten UV beruht.

Um dies zu prüfen, führten Hirshman und Henzler (1998) eine konzeptuelle Replikation des Experiments von Strack und Förster (1995) mit persönlichkeitsbeschreibenden Adjektiven als Lernmaterial durch. Vor dem abschließenden RK-Rekognitionstest wurde einer Vpn-Gruppe gesagt, im Gedächtnistest seien 30% der Items alt, während der anderen Gruppe gesagt wurde, 70% der Items seien alt (statt 50% wie bei Strack und Förster). Tatsächlich umfaßte der Rekognitionstest 50% alte Items bei insgesamt 80 Items. Von den 40 alten Items wurden 20 in der vorhergehenden Lernphase für jeweils 500 Millisekunden, die anderen 20 Items für 2 Sekunden dargeboten. Die Worthäufigkeit wurde nicht manipuliert.

Die relativen Antworthäufigkeiten für die verschiedenen Versuchsbedingungen sind in Tabelle 9.2 zusammengefaßt. Die absoluten Antworthäufigkeiten kann man daraus rekonstruieren, indem man jede relative Häufigkeit mit dem Produkt von Vpn und Items für die betreffende Bedingung multipliziert. Für die beiden Gruppen alter Items sind das jeweils 32 Vpn · 20 Items = 640, für neue Items 32 Vpn · 40 Items = 1280.

Tabelle 9.2:
Relative Häufigkeiten von R-, K- und N-Antworten auf zuvor kurz bzw. lang dargebotene Zielitems und Distraktoritems unter zwei Antworttendenzbedingungen. Die Daten wurden aus Tabelle 1 bei Hirshman und Henzler (1998, S. 63) übernommen.

	Alte Items				Distraktoren	
	500 Millisekunden		2 Sekunden			
Antwort	70% alt	30% alt	70% alt	30% alt	70% alt	30% alt
Remember (R)	.21	.11	.35	.27	.11	.05
Know (K)	.43	.25	.42	.33	.35	.13
New (N)	.36	.64	.23	.40	.54	.82

Anders als Strack und Förster (1995) fanden Hirshman und Henzler (1998) einen auf dem Niveau $\alpha = .05$ signifikanten Effekt der Antworttendenzmanipulation sowohl auf K- als auch auf R-Antworten. Zusätzlich wirkte sich die Darbietungsdauer signifikant auf R-, nicht aber auf K-Antworten aus. Aus den individuellen Trefferraten und den Raten falschen Alarms wurden zudem A'-Maße der Sensitivität errechnet. Erwartungsgemäß unterschieden sie sich bei varianzanalytischer Auswertung für R-Antworten (Antwortkriterium c_r) und kombinierte R-

und K-Antworten (Antwortkriterium c_k) nicht signifikant voneinander und wurden auch durch die Antworttendenzmanipulation nicht beeinflußt. Hingegen war A' für die 2-Sekunden-Bedingung signifikant größer als für die 500-Millisekunden-Bedingung.

Eine weitere einschlägige Untersuchung stammt von Hicks und Marsh (1999). Sie manipulierten Antworttendenzen, indem sie bei einer Versuchsgruppe von 70 Vpn den RK-Rekognitionstest sequentiell durchführten (erst die Alt-Neu-Entscheidung, danach die Remember-Know-Entscheidung für als alt beurteilte Items), bei einer zweiten gleich großen Versuchsgruppe dagegen als simultanen 3AFC-Rekognitionstest mit den Alternativen R („remember"), K („know") und N („new"). Jeweils die Hälfte der Vpn beider Versuchsgruppen bekamen 56 Nomen für jeweils 1 Sekunde, die andere Hälfte für jeweils 4.5 Sekunden dargeboten. Der nach einem 12-minütigen Retentionsintervall applizierte RK-Rekognitionstest umfaßte alle 56 alten Items und zusätzlich 56 Distraktoritems.

Die relativen Antworthäufigkeiten dieses Experiments sind in Tabelle 9.3 zusammengefaßt. Wiederum kann man die absoluten Antworthäufigkeiten daraus ableiten, indem man jede relative Häufigkeit mit dem Produkt von Vpn (35) und Items pro Bedingung (56) multipliziert, d.h. mit 1960.

Tabelle 9.3:
Relative Häufigkeiten von R-, K- und N-Antworten auf zuvor kurz und lang dargebotene Zielitems und Distraktoritems bei simultanem und sukzessivem RK-Test. Die Daten wurden aus Tabelle 1 und 2 bei Hicks und Marsh (1999, S. 120) übernommen.

	1 Sekunde Darbietung				4,5 Sekunden Darbietung			
	Zielitems		Distraktoren		Zielitems		Distraktoren	
Antwort	3 AFC	sukzess.	3 AFC	sukzess.	3 AFC	sukzess.	3 AFC	sukzess.
Remember (R)	.53	.41	.10	.05	.54	.49	.06	.02
Know (K)	.24	.30	.21	.16	.30	.30	.14	.10
New (N)	.23	.29	.69	.79	.16	.21	.80	.88

Erwartungsgemäß induzierte der simultane 3AFC-Test eine liberale Antworttendenz relativ zur sequentiellen Testprozedur, und zwar sowohl bei reinen R-Antworten als auch bei allen „Alt"-Antworten: Die Raten falschen Alarms waren beim simultanen 3AFC-Test grundsätzlich größer. Erklären kann man das damit, daß die Vpn dazu tendieren, das Vertrautheitskontinuum relativ zu gegebenen Antwortkategorien in annähernd gleich große Bereiche einzuteilen (Hicks & Marsh, 1999). Beim simultanen 3AFC-Test mit drei Antwortkategorien wird das Antwortkriterium für eine K-Antwort daher eher in den unteren Bereich der Vertrautheitsverteilung plaziert, während das Antwortkriterium für eine „Alt"-

Antwort beim sequentiellen Test, der in jedem Schritt nur zwei Antwortalternativen anbietet, eher im mittleren Bereich liegen wird.

Zusätzlich zu Maßen der Antworttendenz leiteten Hicks und Marsh (1999) Schätzwerte für die Sensitivität (*A´ und d´*) aus Treffern und falschen Alarmen ab. Hypothesenkonform reflektierten diese Schätzwerte den prognostizierten Effekt der Darbietungsdauer. Dagegen unterschieden sie sich dagegen kaum zwischen dem sequentiellen und dem simultanen Test, eher schon zwischen reinen R-Antworten und allen „Alt"-Antworten. Letzterer Befund, der für die Gültigkeit des Kontinuum-Modells sehr kritisch ist, wurde inferenzstatistisch allerdings nicht geprüft.

Faßt man alle geschilderten Untersuchungen zusammen, so entsteht der Eindruck, daß zwar eine Menge deskriptiv-statistischer Evidenz zugunsten des Kontinuum-Modells gesammelt wurde, die jedoch dessen Gültigkeit nicht zweifelsfrei belegt (vgl. Hirshman, 1998). Hierauf haben Gardiner und Gregg (1997) in ihrer Kritik am Kontinuum-Modell ebenfalls hingewiesen. Sie gestehen ein, daß die Differenzen der aus reinen R-Antworten und aus allen „Alt"-Antworten abgeleiteten *A´*-Schätzer im allgemeinen klein sind, was im Einklang mit den Prognosen des Kontinuum-Modells steht. Gleichzeitig betonen sie jedoch, daß die Differenzen recht konsistent sind, da der für alle „Alt"-Antworten berechnete *A´*-Schätzwert fast immer größer als der für reine R-Antworten ausfällt. Gardiner und Gregg (1997) haben zusätzliche Experimente berichtet, die in die gleiche Richtung weisen und darüber hinaus sogar signifikante Unterschiede zwischen individuell berechneten *A´*-Schätzwerten für alle „Alt"-Antworten und reine R-Antworten lieferten. Begünstigend für ein solches Ergebnismuster scheinen nach ihren Befunden experimentelle Bedingungen zu sein, die geringe Raten von R-Antworten garantieren, beispielsweise eine hohe Darbietungsgeschwindigkeit kombiniert mit einer flachen Verarbeitung in der Enkodierungsphase.

Gardiner und Gregg (1997) haben weiterhin betont, daß die Antworthäufigkeiten der einfachen RK-Prozedur in Vergleichsstudien mit den Antworthäufigkeiten einer dreistufigen Konfidenzratingskala nicht einmal grob übereinstimmten und durch verschiedene unabhängige Variablen auch qualitativ anders beeinflußt wurden. Sie schlußfolgerten, daß es nicht gerechtfertigt sei, RK-Urteile als Maße der Vertrautheit aufzufassen, die Konfidenzratings vergleichbar seien. Dieses Argument ist allerdings nicht sehr überzeugend, denn der Grund für die von Gardiner und Gregg geschilderten Diskrepanzen zwischen RK-Urteilen und Konfidenzratings kann einfach darin begründet liegen, daß die Antwortkriterien für die Konfidenzratings nicht mit denen der RK-Prozedur (d.h. c_k und c_r) übereinstimmen. Reanalysen einiger Experimente, die Inoue und Bellezza (1998) berichtet haben, deuten tatsächlich an, daß ein Kontinuum-Modell mit ungefähr gleichen Sensitivitätsindizes *A´* sowohl für RK-Urteile als auch für Konfidenzratings gilt,

wobei die Antwortkriterien der Konfidenzratingstufen unter denen der RK-Antwortkategorien liegen. Konfidenzratings scheinen demnach eine liberalere Antworttendenz aufzuweisen, ähnlich wie simultane 3AFC-Tests im Vergleich zu einer sequentiellen RK-Prozedur (Hicks & Marsh, 1999). Unterschiedliche Antwortkriterien können prinzipiell alle beobachteten Diskrepanzen zwischen Konfidenzratings und RK-Urteilen erklären (vgl. Inoue & Bellezza, 1998).

Mehr Klarheit zur Frage der Gültigkeit des Kontinuum-Modells können nur Modellanpassungstests bringen. Das einfache Kontinuum-Modell, welches normalverteilte Vertrautheitswerte mit homogenen Varianzen für alte und neue Items annimmt, läßt sich prinzipiell an jedem Zielitem-Distraktoritem-Datensatz überprüfen. Die meisten Proponenten des Kontinuum-Modells scheinen jedoch die Modellvariante von Hirshman und Master (1997) zu bevorzugen, die eine größere Vertrautheitsstreuung bei alten Items im Vergleich zu neuen zuläßt und deshalb einen Parameter mehr umfaßt (vgl. z.B. Hicks & Marsh, 1999; Inoue & Bellezza, 1998). Der Nachteil dieser Modellvariante ist, daß sie für einen einfachen Zielitem-Distraktoritem-Datensatz saturiert und deshalb nicht mehr prüfbar ist. Zur Modellprüfung kommen deshalb nur solche Datensätze in Frage, in denen die Antwortkriterien in mindestens zwei Stufen variiert wurden. Derartige Datensätze liefern mindestens acht freie Kategoriewahrscheinlichkeiten, nämlich zwei für Zielitems und zwei Distraktoren für jede von zwei Antworttendenzbedingungen. Da die Manipulation von Antworttendenzen nur die Lage der beiden Antwortkriterien c_r und $c_{k,}$ nicht aber die Vertrautheitsverteilungen selbst affizieren sollte, umfaßt das anzupassenden Modell bei zwei Antworttendenzbedingungen lediglich sechs Parameter (nämlich d', σ und zusätzlich $2 \cdot 2 = 4$ Antwortkriterien). Für einen solchen Versuchsplan ist das Modell somit bei $df = 2$ testbar.

Drei der oben geschilderten Untersuchungen haben Antworttendenzen in der einfachen RK-Prozedur so manipuliert, daß die Hirshman-Master-Modellvariante prüfbar wird, nämlich Strack und Förster (1995), Hirshman und Henzler (1998) und Hicks und Marsh (1999). Die relativen Antworthäufigkeiten dieser Experimente sind in den Tabellen 9.1 bis 9.3 zusammengefaßt. Alle drei Experimente wurden als intentionale Lernexperimente mit relativ kurzen Retentionsintervallen durchgeführt. Es ist deshalb anzunehmen, daß auch bei den „schweren" Experimentalbedingungen (d.h. bei sehr kurzen Darbietungszeiten bzw. bei hochfrequenten Wörtern) nicht die Voraussetzungen erfüllt waren, welche nach den Ergebnissen von Gardiner und Gregg (1997) besonders kritisch für das Kontinuum-Modell sind. Tatsächlich ist die Rate der Remember-Urteile bei Zielitems in allen geschilderten experimentellen Bedingung größer als 0.20, so daß kaum von Experimenten mit „little or no remembering" (Gardiner & Gregg, 1997) gesprochen werden kann.

Um einen möglichst strengen Test des Kontinuum-Modells zu gewährleisten, habe ich ein weiteres Experiment durchgeführt, dessen Ziel es u.a. war, sehr niedrige Anteile von Remember-Urteilen unter wenigstens einer experimentellen Bedingung sicherzustellen. Im nächsten Abschnitt 9.1.1.2 werde ich dieses Experiment und die daraus resultierenden relativen und absoluten Antworthäufigkeiten zunächst schildern, bevor ich in Abschnitt 9.1.1.3 Modellanpassungstests und Parameterschätzergebnisse für alle vier Experimente gemeinsam darstelle.

9.1.1.2 Eine neue Untersuchung zum Kontinuum-Modell

Anders als die im letzten Abschnitt geschilderten Untersuchungen war das in diesem Abschnitt zu schildernde Experiment kein intentionales Lernexperiment. Die Vpn bekamen zwei Wortlisten mit jeweils 25 Wörtern schriftlich vorgelegt und wurden instruiert, diese Wortlisten in unterschiedlicher Weise zu bearbeiten. Das Wortmaterial war mit dem von Buchner et al. (1995) verwendeten Material identisch. Es bestand ausschließlich aus fünfbuchstabigen deutschen Nomen. Die beiden Bearbeitungsbedingungen entsprachen weitgehend den von Gardiner, Java und Richardson-Klavehn (1996, Experiment 1) verwendeten „tiefen" und „flachen" Verarbeitungsbedingungen: Für eine Wortliste wurde gefordert, daß die Vpn zu jedem Nomen ein anderes Nomen aufschrieben, das ihnen spontan dazu einfiel („tiefe" Verarbeitungsbedingung). Für die zweite Wortliste wurde verlangt, daß die Vpn zu jedem Wort zwei Buchstaben notierten, die nicht in dem Nomen enthalten waren („flache" Verarbeitungsbedingung). Die notierten Buchstaben durften sich bei aufeinanderfolgenden Worten nicht wiederholen. Pro Wort wurden 5 Sekunden Zeit veranschlagt. Nach jeweils 5 Sekunden wurde das von einem Tonband gesteuerte akustische Signal „weiter!" dargeboten, woraufhin sich die Vpn dem nächsten Wort zuzuwenden hatten. Das akustische Signal diente einerseits dem Zweck, daß alle Vpn gleich viel Zeit für die Bearbeitung jedes Items aufwendeten. Darüber hinaus fungierte es als akustische Störbedingung, die das langfristige Behalten des zuletzt bearbeiteten Wortes erschweren sollte. Insgesamt dauerte die Akquisitionsphase etwa acht Minuten.

Nach einem zweieinhalbstündigen Retentionsintervall, in dem eine Lehrveranstaltung stattfand, die thematisch keine Berührungspunkte zur Gedächtnispsychologie aufwies, folgte der für die Vpn überraschende schriftliche Rekognitionstest. Der Test umfaßte die 50 alten Items aus der Akquisitionsphase und zusätzlich 50 neue Distraktoritems in zufälliger Reihenfolge. Items und Versuchsbedingungen wurden über die Vpn hinweg so ausbalanciert, daß jedes Item gleich häufig in jeder Versuchsbedingung (flache Verarbeitung, tiefe Verarbeitung, Distrak-

toritem) vorkam. Auch die Reihenfolge, in der die tiefe und die flache Verarbeitungsbedingung absolviert wurden, wurde über die Vpn hinweg ausbalanciert.

Zwei Testformate fanden Verwendung: Die Vpn wurden zufällig einer einfachen RK-Prozedur (RK-Test) oder einer erweiterten RK-Prozedur (RKG-Test) zugeteilt, bei der als alt beurteilte Items nicht nur in die Kategorien „bewußt erinnert" (R) und „gewußt" (K), sondern zusätzlich auch noch in die Kategorie „geraten" (G) eingruppiert werden konnten. Die korrekte Verwendung dieser Antwortkategorien wurde den Vpn vor dem Test ausführlich erläutert. Die verwendete Instruktion entsprach der von Gardiner favorisierten Instruktion (vgl. Gardiner & Java, 1990, 1993; Gardiner et al., 1996, 1997). Sowohl für den einfachen RK-Test als auch für den erweiterten RKG-Test wurde eine sukzessive Testprozedur gewählt: Die Vpn sollten zunächst die Wörter einkreisen, die sie für „alt" hielten; danach sollten sie durch Angabe des Buchstabens R oder K (beim erweiterten Test: R, K oder G) kundtun, welche Erinnerungserfahrung sie mit dem betreffenden Wort verbanden. Leider vergaßen einige Vpn, zu einigen eingekreisten Wörtern eine Zusatzangabe zu machen, wodurch es zu geringfügigen Datenverlusten kam (vgl. Tabelle 9.4).

Jede der beiden Versuchsgruppen wurde noch einmal zufällig in zwei Untergruppen aufgeteilt, denen unterschiedliche Antworttendenzen induziert wurden. Dies geschah dadurch, daß den drei besten Vpn im Gedächtnistest Geld- oder Buchpreise im Wert von 10 DM bis 100 DM in Aussicht gestellt wurden. Die Güte der Gedächtnisleistung wurde dabei anhand einer bestimmten Punktwertung gemessen, die den Vpn kurz vor Durchführung des Gedächtnistests mitgeteilt wurde. Vpn der liberalen Antworttendenzbedingung wurde gesagt, daß sie für jede richtige „Alt"-Antwort (d.h. für jeden Treffer) vier Pluspunkte bekämen, für jeden „Verpasser" dagegen vier Minuspunkte. Falsche Alarme wurde demgegenüber mit nur einem Minuspunkt bestraft und korrekte Zurückweisungen mit einem Pluspunkt belohnt. Maßgeblich für die Punktvergabe waren also nur die Alt-Neu-Entscheidungen, nicht das R-, K- oder eventuell das G-Urteil nach einer „Alt"-Einschätzung.

Versuchspersonen der konservativen Antworttendenzbedingung bekamen demgegenüber kurz vor dem Gedächtnistest die umgekehrte Instruktion: Falsche Alarme hatten vier Punkte Abzug zur Folge, korrekte Zurückweisung vier Pluspunkte. Treffer brachten in dieser Bedingung nur einen Pluspunkt, Verpasser dagegen einen Minuspunkt. Insgesamt war es somit ratsamer, im Zweifelsfalle eher „neu" zu urteilen, weil man dadurch im Durchschnitt 1.5 Punkte gewinnen konnte. Genau umgekehrt war es in der liberalen Bedingung: „Neu"-Urteile gingen hier im Durchschnitt mit 1.5 Minuspunkten, „Alt"-Urteile dagegen mit 1.5 Pluspunkten einher.

Insgesamt wurde also ein komplett gekreuzter dreifaktorieller Versuchsplan realisiert, wobei eine unabhängige Variable innerhalb der Vpn variiert wurde (UV1: tiefe versus flache Itemverarbeitung in der Akquisitionsphase) und zwei unabhängige Variablen zwischen den Vpn (UV2: einfacher RK-Test versus erweiterter RKG-Test; UV3: liberale versus konservative Antworttendenzbedingung). Als abhängige Variablen wurden die Antworthäufigkeiten in den beiden RK-Gedächtnistests getrennt nach den Bedingungen erhoben.

Das Experiment fand in Form von drei Gruppenuntersuchungen statt, die an Einführungslehrveranstaltungen am Psychologischen Institut der Universität Bonn (durchgeführt im Wintersemester 1996/1997 und 1997/1998) und an einen Blockkurs einer Heilpraktikerschule in Bonn gekoppelt waren (durchgeführt im Sommersemester 1997). Die insgesamt 126 Vpn waren also entweder Studienanfänger im Fach Psychologie (N = 99; Durchschnittsalter ca. 21 Jahre) oder Heilpraktikerschüler (N = 27; Durchschnittsalter ca. 30 Jahre). Sie nahmen an der Untersuchung freiwillig teil, wobei den drei besten Teilnehmern pro Gruppe ein Geld- oder ein Buchpreis angekündigt wurde. Es blieb zu Beginn der Untersuchung zunächst unspezifiziert, nach welchem Leistungskriterium die Preise vergeben werden. Fast alle Besucher der genannten Lehrveranstaltungen nahmen daraufhin am Experiment teil. Lediglich vor dem zweiten Teil des Experiments — dem unangekündigten Gedächtnistest, der nach der zweieinhalbstündigen Lehrveranstaltung durchgeführt wurde — brachen einige Studierende, die am ersten Teil teilgenommen hatten, aus Termingründen die Untersuchung ab (ca. 10% des Gesamtteilnehmerkreises). Ihre Daten blieben bei der Auswertung und bei der Vergabe der Preise unberücksichtigt.

Tabelle 9.4:
Relative Häufigkeiten (in Klammern) und absolute Häufigkeiten von R-, K- und N-Antworten (ohne Klammern) auf zuvor tief bzw. flach verarbeitete Zielitems und Distraktoritems bei Induktion von Antworttendenzen zugunsten einer „Alt"- bzw. einer „Neu"-Antwort über Manipulation von Auszahlungsmatrizen. Die Schilderung der zugrundeliegenden Untersuchung erfolgt im Text.

	Alte Items				Distraktoren	
	flache Verarbeitung		tiefe Verarbeitung			
Antwort	pro alt	pro neu	pro alt	pro neu	pro alt	pro neu
Remember (R)	166 (.22)	141 (.17)	325 (.42)	325 (.39)	279 (.18)	220 (.13)
Know (K)	184 (.24)	162 (.20)	146 (.19)	133 (.16)	344 (.22)	282 (.17)
New (N)	421 (.55)	514 (.63)	295 (.39)	365 (.44)	920 (.60)	1130 (.69)

Die Vpn wurden den vier Gedächtnistestbedingungen, die zwischen den Vpn variiert wurden, zufällig zugeteilt, wobei eine Gleichverteilung über die Bedingungen nicht erzwungen wurde. Insgesamt absolvierten 64 Vpn den einfachen RK-Test, davon 31 in der liberalen Antworttendenzbedingung („pro alt") und 33

Vpn in der konservativen Antworttendenzbedingung („pro neu"). Weitere 62 Vpn bearbeiteten den erweiterten RKG-Test, davon 36 in der liberalen Bedingung und 26 in der konservativen Bedingung. Für jede der drei Gruppenuntersuchungen wurden Preise für die drei Besten der Punktwertung vergeben.

Vor der eigentlichen Datenanalyse wurde mittels Varianzanalysen geprüft, ob sich die drei Untersuchungsgruppen aus dem Wintersemester 1996/1997, dem Sommersemester 1997 und dem Wintersemester 1997/1998 hinsichtlich der Antworthäufigkeiten in den Gedächtnistests unterschieden. Für keine einzige Antwortkategorie wurde ein Haupteffekt der Untersuchungsgruppe oder eine Wechselwirkung der Untersuchungsgruppe mit einer anderen UV relativ zum Niveau $\alpha = .05$ gefunden. Deshalb wurden die Daten aller drei Untersuchungsgruppen aggregiert.

Tabelle 9.4 stellt die aggregierten Antworthäufigkeiten für die einfache RK-Prozedur zusammen. Die Ergebnisse zur erweiterten RKG-Prozedur werden erst in Abschnitt 9.2 berichtet (vgl. Tabelle 9.8). Die maximale Summe der Antworthäufigkeiten beträgt für flach und tief verarbeitete alte Items in Tabelle 9.3 jeweils 25 Items · 31 Vpn = 775 (liberale Antworttendenzbedingung) bzw. 25 Items · 33 Vpn = 825 (konservative Bedingung). Für Distraktoritems ist die maximale Summe jeweils doppelt so groß. Da — wie schon erwähnt — einige Vpn bei einigen Items das RK-Urteil vergaßen, wurden die Maximalsummen nicht ganz erreicht.

Schon ohne inferenzstatistische Analysen ist erkennbar, daß die Enkodierungsbedingung vor allem auf die R-Urteile einwirkt: Bewußte Erinnerungen kommen bei tiefer Verarbeitung (freie Assoziation von Nomen) erwartungsgemäß häufiger vor als bei oberflächlicher Verarbeitung in der Akquisitionsphase (Entdecken von zwei Buchstaben). Genau das Gegenteil gilt für „Neu"-Urteile, während K-Antworten von der Verarbeitungstiefemanipulation kaum betroffen zu sein scheinen. Bei varianzanalytischer Auswertung der Antworthäufigkeiten[96] wird der Verarbeitungstiefeeffekt allerdings für N-Urteile ($F(1, 62) = 85.41$), für R-Urteile ($F(1,62) = 126.53$) und — in umgekehrter Richtung — auch für K-Urteile ($F(1, 62) = 8.98$) auf dem 5%-Niveau signifikant. Letzteres weicht von dem von Gardiner et al. (1996, Experiment 1) berichteten Ergebnismuster ab, wobei allerdings zu bedenken ist, daß dort eine erweiterte RK-Prozedur mit der zusätzlichen Antwortkategorie „guess" Verwendung fand (vgl. Abschnitt 9.2).

Auch die Antworttendenzmanipulation zeigt deskriptiv die erwarteten Effekte sowohl auf R- als auch auf K-Antworten, wobei der Effekt allerdings vergleichsweise schwach ausfällt. Dies gilt insbesondere für die tiefe Verarbei-

[96] Es wurde pro Vp für jede der drei Antwortkategorien R, K und N die Häufigkeit ausgezählt. Für jede der drei Häufigkeitsvariablen wurde dann eine 2×2-Varianzanalyse mit einem meßwiederholten Faktor (Verarbeitungstiefe) und einem nichtmeßwiederholten Faktor (Antworttendenzbedingung) durchgeführt.

tungsbedingung. Varianzanalytisch zeigt sich weder für R- noch für K-Antworten ein auf dem 5%-Niveau signifikanter Haupteffekt; für N-Antworten wird der kritische Wert für das 5%-Niveau knapp verfehlt ($F(1, 62) = 3.00$). Wechselwirkungen zwischen Verarbeitungstiefe und Antworttendenzbedingungen werden für keine AV signifikant. Distraktoritems reflektieren die Antworttendenzmanipulation in der erwarteten Richtung. Führt man einseitige *t*-Tests durch, so wird die *t*-Statistik für N-Antworten ($t(62) = 1.93$) und für R-Antworten ($t(62) = 1.85$) auf dem Niveau $\alpha = .05$ signifikant, während die *t*-Statistik für K-Antworten ($t(62) = 1.52$) das 5%-Niveau knapp verfehlt.

Offensichtlich zeigen also sowohl die Verarbeitungstiefemanipulation als auch die Antworttendenzmanipulation die erwarteten Effekte auf die Antworthäufigkeiten, wobei diese allerdings nicht durchgängig statistisch signifikant werden. Interessanterweise scheint in dieser Untersuchung die R-Antwortkategorie mindestens so stark durch Antworttendenzen affiziert zu werden wie die K-Kategorie. Dies steht zwar im Widerspruch zum Ergebnis von Strack und Förster (1995), deckt sich aber mit dem Ergebnis von Hirshman und Henzler (1998). Wir müssen schließen, daß in Abhängigkeit davon, welche Antworttendenzmanipulationen vorgenommen werden, K-Antworten, R-Antworten oder auch beide Antwortkategorien mit Antworttendenzen kontaminiert sein können. Die Bedeutung dieser Ergebnisse für das Kontinuum-Modell wird im nächsten Abschnitt analysiert.

9.1.1.3 Statistische Evaluation des Kontinuum-Modells

Dieser Abschnitt hat die statistische Evaluation zweier Varianten des Kontinuum-Modells zum Gegenstand. Grundlage hierfür sind die im letzten Abschnitt berichteten Daten (vgl. Tabelle 9.4) sowie die in den Tabellen 9.1 bis 9.3 wiedergegebenen Datensätze von Strack und Förster (1995), Hirshman und Henzler (1998) und Hicks und Marsh (1999). Die beiden Varianten des Kontinuum-Modells, die im folgenden betrachtet werden, sind (a) das einfache Modell, wie es durch die Modellgleichungen (9.1.1-1) bis (9.1.1-4) definiert wird, und (b) das erweiterte Modell von Hirshman und Master (1997), welches ungleiche Streuungen der Vertrautheitsverteilungen von Distraktoritems und alten Items zuläßt. Die Modellgleichungen des erweiterten Modells werden hier nicht extra aufgeführt; sie ergeben sich aus den Modellgleichung des einfachen Kontinuum-Modells, indem man in Gleichung (9.1.1-1) den Term $\Phi(d' / 2 - c_r)$ durch $\Phi((d' / 2 - c_r) / \sigma)$ und in Gleichung (9.1.1-3) analog $\Phi(d' / 2 - c_k)$ durch $\Phi((d' / 2 - c_k) / \sigma)$ ersetzt. Hierbei ist σ die Streuung der Vertrautheitsverteilung für alte Items. Folglich ist $(d' / 2 - c_r) / \sigma$ nichts anderes als der *z*-Wert zum Antwortkriterium c_r und analog $(d' / 2 - c_k) / \sigma$ der entsprechende *z*-Wert zum Antwortkriteriums c_k. Die beiden Modellgleichungen (9.1.1-2) und

(9.1.1-4) für Distraktoritems gelten unverändert, d.h. die Streuung der Distraktoritemverteilung wird auch im erweiterten Modell auf 1 fixiert.

Tabelle 9.5:
G^2-Modellanpassungstatistiken für das einfache Kontinuum-Modell (3. Spalte), ML-Parameterschätzungen für das saturierte erweiterte Kontinuum-Modell (Spalten 4 - 7) und G^2-Modellanpassungsstatistiken für verschiedene restringierte Varianten des erweiterten Kontinuum-Modells (Spalten 8 und 9) bei Anpassung der Modelle an Datensätze aus vier verschiedenen Untersuchungen. Statistiken mit einem * sind auf dem Niveau $\alpha = .05$ signifikant; mit ** markierte Statistiken sind auf dem Niveau $\alpha = .001$ signifikant.

	einfaches	erweitertes Kontinuum-Modell					
	Modell	ML-Parameterschätzungen				Restr. 1	Restr. 2
Experiment	$G^2_{(1)}$	d'	σ	c_k	c_r	$G^2_{(2)}$	$G^2_{(6)}$
Strack & Förster (1995)							
häufige W. 50% alt	4.89*	1.12	1.51	0.08	1.46		
häufige W. 30% alt	0.06	1.23	1.08	0.71	1.44	0.92	
seltene W. 50% alt	3.85*	1.75	1.48	–0.20	1.02		
seltene W. 30% alt	0.31	1.48	1.15	0.27	1.09	1.96	19.78*
Hirshman & Henzler (1998)							
500 msec 70% alt	0.18	0.45	0.97	-0.12	1.00		
500 msec 30% alt	1.66	0.62	0.84	0.61	1.34	2.49	
2 s 70% alt	0.00	0.84	1.00	-0.32	0.81		
2 s 30% alt	1.78	1.13	0.84	0.35	1.08	16.53**	123.53**
Hicks & Marsh (1999)							
1 s 3AFC Test	3.67	1.37	1.19	-0.19	0.60		
1 s sukz. Test	0.68	1.40	1.07	0.11	0.94	2.37	
4,5 s 3AFC Test	6.54*	1.63	0.80	0.02	0.74		
4,5 s sukz. Test	0.33	2.03	1.06	0.16	1.04	26.63**	263.48**
eigene Daten aus Tabelle 9.4							
flach pro alt	0.00	0.13	0.99	0.18	0.85		
flach pro neu	0.02	0.18	0.98	0.41	1.01	0.51	
tief pro alt	5.20*	0.65	1.38	-0.08	0.59		
tief pro neu	5.75*	0.71	1.47	0.15	0.75	0.77	183.17**

Die Anpassung der beiden Modellvarianten an die o.g. Datensätze erfolgte mit der ML-Methode. Praktisch bewerkstelligt wurde die ML-Parameterschätzung

durch Minimierung der G^2-Anpassungsstatistik für einen gegebenen Datensatz mit dem PRAXIS-Programm von Gegenfurtner (1992). Hierbei resultiert quasi als Nebenprodukt auch gleich die G^2-Modellanpassungsstatistik des betreffenden Modells (vgl. dazu die Abschnitte 4.2.1 und 4.7).

Das einfache Kontinuum-Modell wurde an jeden einzelnen Distraktoritem-Zielitem-Datensatz aus den vier o.g. Untersuchungen getrennt angepaßt. Das Modell ist dann — wie schon erwähnt — bei $df = 1$ testbar. Die entsprechenden Modellanpassungsstatistiken sind in der dritten Spalte von Tabelle 9.5 aufgeführt. Wie man sieht, sind fünf der insgesamt 16 Modellanpassungsstatistiken auf dem Niveau $\alpha = .05$ signifikant. Keines der signifikanten Ergebnisse ist für sich genommen ein zwingender Anlaß, das einfache Kontinuum-Modell zu verwerfen. Abgesehen von Strack und Försters (1995) Experiment sind die zugrundeliegenden Stichprobenumfänge nämlich so groß, daß schon sehr kleine Abweichungen vom Modell mit hoher Teststärke aufgedeckt werden, wenn $\alpha = .05$ zugrunde gelegt wird. Dennoch erscheint es nicht ratsam, allen Folgeanalysen das einfache Kontinuum-Modell zugrundezulegen. Metaanalytisch betrachtet sind 5 von 16 Datensätzen, die auf dem Niveau $\alpha = .05$ signifikante Modellanpassungsstatistiken liefern, einfach zu viel, um an der Nullhypothese der Modellgültigkeit für alle 16 Datensätze festzuhalten (vgl. Bredenkamp, 1980, S. 35). Unter der globalen Nullhypothese, daß nur 5% aller Modellanpassungsstatistiken signifikant werden, ist — wie man unter Verwendung der Binomialverteilung errechnen kann — die Wahrscheinlichkeit für 5 oder mehr signifikante Ergebnisse bei 16 durchgeführten Tests lediglich $p = 0.009$. In Anbetracht dieses Ergebnisses erscheint es ratsam, die Folgeanalysen auf der Grundlage des erweiterten Kontinuum-Modells von Hirshman und Master (1997) durchzuführen.

Da das erweitere Modell für einzelne Zielitem-Distraktoritem-Datensätze saturiert ist, sind die Modellanpassungsstatistiken trivialerweise perfekt. Deshalb werden in Tabelle 9.5 nur die ML-Parameterschätzergebnisse aufgeführt und die G^2-Statistiken fortgelassen. Wie man sieht, reflektieren die Schätzwerte die experimentellen Manipulationen in der erwarteten Richtung, d.h. die Enkodierungs- und Wortmaterialmanipulationen affizieren in erster Linie d', die Antworttendenzmanipulationen dagegen die Lage der Antwortkriterien. Generell besteht eher eine Tendenz zu konservativem Antwortverhalten, da fast alle Antwortkriterien im positiven Bereich liegen. Deutlich wird auch, daß die im vierten Experiment angestrebte Erschwerung der Erinnerungsaufgabe (vgl. Abschnitt 9.1.1.2) voll gelungen ist: Die d'-Schätzwerte liegen deutlich unter denen der anderen Experimente und in der flachen Verarbeitungsbedingung sogar sehr nahe bei null. Etwas überraschend ist, daß die Streuung der Zielitemverteilung nicht immer auf einen Wert größer als 1 geschätzt wird. In einigen Fällen liegt der

Schätzwert unter 1, allerdings zumeist nicht signifikant, wie ein Blick auf die Anpassungsindizes zum einfachen Kontinuum-Modell zeigt.

Testbar wird das erweiterte Kontinuum-Modell, wenn man es auf zwei Zielitem-Distraktoritem-Datensätze bezieht, die sich nur in bezug auf Antworttendenzen unterscheiden. Dann kann man nämlich bei $df = 2$ die Hypothese testen, daß sich die Vertrautheitsverteilungen und damit d' und σ zwischen diesen beiden Datensätzen nicht unterscheiden. Die Modellanpassungsstatistiken zu dieser Hypothese sind in der zweitletzten Spalte von Tabelle 9.5 aufgeführt. Wie man sieht, paßt das restringierte Modell in allen vier Experimenten für die Versuchsbedingungen mit niedrigen d'-Schätzwerten gut auf die Daten. Demgegenüber muß es für zwei der vier Versuchsbedingungen mit guten Diskriminationsindizes relativ zum Niveau $\alpha = .001$ zurückgewiesen werden. Dieses Signifikanzniveau ist bei den gegebenen Stichprobenumfängen vernünftig, d.h. triviale Modellverletzungen führen nur mit geringer Wahrscheinlichkeit zu signifikanten Anpassungsstatistiken.

Das Ergebnismuster der Anpassungsstatistiken ist insofern überraschend, als es der Hypothese von Gardiner und Gregg (1997) widerspricht, wonach das Kontinuum-Modell insbesondere bei niedrigen bewußten Erinnerungsraten auf Probleme stoßen sollte. Die hier erzielten Ergebnisse stützen eher die umgekehrte Vermutung: Je höher die bewußten Erinnerungsraten, desto schlechter paßt das erweiterte Kontinuum-Modell auf die Daten. Dieses Ergebnismuster ist aus der Perspektive eines Zwei-Prozeß- oder Zwei-System Modells auch durchaus plausibel: Wenn bewußte Erinnerungsraten sehr niedrig sind, kann es eigentlich kaum überraschen, daß ein Ein-Prozeß-Modell, welches nur die Vertrautheit von Items berücksichtigt und bewußte Erinnerungen als zweite Erinnerungsquelle nicht in Rechnung stellt, dennoch recht gut auf die Daten paßt. Die Grenzen eines solchen Modells sollten umso eher aufzuzeigen sein, je mehr bewußte Erinnerungen zur Rekognition von Items beitragen. Genau dies deutet sich in o.g. Daten an. Daß es sich hierbei um keine zufälligen Modellverletzungen handelt, wird bei der Betrachtung der bedingungsspezifischen d'-Schätzwerte zum unrestringierten Modell deutlich: Der d'-Schätzwert bei konservativem Antwortbias ist mit nur einer Ausnahme immer größer als der entsprechende d'-Schätzwert bei liberalem Antwortbias. Die Residuen zeigen also ein systematisches Muster, welches die Hypothese einer generellen Gültigkeit des erweiterten Kontinuum-Modells erheblich in Frage stellt.

Ganz rechts in Tabelle 9.5 sind die Modellanpassungsstatistiken für ein noch stärker restringiertes erweitertes Kontinuum-Modell aufgeführt. Dieses Modell nimmt zusätzlich zu dem zuletzt geschilderten Modell an, daß auch die Enkodierungs- bzw. Worthäufigkeitsbedingungen keine Effekt auf d' bzw. σ ausüben. Erwartungsgemäß ist dieses Modell für alle vier Experimente zurückzuweisen.

Dies bedeutet, daß die Manipulation der Enkodierungsbedingungen bzw. der Worthäufigkeit — modellimmanent betrachtet — signifikante Auswirkungen auf die Vertrautheitsverteilung der Zielitems hat. Allerdings darf dieses Ergebnismuster nur für das erste und das vierte Experiment in Tabelle 9.5 im Sinne des Kontinuum-Modells interpretiert werden, da das Modell für die anderen beiden Experimente aufgrund des vorgeordneten Tests bereits verworfen werden mußte.

Zusammenfassend kann festgehalten werden, daß weder das einfache noch das erweiterte Kontinuum-Modell eindeutige Bestätigung erfuhren. Das erweiterte Modell beschreibt zwar das Experiment von Strack und Förster (1995) und mein eigenes Experiment aus Abschnitt 9.1.1.2 gut, nicht jedoch die Experimente von Hirshman und Henzler (1998) sowie Hicks und Marsh (1999). Die naheliegendste Erklärung dieses scheinbaren Widerspruchs scheint das relative Gewicht bewußter Erinnerungsprozesse in einem Experiment zu sein: Je größer die Bedeutung bewußter Erinnerungsprozesse ist, desto schlechter paßt das Kontinuum-Modell. Umgekehrt bietet es eine umso bessere Beschreibung der Daten, je eher bewußte Erinnerungen vernachlässigt werden können.

9.1.2 Das diskrete Zustandsmodell

Die im letzten Abschnitt aufgezeigten Grenzen des Kontinuum-Modells legen den Versuch nahe, ein alternatives Meßmodell zur einfachen RK-Prozedur zu entwickeln, das im Gegensatz zum Kontinuum-Modell auf einer Zwei-Prozeß-Annahme basiert. Ein solches diskretes Zustandsmodell wurde von Erdfelder, Brandt und Bröder (1995) vorgestellt. Konzeptuell ist es sehr eng an die ursprüngliche Idee von Tulving (1985) angelehnt, wonach „Remember"- und „Know"-Urteile für unterschiedliche Gedächtnissysteme sensitiv sind. Der Unterschied zur traditionellen Auswertung der RK-Prozedur besteht lediglich darin, daß nicht länger von einer prozeßreinen Erfassung der zugrundeliegenden Gedächtnissysteme ausgegangen wird.

Bei der Konstruktion des diskreten Zustandsmodells wurde berücksichtigt, daß Zwei-Hochschwellen-Modelle der Rekognition bei Modellevaluationen in der Regel deutlich besser abgeschnitten haben als Ein-Hochschwellen-Modelle (vgl. Abschnitt 3.5.2). Folglich ist die Forderung sinnvoll, daß sich das zu konstruierende Modell bei Zusammenfassung von R- und K-Urteilen zu einer „Alt"-Kategorie auf das Zwei-Hochschwellen-Modell der Rekognition reduzieren soll. Eine analoge Forderung für Quellendiskriminationsmodelle führt bekanntlich zum 2HTSM-Modell, welches sich in Validierungsstudien anderen Modellen gegenüber als überlegen erwiesen hat (Bayen et al., 1996; vgl. auch Abschnitt 8.2). Auch für Prozeßdissoziationsmodelle haben sich Zwei-Hochschwellen-Konzep-

tionen im Vergleich zu Ein-Hochschwellen-Konzeptionen insgesamt als überlegen erwiesen (Erdfelder & Buchner, 1995; Vaterrodt-Plünnecke et al., 1996).

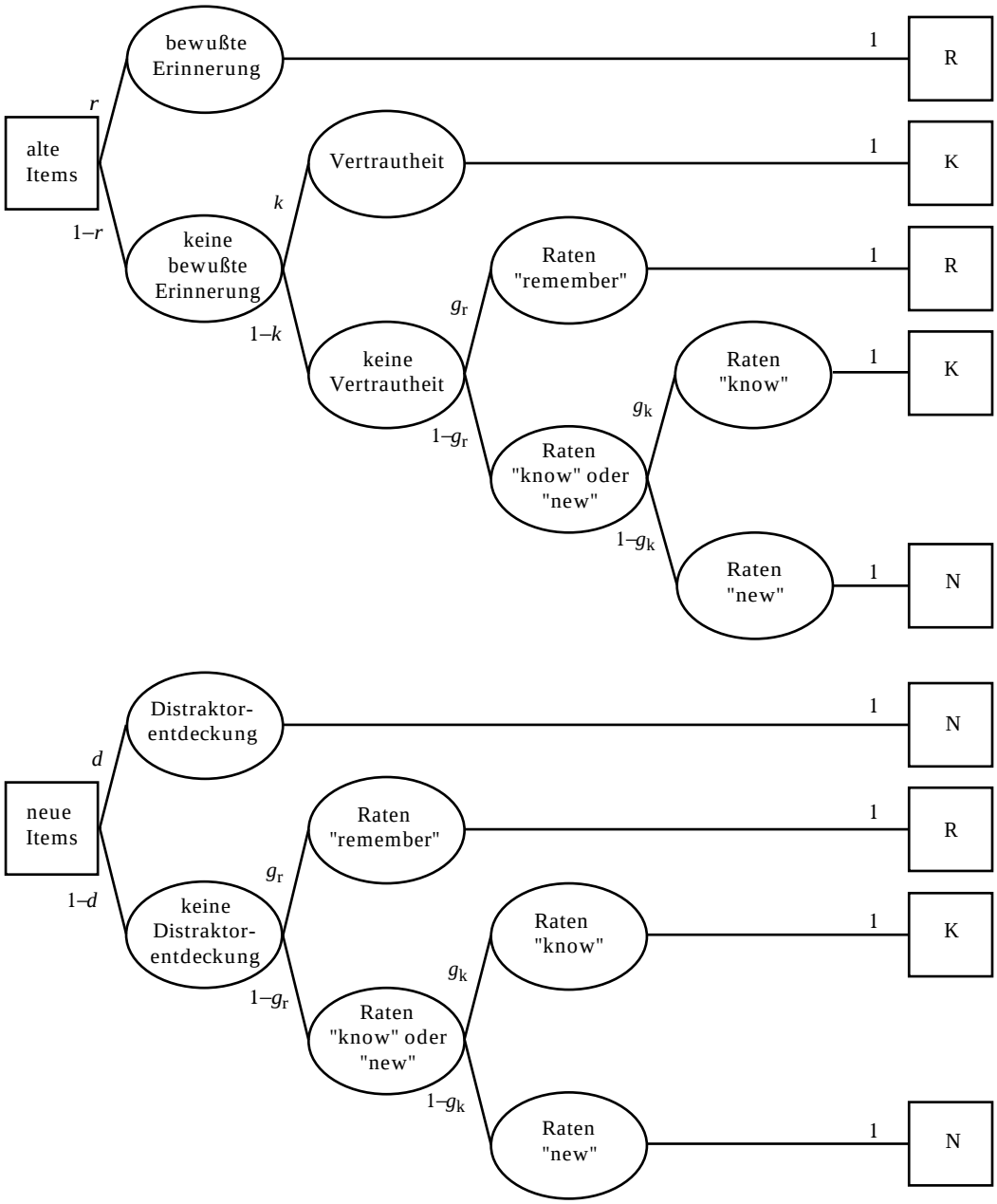

Abbildung 9.1:
Das diskrete Zustandsmodell zur einfachen Remember-Know-Prozedur. Erläuterungen im Text.

Abbildung 9.1 illustriert das diskrete Zustandsmodell zur einfachen RK-Prozedur in Form eines binären Verarbeitungsbaumdiagramms. Der Grundgedanke ist, daß immer zunächst das episodische Gedächtnis durchsucht wird, was bei alten Items mit Wahrscheinlichkeit r zu einer bewußten Erinnerung führt. Dies manifestiert sich — im Einklang mit Tulvings (1985) Annahmen — in einer R-Antwort. Gelingt die bewußte Erinnerung eines alten Items nicht (mit Wahrscheinlichkeit $1-r$), dann kann das Item mit der bedingten Wahrscheinlichkeit k dennoch ein

Vertrautheits- oder Bekanntheitsgefühl auslösen, das zu einer K-Antwort führt. Scheitern beide Prozesse, dann kommt es nicht — wie Tulving, Gardiner und Mitarbeiter implizit voraussetzen — zwangsläufig zu einer N-Antwort, sondern es wird in Abhängigkeit von Antworttendenzen geraten, wobei g_r und g_k die (bedingten) Ratewahrscheinlichkeiten für die R- und die K-Kategorie sind. Diese Rateprozesse sind letztlich verantwortlich dafür, daß R- und K-Antworten nicht als prozeßreine Maße der beiden Gedächtnissysteme aufgefaßt werden können.

Für neue Items laufen die gleichen Prozesse wie für alte ab, die jedoch per definitionem nicht zu einer bewußten Erinnerung oder zu einem speziellen Vertrautheitsgefühl führen können, das auf eine in der unmittelbaren Vergangenheit liegende Lernerfahrung hindeutet. Für neue Items gilt daher $r = k = 0$, so daß die betreffenden Äste im Diagramm hier nicht auftauchen. Umgekehrt kann jedoch das Fehlen sowohl einer bewußten Erinnerung als auch eines Vertrautheitsgefühls dazu führen, daß die Vp korrekt schlußfolgert, das Item müsse neu sein (Strack & Bless, 1994). Dies geschieht mit Wahrscheinlichkeit d und führt zu einer Distraktorentdeckung. Gelingt eine Distraktorentdeckung mit Wahrscheinlichkeit $(1–d)$ dagegen nicht, so laufen die gleichen Rateprozesse wie im Falle alter Items ab.

Wie man sich leicht klarmachen kann, hat das diskrete Zustandsmodell die gewünschte Eigenschaft, daß es sich bei Zusammenfassung der Antwortkategorien R und K zu einer singulären „Alt"-Kategorie auf das 2HT-Modell der Rekognition reduziert. Unter der Voraussetzung, daß $d = r + (1–r)\cdot k$ gilt, ergeben sich die Parameter des 2HT-Modells gemäß

$$\begin{aligned} r_{2ht} &= r + (1-r)\cdot k \\ b_{2ht} &= g_r + (1-g_r)\cdot g_k \end{aligned} \qquad (9.1.2\text{-}1)$$

aus den Parametern des diskreten Zustandsmodells. Man sieht unmittelbar, daß das diskrete Zustandsmodell sowohl den Rekognitions- als auch den Rateparameter des 2HT-Modells in jeweils zwei Parameter zerlegt.

Das in Abbildung 9.1 dargestellte diskrete Zustandsmodell wurde anhand der gleichen vier Datensätze evaluiert, die im letzten Abschnitt der Evaluation des Kontinuum-Modells dienten. Da das Zustandsmodell ein binäres MVB-Modell ist, wurde erneut das Programm AppleTree von Rothkegel (1998; 1999) für die Datenanalysen herangezogen. Die Modelleingabe- und die Datendatei (EQN- bzw. MDT-Datei) sind in Anhang 12 wiedergegeben.

Die Ergebnisse der ML-Analysen sind in Tabelle 9.6 zusammengefaßt. Da das unrestringierte diskrete Zustandsmodell für zwei Antworttendenzbedingungen zehn Parameter umfaßt und bei nur acht freien Kategoriewahrscheinlichkeiten nicht identifizierbar ist, werden in Tabelle 9.6 die ML-Parameterschätzungen zu einer restringierten Modellvariante angegeben, die gleiche Distraktorentdeckungsraten, gleiche bewußte Erinnerungsraten und gleiche Vertrautheitswahrschein-

lichkeiten über zwei Antworttendenzbedingungen hinweg annimmt. Die numerisch identischen *d*-, *r*- und *k*-Schätzwerte in aufeinanderfolgenden Zeilen der Tabelle 9.6 sind eine triviale Folge dieser Restriktionen und sollten daher nicht inhaltlich interpretiert werden.

Tabelle 9.6:
ML-Parameterschätzungen (Spalten 3 - 7) und G^2-Modellanpassungsstatistiken für das restringierte diskrete Zustandsmodell (Spalte 8) bei Anpassung des Modells an Datensätze aus vier verschiedenen Untersuchungen. Die Parameter *d*, *r* und *k* unterliegen Gleichheitsrestriktionen zwischen den Antworttendenzbedingungen. In Spalte 9 ist die Anpassungsstatistik zu dem Submodell aufgeführt, das gleiche *d*-, *r*- und *k*-Parameter für alle vier Bedingungen eines Experiments annimmt. Statistiken mit einem * sind auf dem Niveau $\alpha = .05$ signifikant; mit ** markierte Statistiken sind auf dem Niveau $\alpha = .001$ signifikant. Schätzungen des Distraktorentdeckungsparameters, die auf dem Niveau a = .05 von der Nullhypothese $d = 0$ abweichen, sind mit + markiert.

Experiment	ML-Parameterschätzungen					Restr. 1	Restr. 2
	d	*r*	*k*	g_r	g_k	$G^2_{(1)}$	$G^2_{(5)}$
Strack & Förster (1995)							
häufige W. 50% alt	0.23	0.23	0.23	0.03	0.32		
häufige W. 30% alt	0.23	0.23	0.23	0.02	0.10	1.92	
seltene W. 50% alt	0.39	0.40	0.29	0.05	0.38		
seltene W. 30% alt	0.39	0.40	0.29	0.05	0.21	4.00*	36.98**
Hirshman & Henzler (1998)							
500 msec 70% alt	0.17+	0.08	0.12	0.14	0.48		
500 msec 30% alt	0.17+	0.08	0.12	0.06	0.17	3.70	
2 s 70% alt	0.13	0.26	0.33	0.13	0.46		
2 s 30% alt	0.13	0.26	0.33	0.06	0.16	3.11	115.00**
Hicks & Marsh (1999)							
1 s 3AFC Test	0.56+	0.36	0.00	0.24	0.54		
1 s sukz. Test	0.56+	0.36	0.00	0.11	0.48	33.72**	
4,5 s 3AFC Test	0.55+	0.49	0.44	0.14	0.35		
4,5 s sukz. Test	0.55+	0.49	0.44	0.04	0.23	2.97	166.22**
eigene Daten aus Tabelle 9.4							
flach pro alt	0.00	0.05	0.04	0.18	0.27		
flach pro neu	0.00	0.05	0.04	0.13	0.20	0.05	
tief pro alt	0.00	0.31	0.07	0.18	0.27		
tief pro neu	0.00	0.31	0.07	0.13	0.20	0.03	137.29**

Spalte 8 zeigt, daß das restringierte Modell nur in einem Fall auf dem Niveau $\alpha = .001$ zurückgewiesen werden muß, nämlich für die kurze Darbietungsbedingung bei Hicks und Marsh (1999). Dies steht im Zusammenhang mit der Tatsache, daß der Vertrautheitsparameter k — wie alle Modellparameter — auf das Intervall [0, 1] beschränkt ist, welches im gegebenen Fall offenbar für eine gute Modellanpassung nicht ausreicht: Aufgrund der Ratewahrscheinlichkeiten sind signifikant mehr K-Antworten auf Zielitems zu erwarten, als empirisch beobachtet wurden. Der Vertrautheitsparameter müßte folglich auf einen negativen Wert geschätzt werden, um diesem Ergebnismuster gerecht zu werden. Da der Schätzwert auf dem Rand des Parameterraumes „festgehalten" wird, resultiert eine schlechte Anpassungsstatistik.

Abgesehen von der 1s-Versuchsbedingung von Marsh und Hicks (1999) ist auch die Anpassungsgüte für die seltenen Wörter bei Strack und Förster (1995) nur mäßig, was aber bei den gegebenen Stichprobenumfängen noch akzeptabel ist. Alle anderen Datensätze werden gut angepaßt, so daß das diskrete Zustandsmodell insgesamt etwas besser abschneidet als das Kontinuum-Modell.

Die Distraktorentdeckungsparameter werden für die meisten Datensätze auf einen Wert größer als null geschätzt, sind jedoch nur für drei Datensätze auf dem Niveau $\alpha = .05$ signifikant größer als null. Auf dem Niveau $\alpha = .01$, das man wegen der großen Stichprobenumfänge eigentlich vorziehen sollte, ist keiner der d-Schätzwerte von null verschieden. Konkret bedeutet das, daß eine Ein-Hochschwellen-Variante des diskreten Zustandsmodells, die aus dem Modell in Abbildung 9.1 durch die Restriktion $d = 0$ hervorgeht, die vorliegenden Daten ebenfalls recht gut anpassen kann. Dennoch will ich mich auf die Darstellung der Ergebnisse zur Zwei-Hochschwellen-Version beschränken, da die Ein-Hochschwellen-Version nur unter speziellen Randbedingungen gut abschneiden dürfte. Erfüllt sind diese Randbedingungen beispielsweise beim Datensatz aus Abschnitt 9.1.1.2: Aufgrund des relativ langen Retentionsintervalls und nichtintentionaler Lernbedingungen kann das Fehlen einer bewußten Erinnerung oder eines Vertrautheitsgefühls nicht unbedingt Anlaß für den Schluß sein, daß das betreffende Item neu sein müsse (vgl. Strack & Bless, 1994). Die Konsequenz ist, daß die Distraktorentdeckungsraten auf null sinken. Man hält es offenbar bei jedem neuen Item für denkbar, daß es alt sein könnte und lediglich aufgrund mangelnder Aufmerksamkeit in der Akquisitionsphase oder aufgrund des langen Retentionsintervalls vergessen wurde.

Die Parameterschätzungen zeigen ansonsten das erwartete Muster: Die Antworttendenzmanipulationen affizieren die Ratewahrscheinlichkeiten in der erwarteten Richtung, die Worthäufigkeits- und Enkodierungsmanipulationen beeinflussen die Gedächtnisparameter. In der letzten Spalte der Tabelle 9.6 sind die G^2-Anpassungsstatistiken einer noch stärker restringierten Modellvariante aufge-

führt, die gleiche *d*-, *r*- und *k*-Parameter über alle vier Bedingungen eines Experiments annimmt. Wie unschwer zu erkennen ist, muß diese Restriktion für jedes Experiment klar zurückgewiesen werden. Prüft man den Effekt der Gleichheitsrestriktion für jeden der drei Parameter separat, so wird deutlich, daß gleiche *d*-Parameter für jedes der vier Experimente angenommen werden können (alle $\Delta G^2_{(1)} < 1$), während die Gleichsetzungsrestriktionen für die *r*-Parameter immer zu einer auf dem Niveau $\alpha = 0.001$ signifikanten Verschlechterung der Anpassungsgüte führen (alle $\Delta G^2_{(1)} > 10.82$)[97]. Erwartungsgemäß ergibt sich ein differenzierteres Muster für Gleichsetzungsrestriktionen zum *k*-Parameter: Für die Daten von Strack und Förster (1999) sowie den in Abschnitt 9.1.1.3 berichteten Datensatz sind sie haltbar (beide $\Delta G^2_{(1)} < 1$), für die anderen beiden Experimente dagegen nicht (beide $\Delta G^2_{(1)} > 10.82$). Die Darbietungsdauer eines Items in der Enkodierungsphase affiziert also spätere bewußte Erinnerungen und zusätzlich die Vertrautheit, während Effekte der Worthäufigkeit und der Verarbeitungstiefe auf bewußte Erinnerungen beschränkt zu sein scheinen. Dies harmoniert gut mit den in der Literatur zum impliziten Gedächtnis berichteten Ergebnismustern.

Zusammenfassend kann festgehalten werden, daß das diskrete Zustandsmodell — mit einer Ausnahme, die noch näher aufgeklärt werden muß — recht gut auf die Daten von vier Experimenten zur einfachen RK-Prozedur paßt. Die Ergebnisse der Parameterschätzungen sind auf dem Hintergrund der Literatur zu Zwei-Prozeß-Theorien des Gedächtnisses plausibel. Da das Kontinuum-Modell größere Schwierigkeiten mit denselben Daten hat, sprechen die Befunde eher für das diskrete Zustandsmodell.

9.2 Modelle zu erweiterten Remember-Know-Prozeduren

Die vier in Abschnitt 9.1 referierten Experimente haben den Nachweis erbracht, daß alle drei Antwortkategorien der einfachen Remember-Know-Prozedur potentiell mit Ratetendenzen kontaminiert sind. Anders sind die Effekte verschiedener Antworttendenzmanipulationen auf die Antworthäufigkeiten nicht zu erklären. Eine Möglichkeit, hierauf zu reagieren, bilden Meßmodelle, welche die beobachteten Antworthäufigkeiten in Beiträge verschiedener Gedächtnis- und Urteilsprozesse dekomponieren. Diesem Ansatz sind das Kontinuum-Modell und das diskrete Zustandsmodell zuzurechnen, die in den letzten Abschnitten dargestellt wurden. Eine zweite Möglichkeit ist die Änderung der Prozedur mit dem Ziel, neue Beobachtungskategorien zu finden, die zumindest teilweise nicht durch Ratetendenzen affiziert werden. Gardiner et al. (1997) haben diesen Weg beschritten

[97] Man beachte, daß die ΔG^2-Statistik zum Datensatz von Hicks und Marsh (1999) nicht interpretierbar ist, da das Ausgangsmodell bereits zurückgewiesen werden mußte (vgl. Abschnitt 4.3.2 und 4.3.2.2).

und die erweiterte RKG-Prozedur mit den vier Antwortkategorien „remember“ (R), „know” (K), „guess” (G) und „new” (N) als Weiterentwicklung der einfachen RK-Prozedur vorgeschlagen. Die G-Antwortkategorie sollte es den Vpn ermöglichen, jene Gedächtnisurteile zu markieren, die subjektiv aus dem Gefühl des Ratens heraus erwachsen, d.h. mit großer Unsicherheit einhergehen.

Diese erweiterte RKG-Prozedur wurde inzwischen in einer Reihe von Experimenten eingesetzt (z.B. Gardiner et al., 1996; Gardiner & Radomski, 1999; Gardiner, Ramponi & Richardson-Klavehn, 1998). Gardiner et al. (1997) replizierten das Experiment von Strack und Förster (1995) mit der erweiterten Prozedur und fanden zum Teil Effekte der Antworttendenzmanipulation auf G-Antworten, nicht aber auf R- und K-Antworten, die von primärem Interesse sind. Sie argumentierten, daß die R- und die K-Antworthäufigkeiten der erweiterten RKG-Prozedur als von Antworttendenzen unkontaminiert aufgefaßt werden können. In eine ähnliche Richtung weisen neuere Befunde von Gardiner et al. (1998), die ausführliche verbale Berichte über Erinnerungserfahrungen bei Benutzung der erweiterten RKG-Prozedur analysiert haben. R-Urteile gehen demnach fast immer mit bewußten Erinnerungen an anstrengende Enkodierungstätigkeiten oder an spontane Gedankenassoziationen während der Enkodierungsphase einher, speziell Assoziationen in bezug auf die eigene Person. K-Urteile sind demgegenüber mit Vertrautheitsgefühlen oder Gefühlen des „just knowing” assoziiert. Während R- und K-Urteile mit Erinnerungsgewißheit verknüpft sind, gilt dies für G-Urteile nicht. Hier ist sich die Vp typischerweise sehr unsicher. G-Urteile treten bei alten und neuen Items mit ungefähr gleicher Häufigkeit auf, was Gardiner und Mitarbeiter (1998) als weiteren Hinweis darauf werteten, daß wirklich nur Urteilsprozesse, jedoch keine Gedächtnisprozesse in G-Antworten einmünden.

Somit scheint die erweiterte RKG-Prozedur eine hinreichend feine Aufgliederung möglicher Erinnerungserfahrungen zu liefern, vielleicht mit Ausnahme der K-Kategorie, die insofern auch in der erweiterten Prozedur noch hybrid ist, als sie Gefühle der Vertrautheit und des „just knowing” nicht auseinanderhält. Tatsächlich ist auch eine Aufsplittung der K-Kategorie in Vertrautheit und „just know” schon vorgeschlagen und auch empirisch exploriert worden (vgl. Conway, Gardiner, Perfect, Anderson & Cohen, 1997). Bislang liegen hierzu jedoch nur wenige Daten vor, so daß diese Erweiterung in der vorliegenden Arbeit nicht explizit diskutiert werden soll. Die Überlegungen, die im folgenden dargestellt werden, lassen sich jedoch prinzipiell auch auf eine zweifach erweiterte RK-Prozedur mit den fünf Antwortkategorien „remember“ (R), „familiarity” (F), „just know” (K), „guess” (G) und „new” (N) generalisieren. Ich werde am Schluß der Zusammenfassung dieses Kapitels (Abschnitt 9.3) hierauf kurz zurückkommen.

Zunächst ist kritisch zu prüfen, ob die Behauptung von Gardiner und Mitarbeitern (1997) wirklich zutrifft, daß Probleme der Kontaminierung von Antwortka-

tegorien mit Antworttendenzen durch Erweiterung der einfachen RK-Prozedur um eine G-Antwortkategorie wirksam vermieden werden können. Betrachten wir zunächst das bereits erwähnte Experiment von Gardiner et al. (1997) etwas genauer. Die relativen Antworthäufigkeiten für die verschiedenen Bedingungen des Experiments sind in Tabelle 1 bei Gardiner et al. (1997, S. 392) publiziert. Man kann die absoluten Antworthäufigkeiten daraus schätzen, indem man jede relative Häufigkeit mit dem Produkt aus zugrundeliegenden Items (24 pro Bedingung) und Vpn (ebenfalls 24 pro Bedingung) multipliziert, d.h. mit 576. Die geschätzten absoluten und die relativen Antworthäufigkeiten werden hier in Tabelle 9.7 wiedergegeben.

Tabelle 9.7:
Geschätzte absolute (ohne Klammern) und relative Häufigkeiten von R-, K-, G- und N-Antworten (in Klammern) auf häufige und seltene Zielitems und Distraktoritems bei einer mitgeteilten Basisrate von 50% versus 30% alten Items im RKG-Test. Der tatsächliche Anteil alter Testitems betrug immer 50%. Die relativen Häufigkeiten wurden aus Tabelle 1 bei Gardiner et al. (1997, S. 392) übernommen.

	Häufige Wörter				Seltene Wörter			
	Zielitems		Distraktoren		Zielitems		Distraktoren	
Antwort	50%	30%	50%	30%	50%	30%	50%	30%
Remember (R)	173 (.30)	150 (.26)	17 (.03)	12 (.02)	236 (.41)	213 (.37)	12 (.02)	17 (.03)
Know (K)	127 (.22)	121 (.21)	52 (.09)	63 (.11)	132 (.23)	127 (.22)	52 (.09)	69 (.12)
Guess (G)	69 (.12)	69 (.12)	104 (.18)	58 (.10)	52 (.09)	52 (.09)	.63 (.11)	63 (.11)
New (N)	207 (.36)	236 (.41)	403 (.70)	443 (.77)	156 (.27)	184 (.32)	449 (.78)	427 (.74)

Auf den ersten Blick erkennt man das bereits skizzierte Ergebnismuster: Seltene Zielitems haben höhere R-Antworthäufigkeiten als häufige Zielitems, während die K-Antworthäufigkeiten sich nur marginal unterscheiden. Varianzanalytische *F*-Tests bestätigen diesen Eindruck (Gardiner et al., 1997, S. 392). Weiterhin sind G-Antworten bei häufigen Zielitems häufiger als bei seltenen Zielitems. Auf den zweiten Blick wird deutlich, daß die Antworttendenzmanipulation nur auf G-Antworten bei häufigen Distraktoren statistisch signifikant durchschlägt und sonst praktisch keine Effekte oder sogar Effekte mit falschem Vorzeichen zeigt. Dies ist erstaunlich und läßt vermuten, daß die bloße Mitteilung, 50% versus 30% der Testitems seien alte Items, schlicht nicht wirksam genug war, um substantielle Antworttendenzänderungen zu erzeugen (vgl. auch Hirshman & Henzler, 1998). Wenn dem so ist, dann ist es auch nicht erstaunlich, daß die Manipulation keine signifikanten Effekte auf R- und K-Antworten ausübte.

Es erscheint folglich ratsam, den Effekt einer stärkeren Antworttendenzmanipulation auf die Antwortkategorien der RKG-Prozedur zu explorieren. Geeignet hierfür ist die bereits in Abschnitt 9.1.1.2 geschilderte eigene Untersuchung, die Effekte von Antworttendenzveränderungen als Folge der Manipulation von Aus-

zahlungsmatrizen zum Gegenstand hatte. In zwei randomisierten experimentellen Bedingungen wurde hierbei für einen Teil der Vpn-Gesamtstichprobe die einfache RK-Prozedur verwendet, deren Ergebnisse in Tabelle 9.4 zusammengefaßt sind. Für die anderen 62 Vpn wurde die erweiterte RKG-Prozedur verwendet. Per Zufallsauswahl wurden hiervon 36 Vpn der liberalen Antworttendenzbedingung („pro alt") und 26 Vpn der konservativen Bedingung („pro neu") zugewiesen.

Tabelle 9.8 stellt die absoluten und die relativen Antworthäufigkeiten für die erweiterte RKG-Prozedur zusammen. Die maximale Summe der Antworthäufigkeiten beträgt für flach und tief verarbeitete alte Items in Tabelle 9.8 jeweils 25 Items · 36 Vpn = 900 (Versuchsbedingung „pro alt") bzw. 25 Items · 26 Vpn = 650 (Versuchsbedingung „pro neu"). Für Distraktoritems ist die maximale Summe erneut doppelt so groß. Da einige Vpn bei einigen Items das RKG-Urteil vergaßen, wurden die Maximalsummen nicht ganz erreicht.

Tabelle 9.8:
Relative Häufigkeiten (in Klammern) und absolute Häufigkeiten von R-, K-, G- und N-Antworten (ohne Klammern) auf zuvor tief bzw. flach verarbeitet Zielitems und Distraktoritems bei Induktion von Antworttendenzen zugunsten einer „Alt"- bzw. einer „Neu"-Antwort mittels Manipulation von Auszahlungsmatrizen. Die Schilderung der zugrundeliegenden Untersuchung erfolgte in Abschnitt 9.1.1.2.

	Alte Items				Distraktoren	
	flache Verarbeitung		tiefe Verarbeitung			
Antwort	pro alt	pro neu	pro alt	pro neu	pro alt	pro neu
Remember (R)	184 (.21)	95 (.15)	409 (.46)	233 (.36)	348 (.19)	160 (.12)
Know (K)	120 (.13)	58 (.09)	83 (.09)	72 (.11)	198 (.11)	95 (.07)
Guess (G)	142 (.16)	54 (.08)	99 (.11)	46 (.07)	265 (.15)	80 (.06)
New (N)	450 (.50)	437 (.68)	304 (.34)	291 (.45)	977 (.55)	956 (.74)

Die Enkodierungsmanipulation hat offensichtlich einen starken Effekt auf die R-Antworthäufigkeiten in der erwarteten Richtung, K- und G-Antworten sind dagegen weniger betroffen. Anders als in der Untersuchung von Gardiner et al. (1997) zeigt die Antworttendenzmanipulation einen durchgängigen Effekt auf G-Antworten, erstaunlicherweise aber auch auf die R-Antworten. Wertet man die Antworthäufigkeiten über eine zweifaktorielle ANOVA mit dem meßwiederholten Faktor „Tiefe der Verarbeitung" und dem nichtmeßwiederholten Faktor „Antworttendenz" aus, so werden auf dem Niveau $\alpha = .05$ sowohl der Effekt der Verarbeitungstiefe ($F(1, 60) = 143.31$) als auch der Antworttendenzeffekt bezüglich der R-Antworthäufigkeiten signifikant ($F(1, 60) = 4.38$). Für die K-Antworthäufigkeiten wird nur die Interaktion zwischen beiden UVn signifikant ($F(1, 60) = 4.77$). G-Antworten zeigen einen Haupteffekt der Verarbeitungstiefe ($F(1, 60) = 9.33$), während der Antworttendenzeffekt und der Interaktionseffekt das 5%-Niveau

knapp verfehlen ($F(1, 60) = 3.06$ bzw. 3.25). Vergleicht man die Antworthäufigkeiten auf Distraktoritems zwischen den Antworttendenzbedingungen mittels einseitiger t-Tests, so wird der Effekt für R-Antworten ($t(60) = 5.03$), für G-Antworten ($t(60) = 6.29$) und für N-Antworten ($t(60) = 15.75$) in der prognostizierten Richtung auf dem 5%-Niveau signifikant, nicht jedoch für K-Antworten.

Mit diesen Analysen ist der Nachweis erbracht, daß die bloße Erweiterung der RK-Prozedur auf eine RKG-Prozedur das Problem der möglichen Kontaminierung von Antwortkategorien mit Antworttendenzen nicht löst. Werden Antworttendenzen stark genug manipuliert, so schlagen sie offenbar sogar auf R-Antworten durch, eventuell sogar noch stärker als auf G- und K-Antworten. Dies entspricht den Ergebnissen, die auch für die einfache RK-Prozedur ermittelt wurden (vgl. Abschnitt 9.1.1.2).

Der denkbare Einwand, daß die Manipulation der Auszahlungsmatrizen Vpn indirekt nahegelegt hat, Antworten zu geben, die nicht ihren wirklichen Erinnerungserfahrungen entsprechen (vgl. Gardiner et al., 1998, S. 286), greift bei Verwendung der RKG-Prozedur nicht: Durch Benutzung der G-Antwortkategorie hätten die Vpn es ja in der Hand gehabt, „Alt"-Antworten zu geben, die weder mit bewußter Erinnerung noch mit Vertrautheit oder „just knowing" assoziiert sind. Niemandem wurde durch die Auszahlungsmatrizen nahegelegt oder suggeriert, speziell die R- oder K-Antworten zu suchen oder zu meiden. Die Vpn hatten durchaus die Möglichkeit, ihre Punktzahl durch geschickte Verwendung der G- bzw. N-Antworten zu maximieren und R- sowie K-Antworten nur für die Fälle zu reservieren, bei denen das subjektive Erinnerungsgefühl eine entsprechende Antwort wirklich gerechtfertigt hat. Dennoch zeigte die Auszahlungsmanipulation Effekte auch auf die R-Antworten. Folglich muß man diese Kategorie — und sehr wahrscheinlich auch die K-Kategorie der RKG-Prozedur — als nach wie vor mit Antworttendenzen kontaminiert auffassen. Eine Lösung des Problems scheint nur mit Hilfe von Meßmodellen möglich zu sein.

Zwei Kandidaten für Meßmodelle zur erweiterten RKG-Prozedur werde ich in den folgenden beiden Abschnitten vorstellen und evaluieren. Das erste Modell ist eine Verallgemeinerung des Kontinuum-Modells aus Abschnitt 9.1.1 auf vier statt drei Antwortkategorien. Es ist Gegenstand von Abschnitt 9.2.1. Abschnitt 9.2.2 behandelt eine analoge Verallgemeinerung des diskreten Zustandsmodells aus Abschnitt 9.1.2 auf die erweiterte RKG-Prozedur.

9.2.1 Das RKG-Kontinuum-Modell

Die erweiterte RKG-Prozedur ist von Hirshman (1998b) mit Argumenten kritisiert worden, die seinen Argumenten gegen die einfache RK-Prozedur weitgehend entsprechen (Hirshman & Master, 1997; Hirshman & Henzler, 1998). Er zeigte, daß die

bisherigen Ergebnisse zur RKG-Prozedur qualitativ mit einem Kontinuum-Modell vereinbar sind, das nicht nur zwei Antwortkriterien für die K- und die R-Kategorie (c_r und c_k), sondern ein zusätzliches drittes Antwortkriterium c_g für die G-Antwortkategorie umfaßt, wobei $c_g < c_k < c_r$. Dieses RKG-Kontinuum-Modell entspricht einem Signalentdeckungsmodell für eine vierstufige Ratingskala, wobei N, G, K und R die Stufen der Skala in aufsteigender Folge darstellen (vgl. Macmillan & Creelman, 1991, Kap. 3).

Eine Anpassung des RKG-Kontinuum-Modells an Daten der erweiterten Prozedur wurde bislang allerdings noch nicht vorgenommen. Hirshman (1998b) zeigte lediglich, daß die Richtungen der in der Literatur berichteten Effekte auf R-K- und G-Antworten durch geeignete Wahl des Sensitivitätsparameters *d´ und* der drei Antwortkriterien c_g, c_k und c_r reproduziert werden können. „The exercise is entirely post hoc", wie Gardiner, Richardson-Klavehn und Ramponi (1998, S. 285) in ihrer Antwort auf Hirshman (1998b) zu Recht feststellten. Von einem strengen Test des RKG-Kontinuum-Modells kann bislang noch nicht gesprochen werden.

Um diese Lücke zu schließen, werde ich im folgenden das RKG-Kontinuum-Modell in der einfachen Form (mit gleichen Vertrautheitsstreuungen für Distraktoren und Zielitems) und in der erweiterten Hirshman-Master-Version (mit beliebigen Vertrautheitsstreuungen) Modellanpassungstests unterziehen. Wie in Abschnitt 9.1.1 festgestellt wurde, läßt sich das Kontinuum-Modell zur einfachen RK-Prozedur für einen Zielitem-Distraktoritem-Datensatz nur in der Variante mit homogenen Vertrautheitsstreuungen für Zielitems und Distraktoren testen, nicht dagegen in der erweiterten Version nach Hirshman und Master (1997), die ungleiche Streuungen erlaubt. Letzteres Modell ist für einen einzelnen RK-Datensatz saturiert und kann nur anhand von Experimenten geprüft werden, die beispielsweise Antworttendenzen manipulieren. Für die RKG-Kontinuum-Modelle, die Gegenstand des vorliegenden Abschnitts sind, sieht dies etwas anders aus. Da nunmehr jeweils vier Antwortkategorien für Zielitems und Distraktoren gegeben sind (R, K, G und N), gibt es $2 \cdot 3 = 6$ funktional unabhängige Kategoriewahrscheinlichkeiten. Das RKG-Kontinuum-Modell mit homogenen Streuungen umfaßt einen Sensitivitätsparameter *d´* und drei Antwortkriterien c_g, c_k und c_r, ist also bei $df = 6 - 4 = 2$ testbar. Das Modell mit freier Streuung σ für die Zielitems hat zwar einen Parameter mehr, ist aber immer noch bei $df = 1$ testbar. Prinzipiell eignet sich also jeder RKG-Datensatz zur Prüfung beider Varianten des RKG-Kontinuum-Modells.

Dennoch werde ich im folgenden keine umfassende Reanalyse publizierter Datensätze zur RKG-Prozedur präsentieren, sondern mich stattdessen auf die beiden einzigen mir bekannten Experimente konzentrieren, die (a) die RKG-Prozedur benutzt haben und (b) zusätzlich Antworttendenzen experimentell manipuliert haben. Es handelt sich um die in den Tabellen 9.7 (Gardiner et al., 1997) und 9.8

(eigene Untersuchung) zusammengefaßten Experimente. Wie deutlich werden wird, reichen diese beiden Datensätze bereits aus, um ein Urteil zum RKG-Kontinuum-Modell zu fällen.

Die Modellgleichungen der RKG-Kontinuum-Modelle werden hier nicht explizit aufgeführt. Die Modellgleichungen zum Modell mit homogenen Streuungen ergeben sich unmittelbar durch Erweiterung der Gleichungen (9.1.1-1) bis (9.1.1-4) um je eine Gleichung zum Antwortkriterium c_g für Zielitems und Distraktoren. Das erweiterte RKG-Modell nach Hirshman und Master (1997), welches ungleiche Streuungen der Vertrautheitsverteilungen von Distraktoritems und alten Items zuläßt, ergibt sich durch Einbeziehung des σ-Parameters in die Modellgleichungen — ganz analog zum Vorgehen beim erweiterten Kontinuum-Modell (vgl. dazu Abschnitt 9.1.1.3).

Die Anpassung der beiden Modellvarianten an die beiden o.g. Datensätze erfolgte wiederum mit der ML-Methode unter Benutzung des PRAXIS-Programms von Gegenfurtner (1992). Sowohl das einfache als auch das erweiterte RKG-Kontinuum-Modell wurde für jede Versuchsbedingung getrennt an die Häufigkeitsdaten aus den Tabellen 9.7 und 9.8 angepaßt. Die Ergebnisse sind in Tabelle 9.9 zusammengefaßt.

Tabelle 9.9:
G^2-Modellanpassungsstatistiken für das einfache RKG-Kontinuum-Modell (3. Spalte) sowie ML-Parameterschätzungen (Spalten 4 - 8) und G^2-Modellanpassungsstatistiken (Spalte 9) für das σ-erweiterte RKG-Kontinuum-Modell bei Anpassung der Modelle an die Daten zur RKG-Prozedur aus Tabelle 9.7 und 9.8. Statistiken mit einem * sind auf dem Niveau α = .05 signifikant; mit ** markierte Statistiken sind auf dem Niveau α = .001 signifikant.

Experiment	einfaches Modell	erweitertes RKG-Kontinuum-Modell					
		ML-Parameterschätzungen					
	$G^2_{(2)}$	d'	σ	c_g	c_k	c_r	$G^2_{(1)}$
Gardiner et al. (1997)							
häufige W. 50% alt	25.71**	0.21	1.18	–0.22	0.55	1.61	21.84**
häufige W. 30% alt	29.59**	0.22	1.30	–0.09	0.75	1.71	20.96**
seltene W. 50% alt	15.40**	0.45	1.31	–0.58	0.54	1.82	0.06
seltene W. 30% alt	19.23**	0.34	1.36	–0.42	0.53	1.64	4.05*
eigene Daten aus Tabelle 9.8							
flach pro alt	269.93**	–22.66	52.73	11.74	11.45	12.19	181.54**
flach pro neu	66.82**	–1.49	3.04	1.38	1.44	1.89	55.86**
tief pro alt	26.43**	0.20	1.20	–0.29	0.07	0.74	20.84**
tief pro neu	2.13	0.22	1.12	–0.04	0.53	1.06	0.41

Das einfache Modell mit homogenen Vertrautheitsstreuungen paßt offensichtlich nur auf eine von acht Versuchsbedingungen und muß für die anderen sieben Bedingungen klar zurückgewiesen werden. Daß dies nur zum geringen Teil an der Annahme homogener Streuungen liegt, wird in der letzten Spalte der Tabelle deutlich: Auch das erweiterte RKG-Modell mit einem zusätzlichen freien Streuungsparameter σ für Zielitems muß für mehr als die Hälfte aller Versuchsbedingungen verworfen werden. Hinzu kommt, daß die ML-Parameterschätzungen zumindest für die flach verarbeiteten Items aus Tabelle 9.8 mit negativen d'-Schätzungen und extremen Zielitemstreuungen sehr unplausibel sind.

Verallgemeinert man das RKG-Kontinuum-Modell auf zwei Antworttendenzbedingungen mit der zusätzlichen Restriktion, daß die Antworttendenzen zwar Effekte auf die drei Antwortkriterien, aber keine Effekte auf d' und σ haben dürfen — dies ist ein naheliegender, aber recht strenger Modelltest —, so ergibt sich sowohl für die häufigen Wörter von Gardiner et al. ($G^2_{(4)} = 43.89$) als auch für die flach ($G^2_{(4)} = 293.45$) und die tief verarbeiteten Wörter aus Tabelle 9.8 ($G^2_{(4)} = 21.54$) ganz klare Evidenz gegen das Modell (bei $\alpha = .001$). Lediglich für die seltenen Wörter des Gardiner-et-al.-Experiments übersteht das Modell den Modelltest ($G^2_{(4)} = 6.15$).

Offensichtlich ist die Annahme falsch, daß die Antwortkategorien N, G, K, und R auf einem Kontinuum normalverteilter Vertrautheitsstärken lokalisiert werden können, eine Annahme, die für die N-, die K- und die R-Kategorie der einfachen RK-Prozedur mit Abstrichen noch vertretbar schien (vgl. Abschnitt 9.1.1.3). Denkbar ist, daß die Anpassungsprobleme im Rahmen eines verteilungsfreien RKG-Kontinuum-Modells, das auf die Normalverteilungsannahme für Vertrautheitswerte verzichtet, etwas reduziert werden können. Erfolgversprechend erscheint dieser Weg jedoch nicht, da die hier berichteten G^2-Werte einfach zu groß sind. Plausibler erscheint es, einen alternativen Modellierungszugang zu suchen, der die Kontinuum-Idee zugunsten der Idee eines diskreten Zustandsraumes fallen läßt. Diesem Ansatz wende ich mich im nächsten Abschnitt zu.

9.2.2 Das diskrete RKG-Zustandsmodell

Eine Alternative zum RKG-Kontinuum-Modell ergibt sich zwanglos aus dem diskreten Zustandsmodell, das in Abschnitt 9.1.2 behandelt wurde. Die Grundidee ist, den Zustandsraum dieses Modells beizubehalten und lediglich das Meßmodell auf die zusätzliche G-Antwortkategorie zu erweitern. Bei alten Items ist man also entweder (a) im Zustand bewußter Erinnerung (mit Wahrscheinlichkeit r), (b) im Zustand der Vertrautheit (mit Wahrscheinlichkeit $(1–r)\cdot k$) oder (c) im Zustand der Unsicherheit (mit Wahrscheinlichkeit $(1–r)\cdot(1–k)$). In den ersten beiden Zuständen

werden R- bzw. K-Antworten gegeben, genau wie im diskreten Zustandsmodell. Lediglich im Unsicherheitszustand kommt die Antwortkategorie G als zusätzliche — und vielleicht sogar als wichtigste — Ratekategorie zu Geltung. Es wird jedoch keine Annahme darüber gemacht, wie stark die Neigung ist, G zu antworten, wenn man unsicher ist. Alle vier Antworten sind hier prinzipiell möglich, was konkret die Hinzunahme eines weiteren Rateparameters g_g impliziert.

Analoge Überlegungen gelten für die Distraktoritems. Auch hier kann die Antwortkategorie G als zusätzliche Option nur im Unsicherheitszustand gewählt werden. Alle anderen Annahmen entsprechen dem diskreten Zustandsmodell aus Abschnitt 9.1.2.

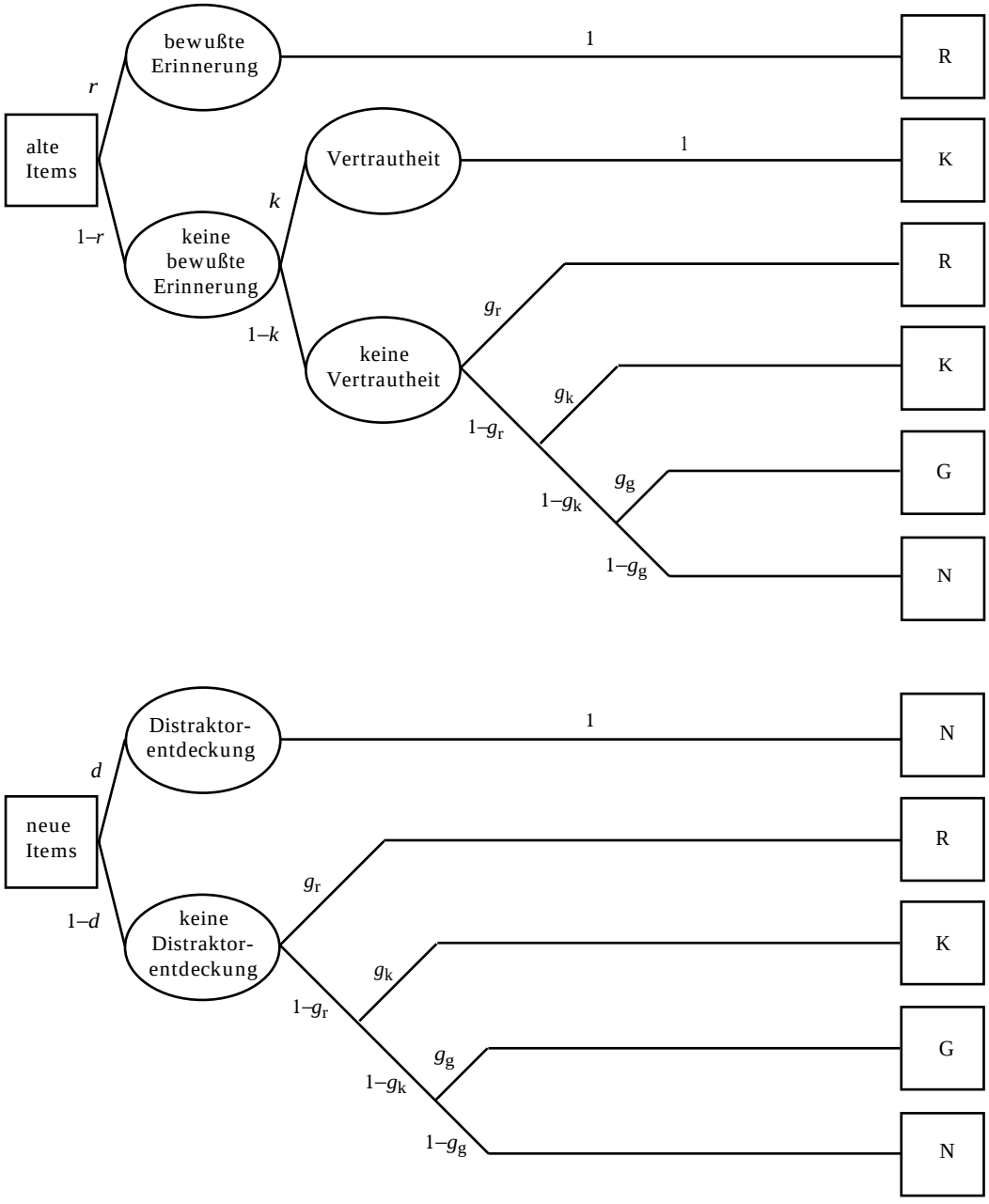

Abbildung 9.2:
Das diskrete RKG-Zustandsmodell zur RKG-Prozedur. Erläuterungen im Text.

Abbildung 9.2 illustriert das diskrete RKG-Zustandsmodell in Form eines binären Verarbeitungsbaumdiagramms. Auch dieses Modell reduziert sich unter der Voraussetzung, daß $d = r + (1–r)\cdot k$ gilt, auf das 2HT-Modell der Rekognition, wenn man die drei Antwortkategorien R, K und G zu einer „Alt"-Kategorie zusammenfaßt. Die Parameter des 2HT-Rekognitionsmodells ergeben sich dann wie folgt aus den Parametern des diskreten RKG-Zustandsmodells:

$$\begin{aligned} r_{2ht} &= r + (1-r)\cdot k \\ b_{2ht} &= g_r + (1-g_r)\cdot g_k + (1-g_r)\cdot(1-g_k)\cdot g_g \end{aligned} \qquad (9.2.2\text{-}1)$$

Offensichtlich wird also nur der Rateparameter b_{2ht} in drei statt bisher zwei neue Rateparameter aufgesplittet, sonst ändert sich nichts.

Das in Abbildung 9.2 dargestellte Modell wurde anhand der gleichen Datensätze evaluiert, die im letzten Abschnitt der Evaluation des RKG-Kontinuum-Modells dienten. Wiederum wurde das Programm „AppleTree" von Rothkegel (1998; 1999) für die Datenanalysen herangezogen. Die Modelleingabe- und die Datendatei (EQN- bzw. MDT-Datei) werden in Anhang 13 wiedergegeben.

Tabelle 9.10:
ML-Parameterschätzungen (Spalten 3 - 8) und G^2-Modellanpassungsstatistiken für das restringierte diskrete RKG-Zustandsmodell (Spalte 9) bei Anpassung des Modells an Datensätze aus zwei Untersuchungen. Die Parameter *d*, *r* und *k* unterliegen Gleichheitsrestriktionen zwischen den Antworttendenzbedingungen. Statistiken mit einem * sind auf dem Niveau $\alpha = .05$ signifikant. Schätzungen des Distraktorentdeckungsparameters, die auf dem Niveau $\alpha = .05$ von der Nullhypothese $d = 0$ abweichen, sind mit + markiert.

Experiment	ML-Parameterschätzungen						$G^2_{(3)}$
	d	*r*	*k*	g_r	g_k	g_g	
Gardiner et al. (1997)							
häufige W. 50% alt	0.28$^+$	0.26	0.18	0.04	0.14	0.28	
häufige W. 30% alt	0.28$^+$	0.26	0.18	0.03	0.15	0.20	8.17*
seltene W. 50% alt	0.40$^+$	0.37	0.22	0.04	0.17	0.24	
seltene W. 30% alt	0.40$^+$	0.37	0.22	0.04	0.19	0.23	6.53
eigene Daten aus Tabelle 9.8							
flach pro alt	0.10	<0.01	0.01	0.21	0.16	0.25	
flach pro neu	0.10	<0.01	0.01	0.14	0.09	0.10	2.53
tief pro alt	0.14$^+$	0.28	0.04	0.23	0.16	0.26	
tief pro neu	0.14$^+$	0.28	0.04	0.14	0.11	0.11	13.36*

Die Ergebnisse der ML-Analysen sind in Tabelle 9.10 zusammengefaßt. Da das unrestringierte diskrete Zustandsmodell für zwei Antworttendenzbedingun-

gen saturiert und nicht testbar ist, werden in Tabelle 9.10 wie schon in Tabelle 9.6 die ML-Parameterschätzungen und G^2-Anpassungsstatistiken zu einer restringierten Modellvariante angegeben, die gleiche *d*-Parameter, gleiche *r*-Parameter und gleiche *k*-Parameter über zwei Antworttendenzbedingungen hinweg annimmt.

Spalte 9 zeigt, daß das restringierte RKG-Modell in keinem Fall auf dem Niveau $\alpha = .001$ zurückgewiesen werden muß. Die Anpassungsergebnisse sind somit erheblich besser als im Falle des RKG-Kontinuum-Modells und denen zum diskreten Zustandsmodell vergleichbar. Voll befriedigen kann die Anpassungsgüte allerdings auch für das RKG-Zustandsmodell nicht, da die G^2-Statistik mit nur einer Ausnahme über ihrem Erwartungswert bei Modellgültigkeit liegt. Partiell scheint das Problem mit der Gleichheitsrestriktion über die beiden *d*-Parameter zusammenzuhängen. Gibt man diese Gleichheitsrestriktion auf, so sinken die vier G^2-Statistiken von 8.17 auf 6.90, von 6.53 auf 3.14, von 2.53 auf 0.89 und schließlich sogar von 13.36 auf 7.35. Denkbar ist also, daß die Vpn bei Benutzung der erweiterten RKG-Prozedur das Kriterium für eine Distraktorentdeckung in Abhängigkeit von den erwarteten Auszahlungen für „Alt"- und „Neu"-Antworten ändern: Bei einer Tendenz zu „Neu"-Antworten ist das Kriterium liberaler und führt zu mehr Distraktorentdeckungen. Dies ist jedoch nur eine Post-hoc-Erklärung des beobachteten Ergebnismusters. Eine endgültige Bewertung des diskreten RKG-Zustandsmodells sollte erst nach einer Replikation des hier beobachteten Ergebnisses anhand unabhängig gewonnener Daten erfolgen.

Die ML-Parameterschätzungen zum restringierten RKG-Zustandsmodell liegen hinsichtlich der strukturellen Parameter *d*, *r* und *k* und auch hinsichtlich des Rateparameters g_r erstaunlich nah an den korrespondierenden Ergebnissen zum diskreten Zustandsmodell aus Tabelle 9.6: Die Ergebnisse zum Gardiner-et-al.-(1997)-Datensatz entsprechen denen zum Experiment von Strack und Förster (1995), und die Ergebnisse zum eigenen RKG-Datensatz aus Tabelle 9.8 entsprechen denen zum RK-Datensatz aus Tabelle 9.4. Dies ist unter der Annahme zu erwarten, daß die beiden diskreten Zustandsmodelle für die einfache RK- und die erweiterte RKG-Prozedur die gleichen kognitiven Zustände erfassen. Da die Daten in Tabelle 9.4 und Tabelle 9.8 von randomisierten Vpn-Gruppen unter identischen experimentellen Bedingungen generiert wurden, war die Übereinstimmung der Zustandswahrscheinlichkeiten für beide Datensätze zu erwarten. Gleiches gilt nicht unbedingt für die Übereinstimmung zwischen Strack und Förster (1995) auf der einen und Gardiner et al. (1997) auf der anderen Seite, weil hier unterschiedliche Wortmaterialien und Probandengruppen Verwendung fanden. Überraschend ist die hohe Übereinstimmung jedoch auch in diesem Fall nicht, da das Experiment von Gardiner und Mitarbeitern (1997) auf eine möglichste exakte Replikation des Experiments von Strack und Förster (1995) abzielte.

Die Übereinstimmung der *d*-Parameter, der *r*-Parameter und der *k*-Parameter des einfachen diskreten Zustandsmodells und des erweiterten RKG-Zustandsmodells läßt sich im Rahmen eines gemeinsamen Modells für die RK-Daten und die RKG-Daten statistisch prüfen. Ich habe diesen Übereinstimmungstest nur für die eigenen Daten aus Tabelle 9.4 und Tabelle 9.8 durchgeführt, weil aus den o.g. Gründen strenggenommen nur für diese beiden Datensätze parallele Ergebnismuster zu erwarten waren. Die Modellbeschreibungs- und Häufigkeitsdateien (EQN- und MDT-Dateien) zur Durchführung dieses Übereinstimmungstests mit den Programmen AppleTree oder GPT — MBT scheidet hier wegen Überschreitung der Arbeitsspeichergrenze des DOS-Betriebssystems aus — sind in Abschnitt 14 des Anhangs wiedergegeben.

Die Differenzen-G^2-Statistik zum Test der Prozedurunabhängigkeit der insgesamt zehn strukturellen Parameter — *d*, zwei *r*- und zwei *k*-Parameter für die flache und die tiefe Verarbeitungsbedingung, alle fünf Parameter jeweils getrennt geschätzt für die beiden Auszahlungsbedingungen — beträgt $\Delta G^2 = 8.07$ bei $df = 10$ und liefert keinen Anlaß, die Prozedurunabhängigkeitshypothese zu verwerfen. Dies ist umso erstaunlicher, als der Test auf einem Gesamtstichprobenumfang von knapp 12600 basiert und deshalb sehr hohe Teststärke selbst für geringe Abweichungen von der Übereinstimmungshypothese aufweist. Auch die weitergehenden Hypothesen, daß die beiden *k*-Parameter ($\Delta G^2 = 5.56$ bei $df = 2$) und die beiden *r*-Parameter ($\Delta G^2 = 1.95$ bei $df = 2$) nicht von der Antworttendenzmanipulation beeinflußt werden, lassen sich für den kombinierten RK- und RKG-Datensatz aufrechterhalten.

In Analogie zum Konzept τ-äquivalenter Tests aus der klassischen Testtheorie (z.B. Stumpf, 1996) kann man also bezüglich der einfachen RK-Prozedur und der erweiterten RKG-Prozedur von „äquivalenten Prozeduren" sprechen, die in Verbindung mit dem diskreten Zustandsmodell bzw. dem RKG-Zustandsmodell die Erfassung der gleichen „wahren Werte" — hier: der Strukturparameter r und *k* — ermöglichen.

Damit drängt sich eine weitere Frage auf: Wenn die beiden Prozeduren im beschriebenen Sinne äquivalent sind, sind sie dann auch zwangsläufig gleich gut, so daß die Wahl zwischen ihnen nach Belieben erfolgen kann? Das ist ebensowenig der Fall wie in der klassischen Testtheorie. Dort sind τ-äquivalente Tests auch nicht notwendigerweise gleich gut, d.h. „streng parallel" (Stumpf, 1996). Ein Test kann in dem Sinne besser sein, daß er höhere Reliabilität aufweist, auch wenn sich die wahren Werte von denen eines Konkurrenztests nicht unterscheiden. In Analogie dazu kann man äquivalente Prozeduren nach ihrer „Effizienz" hinterfragen, d.h. nach der Größe des Standardschätzfehlers, mit dem Schätzer der interessierenden Parameter behaftet sind. Von zwei äquivalenten Prozeduren ist diejenige

effizienter, die bei gleichem Stichprobenumfang kleinere Standardschätzfehler der interessierenden Parameter garantiert.

Die RKG-Prozedur erweist sich als effizienter als die RK-Prozedur, was damit zusammenhängt, daß das diskrete RKG-Zustandsmodell „restriktiver" als das diskrete Zustandsmodell zur einfachen RK-Prozedur ist. Im Vergleich zum RK-Zustandsmodell bezieht sich das RKG-Modell im einfachsten Fall (ein Zielitem- und ein Distraktoritemdatensatz) auf zwei zusätzliche empirische Kategorien, umfaßt aber nur einen Rateparameter mehr. Wenn mehrere experimentelle Bedingungen simultan analysiert werden, kann daraus eine erhebliche Freiheitsgraddifferenz erwachsen. Analysiert man beispielsweise die RK-Daten aus Tabelle 9.4 und die RKG-Daten aus Tabelle 9.8 mit Modellen, die die Gleichheit der *d*-. der beiden *r*- und der beiden *k-Parameter* über die beiden Antworttendenzbedingungen hinweg behaupten, so hat der entsprechende Modellanpassungstest zum diskreten Zustandsmodell nur $df = 3$ Freiheitsgrade, während der entsprechende Test zum RKG-Zustandsmodell auf $df = 7$ kommt.

Wie man in Tabelle 9.11 anhand eines konkreten Beispiels erkennt, kann der Effekt der Prozedurwahl auf die Standardschätzfehler gerade bei kleinen Stichprobenumfängen sehr drastisch sein. Die Standardschätzfehler des diskreten Zustandsmodells sind dort auch bei einem $N = 768$ pro Antworttendenzbedingung noch größer als die Standardschätzfehler des RKG-Modells bei $N = 48$. Diese Standardschätzfehler wurden unter der Annahme errechnet, daß das Zustandsmodell und das RKG-Zustandsmodell mit antworttendenzunabhängigen Gedächtnisparametern gelten, welche den empirisch geschätzten Werten sehr ähnlich sind, nämlich $d = .10$, $k_{\text{flach}} = .02$, $k_{\text{tief}} = .06$, $r_{\text{flach}} = .01$ und $r_{\text{tief}} = .06$. Für die Rateparameter wurden $g_{\text{r}} = .10$ für die liberale und $g_{\text{r}} = .01$ für die konservative Antworttendenzbedingung zugrunde gelegt. Die entsprechende Werte für g_{k} lauteten .30 bzw. .25 beim diskreten Zustandsmodell und .15 bzw. .10 beim RKG-Zustandsmodell. Der Parameter g_{g}, der nur im RKG-Zustandsmodell vorkommt, betrug .25 für die liberale und .10 für die konservative Antworttendenzbedingung.

Tabelle 9.11:
Standardschätzfehler der *r*- und der *k*-Schätzer für die flache und die tiefe Verarbeitungsbedingung bei Verwendung der einfachen RK-Prozedur und der erweiterten RKG-Prozedur; für verschiedene Stichprobenumfänge. Die zugrundeliegenden Parameterwerte werden im Text erläutert.

	$N = 48$		$N = 96$		$N = 192$		$N = 384$		$N = 768$	
Par.	RK	RKG	RK	RKG	RK	RKG	RK	RKG	RK	RKG
r_{flach}	.681	.121	.481	.086	.340	.061	.240	.043	.170	.030
r_{tief}	.443	.120	.313	.085	.221	.051	.157	.042	.111	.030
k_{flach}	1.77	.114	1.25	.080	.884	.057	.625	.040	.442	.028
k_{tief}	1.73	.134	1.22	.095	.862	.067	.610	.047	.431	.033

Die in Tabelle 9.11 angegebenen Standardschätzfehler beziehen sich jeweils auf die restringierten Modellvarianten, die gleiche Strukturparameter d, k_{flach}, k_{tief}, r_{flach} und r_{tief} über die beiden Antworttendenzbedingungen hinweg annehmen. Die angegebenen Stichprobenumfänge gelten jeweils für eine Antworttendenzbedingung. N = 48 bedeutet also beispielsweise, daß in einer Antworttendenzbedingung 48 Testitems beantwortet werden, davon 12 ursprünglich tief verarbeitete Items, 12 ursprünglich flach verarbeitete Items und 24 neue Distraktoritems. Jede weitere Stufe von N entspricht einer Verdopplung des Stichprobenumfangs bei gleicher Aufteilung auf tief verarbeitete (25%), flach verarbeitete (25%) und neue Items (50%). Schon bei N = 48 sind die Standardschätzfehler des RKG-Zustandsmodells durchaus akzeptabel, was für das diskrete Zustandsmodell zur einfachen RK-Prozedur offenbar nicht gilt.[98]

Obwohl also das Hauptargument, das Gardiner et al. (1997) für die Entwicklung der RKG-Prozedur genannt haben — Dekontaminierung der R- und K-Antwortkategorien von Antworttendenzen — nach den hier diskutierten Befunden zurückgewiesen werden muß, ist die RKG-Prozedur dennoch eine Verbesserung der RK-Prozedur in dem Sinne, daß sie in Verbindung mit einem geeigneten Meßmodell eine effizientere Schätzung der Wahrscheinlichkeit bewußter Erinnerungen (r) und der Vertrautheit ohne bewußte Erinnerung (k) ermöglicht.

Dieses Ergebnis weist Parallelen zu einem Resultat von Garcia-Pérez (1993) auf. Garcia-Peréz hat Multiple-Choice-Tests, bei denen aus einer Menge von Antworten die richtige auszuwählen ist, ohne und mit zusätzlicher Antwortoption „keine der o.g. Antworten ist richtig" auf dem Hintergrund psychometrischer Modelle mit diskretem Zustandsraum untersucht. Er fand, daß die oftmals kritisierte Antwortoption „none of the above" eine Verbesserung des Tests in dem Sinne sein kann, daß sie die Schätzung des Fähigkeitsparameters mit kleinerem Standardschätzfehler ermöglicht. Man sieht, daß Vorsicht bei der voreiligen Bewertung von Prozedurergänzungen am Platze ist. Manche Ergänzung, die psychologisch unergiebig wirkt und keinen zusätzlichen Informationsgewinn mit sich zu bringen scheint — sei es die „Guess"-Antwortkategorie in der erweiterten

[98]Die praktische Berechnung der Standardschätzfehler erfolgt am besten über die inversen Fisher-Informationsmatrizen zu den betrachteten Modellen, wobei die wahren Parameter in der zugrundeliegenden Population einzusetzen sind (vgl. Abschnitt 4.2.5). Die Standardschätzfehler in Tabelle 9.11 wurden mittels „AppleTree" (Rothkegel, 1998; 1999) errechnet, indem zunächst die erwarteten Häufigkeiten zum gegeben N und den fiktiven Populationsparametern generiert wurden. Diese erwarteten Häufigkeiten wurden erneut mit den geeignet restringierten Modellen analysiert, wobei die Ausgabe der inversen Fisher-Informationsmatrix angewählt wurde. Die Modellanpassung führt trivialerweise zu den Parameterwerten, die den erwarteten Häufigkeiten zugrunde gelegt wurden, und natürlich zu residuenfreier Datenanpassung (G^2 = 0). Wichtig ist allein die inverse Fisher-Informationsmatrix. In der Hauptdiagonalen stehen die Varianzen der Schätzer. Die Standardschätzfehler erhält man, indem man jeweils die Wurzel aus den Varianzen zieht.

RKG-Prozedur oder die Antwortoption „none of the above" bei Multiple-Choice-Tests —, kann in Verbindung mit geeigneten Meßmodellen dennoch sehr wertvoll sein, weil sie eine effizientere Schätzung der psychologisch primär interessierenden Modellparameter ermöglicht.

9.3 Zusammenfassung

Die Remember-Know-Prozedur (RK-Prozedur) hat sich im Anschluß an die klassischen Arbeiten von Tulving und Gardiner seit Mitte der achtziger Jahre zu einem einflußreichen Instrument bei der Überprüfung von Zwei-Prozeß- oder Zwei-System-Konzeptionen des menschlichen Gedächtnisses entwickelt. Offensichtlich problematisch ist allerdings — wie oftmals praktiziert — die Benutzung der Häufigkeiten von R- oder K-Antworten als Maße für explizite bzw. implizite Gedächtnisleistungen. Wie von Strack und Förster (1995) sowie Hirshman und Henzler (1998) gezeigt, sind diese Häufigkeiten mit Antworttendenzen kontaminiert. Die Erweiterung der RK-Prozedur um eine zusätzliche „Guess"-Antwortkategorie, die geratene „Alt"-Antworten absorbieren soll (Gardiner et al., 1997), löst dieses Problem nicht. Dies wurde im vorliegenden Kapitel anhand neuer Daten gezeigt. Werden Antworttendenzen stark genug angeregt, so schlagen sie auf R- und K-Antworten auch dann durch, wenn die Möglichkeit besteht, subjektiv geratene „Alt"-Antworten mit „Guess"-Urteilen (G-Antworten) zu markieren. Offensichtlich sind Antworttendenzen und Rateprozesse nur partiell bewußtseinsfähig. Manches, was subjektiv geraten ist, ist objektiv überzufällig korrekt und vice versa (vgl. bereits Peirce & Jastrow, 1884, zitiert nach Kihlstrom et al., 1992).

Potentiell lösbar wird das Kontaminierungsproblem im Rahmen von Meßmodellen, die den simultanen Einfluß von Gedächtnis- und Urteilsprozessen auf R-, K-, N- und ggf. G-Antworten modellieren. Zwei Klassen von Meßmodellen wurden im vorliegenden Kapitel evaluiert: Kontinuum-Modelle und diskrete Zustandsmodelle. Kontinuum-Modelle basieren auf Ein-Prozeß-Konzeptionen des Gedächtnisses und postulieren im Einklang mit der Signalentdeckungsmodellen für Ratingskalen, daß sich die Antwortkategorien der einfachen oder erweiterten RK-Prozedur auf einem Kontinuum der „Vertrautheit" oder der „Gedächtnisspurenstärke" ordnen lassen. Diskrete Zustandsmodelle sind Verallgemeinerungen von Zwei-Hochschwellen-Modellen der Rekognition und basieren auf der Idee, daß sich ein Item in einem von drei Zuständen befindet: (a) bewußt erinnerbar, (b) nicht bewußt erinnerbar, aber vertraut und (c) weder bewußt erinnerbar noch vertraut („Zustand der Unsicherheit"). In Abhängigkeit vom Zustand werden dann R-Urteile bzw. K-Urteile gegeben oder es wird geraten, wobei — je nach Ausprägung von Antworttendenzen — prinzipiell jede Antwortkategorie resultieren kann.

Während das diskrete Zustandsmodell zur einfachen RK-Prozedur in der Evaluation nur minimal besser abschnitt als das Kontinuum-Modell, war das erweiterte diskrete Zustandsmodell zur RKG-Prozedur der Verallgemeinerung des Kontinuum-Modells auf die RKG-Prozedur klar überlegen. Das diskrete RKG-Zustandsmodell lieferte zwar keine durchweg guten Anpassungsindizes, kann aber noch als empirisch angemessen bezeichnet werden. Insgesamt sprechen die Ergebnisse somit eher für Zwei-Prozeß-Modelle und gegen Ein-Prozeß-Modelle des Gedächtnisses. Dies gilt insbesondere bei Verwendung der erweiterten RKG-Prozedur und — wie bei Gültigkeit eines Zwei-Prozeß-Modells zu erwarten — für Versuchsbedingungen mit mutmaßlich hohem Anteil bewußter Erinnerungen.

Eine vergleichende Untersuchung des einfachen und des erweiterten diskreten Zustandsmodells anhand randomisierter Probandengruppen ergab, daß die RK- und die RKG-Prozedur äquivalent in dem Sinne sind, daß sie in Verbindung mit den beiden Zustandsmodellen identische Gedächtnisparameter erfassen. Unter dem Aspekt der Effizienz erwies sich das erweiterte RKG-Modell als überlegen, d.h. es garantiert bei gegebenem Stichprobenumfang kleinere Standardschätzfehler für die zentralen Modellparameter als das diskrete Zustandsmodell zur einfachen RK-Prozedur.

Wenn weitere Evaluationen anhand unabhängig gewonnener Daten das hier gewonnene Bild bestätigen, kann das erweitere diskrete Zustandsmodell in Verbindung mit der RKG-Prozedur als Methode der Wahl bei zukünftigen Remember-Know-Untersuchungen empfohlen werden. Sofern sich der Vorschlag von Conway et al. (1998) durchsetzt, eine weitere Aufsplittung der K-Antwortkategorie in „familiarity” (F) und „just know” (K) vorzunehmen, wäre eventuell eine Erweiterung des Modells auf die zweifach erweiterte RK-Prozedur mit den fünf Antwortkategorien R, F, K, G und N zu erwägen. Dies würde auf die Hinzunahme eines vierten latenten Zustandes hinauslaufen, den man vielleicht „bloßes Wissen” nennen könnte. Während „bewußte Erinnerungen” dem expliziten Gedächtnis und „Vertrautheit” dem impliziten Gedächtnis zuzurechnen sind, würde „bloßes Wissen” zusätzlich das semantische Gedächtnis repräsentieren — wie ursprünglich von Tulving (1985) angestrebt.

Ob diese Prozedurerweiterung wirklich sinnvoll ist, muß von zahlreichen Validierungsuntersuchungen abhängig gemacht werden, die wegen der Vielzahl der involvieren Gedächtnissysteme notgedrungen sehr aufwendig sein müssen. Die von Conway et al. (1998) vorgelegten Ergebnisse lassen einen solchen Ansatz immerhin lohnenswert erscheinen.

10. Modelle zum Falschinformationseffekt

Unter einem Falschinformationseffekt versteht man die Beeinträchtigung von Gedächtnisurteilen zu einem bestimmten Ereignis durch nachträglich dargebotene Falschinformationen. Das Standard-Untersuchungsparadigma hierzu kann anhand der klassischen Arbeit von Loftus et al. (1978) erläutert werden, auf die in Abschnitt 2.6 bereits kurz eingegangen wurde. Zur Erinnerung sei noch einmal der grundlegende Versuchsablauf erläutert: In einer Experimentalbedingung wird zunächst eine Originalinformation (OI) dargeboten (zumeist in visueller Form, d.h. als Bild oder Film), anschließend wird eine versteckte Falschinformation zu einem bestimmten Detail der OI präsentiert (zumeist in Textform) und anschließend folgt ein Gedächtnistest zur OI. Der Experimentalbedingung (EB) werden eine oder mehrere Kontrollbedingungen (KB) gegenübergestellt, die sich von der EB nur dadurch unterscheiden, daß die nachträgliche Falschinformation entweder durch keine, durch neutrale oder durch zur OI konsistente Nachinformation (NI) ersetzt wird. Von einem „Falschinformationseffekt" spricht man, wenn die Gedächtnisleistung für die OI in der EB schlechter ausfällt als in allen Kontrollbedingungen.

In der Untersuchung von Loftus et al. (1978) war die kritische OI ein Dia, das einen roten Datsun-Sportwagen an einer Straßeneinmündung vor einem bestimmten Verkehrsschild zeigte. Für die Hälfte der Vpn war dies ein Stop-Schild, für die andere Hälfte ein Vorfahrt-Achten-Schild. Zusätzlich wurden weitere Dias gezeigt, die einen Unfall darstellten, in den der Datsun verwickelt war. Anschließend wurde mehrere Fragen zu den Dias gestellt. Die kritische Frage lautete für ein Drittel der Vpn „Fuhr ein anderes Auto an dem roten Datsun vorbei, als dieser am Stop-Schild anhielt?" Für ein weiteres Drittel wurde die gleiche Frage gestellt, wobei lediglich „Stop-Schild" durch „Vorfahrt-Achten-Schild" ersetzt wurde. Das letzte Drittel schließlich bekam die Frage ohne Hinweis auf das Verkehrsschild. Da die NI mit der OI komplett gekreuzt wurde, ergaben sich insgesamt sechs Versuchsgruppen mit falscher, konsistenter und neutraler NI für zwei verschiedene Originalformationen. Nach einem bestimmten Retentionsintervall, in dem die Vpn Distraktortätigkeiten zu verrichten hatten, folgte dann ein 2-AFC-Rekognitionstest für die OI, wobei als kritisches Itempaar die beiden Dias dargeboten wurden, die den roten Datsun vor dem Stop-Schild bzw. dem Vorfahrt-Achten-Schild zeigten. Die Vpn mußten entscheiden, welches der beiden Dias mit der OI identisch war. Loftus et al. (1978) fanden in mehreren Experimenten massive Falschinformationseffekte, d.h. mehr inkorrekte Gedächtnisurteile, wenn zuvor nachträgliche Falschinformationen präsentiert worden waren.

Später hat man dieses Paradigma in verschiedener Hinsicht variiert, vor allem in bezug auf den Gedächtnistest. Es zeigte sich immer wieder, daß die Stärke

der gefundenen Falschinformationseffekte entscheidend vom verwendeten Gedächtnistest abhängt (vgl. Abschnitt 2.6). Dies ist eine der vielen Gemeinsamkeiten, die das Paradigma nachträglicher Falschinformation mit dem klassischen Paradigma der retroaktiven Hemmung aufweist. Folgerichtig wird es in manchen Lehrbüchern auch darunter subsumiert (z.B. Baddeley, 1997). Auf der anderen Seite sind die speziellen Eigenschaften der OI (Bild) und der NI (Wort) wie auch die semantische Inkonsistenz zwischen OI und NI eine Besonderheit des Falschinformationsparadigmas, die es von klassischen Untersuchungen zur retroaktiven Hemmung abhebt. Wegen der praktischen Relevanz des Falschinformationseffekts hat er insbesondere in der Psychologie der Zeugenaussage viel Aufmerksamkeit gefunden (Loftus, 1979).

Man kann im wesentlichen zwischen vier Klassen von Theorien zum Falschinformationseffekt unterscheiden (vgl. z.B. Ayers & Reder, 1998; Sporer, 1993):

1) Die von Elizabeth Loftus selbst vorgeschlagene Gedächtnisbeeinträchtigungstheorie (destructive updating theory, memory impairment theory). Nach dieser Theorie wird die Gedächtnisspur der OI durch die falsche NI zerstört oder zumindest modifiziert.
2) Die Koexistenztheorie und die Blocking-Hypothese (Bekerian & Bowers, 1983; Christiaansen & Ochalek, 1983; Schreiber & Sergant, 1998), nach denen die Gedächtnisspur für die OI zwar intakt, aber als Folge nachträglicher Falschinformation schlechter abrufbar ist.
3) Die Quellenkonfusionshypothese (Lindsay & Johnson, 1989), welche den Falschinformationseffekt auf ungenügendes Quellengedächtnis für die OI und die NI zurückführt. Auch das jüngst von Ayers und Reder (1998) vorgeschlagene „source of activation confusion model", das auf einem Netzwerk-Modell des Gedächtnisses basiert, kann als spezielle Quellenkonfusionshypothese aufgefaßt werden.[99]
4) Die Antworttendenzhypothese (no-conflict theory; misinformation acceptance theory), die auf die bereits in Abschnitt 2.6 besprochene Arbeit von McCloskey und Zaragoza (1985) zurückgeht. Nach dieser Theorie ist weder die Gedächtnisspur noch der Gedächtnisabruf noch das Quellengedächtnis beeinträchtigt; es liegt lediglich ein stärkerer Antwortbias zugunsten der Falschinformation in der EB vor.

Gemeinsam ist den genannten Erklärungsansätzen die Annahme, daß die beobachtbaren Gedächtnisurteile immer mehrere kognitive Prozesse zugleich reflektieren, wie z.B. Speicherung der OI, Abruf der OI, Quellendiskriminierung zwischen

[99] Ayers und Reder (1998, p. 17) schreiben: „It is fair to say that, in some cases, the activation based account merely instantiates the assumptions of the source monitoring account in a mechanistic model of memory".

OI und NI sowie Rateprozesse. Die Theorien unterscheiden sich nur darin, welchen der genannten Prozesse sie als primär ursächlich für den Falschinformationseffekt ansehen.

Der von den meisten Autoren favorisierte Ansatz zur Lösung des daraus resultierenden Dekomponierungsproblems ist der designbasierte Ansatz. Dieser stößt jedoch auf Schwierigkeiten, weil (a) der Versuch der „Dekontaminierung" bestimmter Gedächtnismaße von potentiellen Störeinflüssen häufig auf ungeprüften, z.T. fragwürdigen Annahmen basiert und (b) darüber hinaus die unerwünschte Nebenwirkung hat, daß die Sensitivität der verwendeten Gedächtnisindikatoren auch für die eigentlich interessierenden Gedächtnisprozesse herabgesetzt ist (vgl. Abschnitt 2.6).

Diese Randbedingungen lassen den Versuch, das Dekomponierungsproblem modellbasiert zu lösen, geradezu zwingend erscheinen. Es erstaunt deshalb nicht, daß der Falschinformationseffekt schon bald nach seiner experimentellen Etablierung auch mit Hilfe multinomialer Modelle untersucht wurde (Batchelder, Riefer & Walton, 1990; Wagenaar & Boer, 1987). Erstaunlicher ist es schon, daß diese Ansätze bislang insgesamt wenig Einfluß auf die Forschung zum Falschinformationseffekt hatten.

Die Ursachen hierfür sind vielfältig. Zum einen erweist es sich offenkundig als schwer, Modelle zum Falschinformationseffekt zu finden, die eine akzeptable Datenanpassung ermöglichen. Batchelder et al. (1990) haben beispielsweise den Quellendiskriminationstest von Lindsay und Johnson (1989) verwendet, bei dem die Vpn zu jedem Testitem angeben mußten, ob es (a) in der OI und der NI vorkam, (b) nur in der OI vorkam, (c) nur in der NI oder (d) weder in der OI noch in der NI. Sie erhielten für ihre Modelle G^2-Anpassungsstatistiken in der Größenordnung $16.44 \leq G^2 \leq 43.40$ bei nur 2 Freiheitsgraden. Dies ist offensichtlich keine Basis, um mit den Modellen weiterarbeiten zu können.

Wagenaar und Boer (1987) haben demgegenüber akzeptable Modellanpassungsindizes erhalten. Ihre Arbeit weist jedoch andere Probleme auf, die einer Verbreitung des Ansatzes bislang entgegengestanden haben. Zunächst einmal stellten sie nicht *ein* Modell vor, aus dem sich alle Theorien zum Falschinformationseffekt durch Parameterrestriktionen herleiten lassen, sondern zwei statistisch unverbundene Vier-Parameter-Modelle, eines zur Gedächtnisbeeinträchtigungstheorie (Modell 1 mit Parametern p, c, q und d) und eines zur Koexistenzhypothese (Modelle 2 mit Parametern p, c, q und s). Lediglich das dritte Modell zur Antworttendenzhypothese läßt sich durch Parameterrestriktionen aus Modell 1 (mit $d = 0$) oder aus Modell 2 (mit $s = 0$) herleiten.[100] Die Parameter p, c, und q haben in allen

[100]Die EQN-Dateien, die zur Analyse der drei Modelle mittels der Programme MBT, GPT oder AppleTree erforderlich sind, sind ebenso im Anhang (Abschnitt 3 - 6) angegeben wie die MDT-Datei, die die von Wagenaar und Boer berichteten Daten enthält.

Modellen die gleiche psychologische Interpretation, werden aber wegen der unterschiedlichen Modellgleichungen auf verschiedene Werte geschätzt. Als einziges Kriterium zur Entscheidung zwischen den verschiedenen Modellen boten die Autoren globale Pearson χ^2-goodness-of-fit-Werte (X^2) an. Diese unterschieden sich nur minimal zwischen den verschiedenen Modellen, was die Autoren als Bestätigung des parameterärmsten Modells, des Antworttendenzmodells (no-conflict theory), interpretierten.

Dieses Vorgehen setzt zunächst einmal voraus, daß die durchgeführten Modellgeltungstests hinreichende Teststärke zur Evaluation der Modelle einerseits und zur Diskriminierung zwischen den Modellen andererseits aufwiesen. Führt man entsprechende Post-hoc-Teststärkekontrollen mit GPOWER (Buchner et al., 1992) durch, so ergeben sich nur in bezug auf den globalen χ^2-Modellanpassungstest zum „destructive-updating model" und zum „coexistence model" überzeugende Ergebnisse: Bei $df = 5$, $\alpha = .05$ und $N = 562$ beträgt die Teststärke für kleine Modellverletzungen im Sinne von Cohen (1988) ($w = .10$) zwar nur $(1-\beta) = .40$, für moderate Modellverletzungen ($w = .20$) aber immerhin schon .97. Dies ist akzeptabel. Ernüchternde Ergebnisse erhält man jedoch, wenn man den Modellvergleich zwischen Antworttendenzhypothese und den beiden anderen Modellen betrachtet. Dies liegt vor allem daran, daß der Cohensche Effektstärkeparameter w nur wenig durch den Parameter d, der zwischen dem „destructive-updating model" und der Antworttendenzhypothese diskriminiert, wie auch durch den Parameter s, welcher zwischen „coexistence model" und der Antworttendenzhypothese trennt, beeinflußt wird.

Wenn man das Antworttendenzmodell testet, während eines der beiden allgemeineren Modelle gilt, kann man die Größe der Parameter d bzw. s der allgemeineren Modelle so in das Effektstärkemaß w umrechnen wie es in den Abschnitten 4.3.3.1 und 4.3.3.2 bereits geschildert wurde. Führt man die entsprechenden Berechnungen unter Zugrundelegung der von Wagenaar und Boer (1987) gewählten Stichprobenaufteilung auf die drei Untersuchungsbedingungen „falsche Nachinformation" ($n_1 = 250$), „konsistente Nachinformation" ($n_2 = 170$) und „neutrale Nachinformation" ($n_3 = 142$) durch, und nimmt man weiterhin an, daß die anderen Modellparameter $p = .5$, $c = .57$ und $q = .5$ betragen — was die plausibelsten Werte im Lichte der Ergebnisse von Wagenaar und Boer (1987) sind — so ergibt sich, daß w für den Modellvergleich zwischen Antworttendenzhypothese und Koexistenzhypothese erst ab $s > .40$ kleine Effektgrößen ($w > .10$) produziert und maximal (d.h. bei $s = 1.00$) nur $w = .23$ zur Folge hat. Beim Modellvergleich zwischen Antworttendenzhypothese und „destructive-updating model" sieht es noch ungünstiger aus: Hier werden kleine Effektstärken erst ab $d = .70$ erreicht, und die maximale Effektstärke bei $d = 1.00$ beträgt nur $w = .15$.

Die Konsequenzen für die Teststärke des Differenzen-χ^2-Tests mit $df = 1$ und $N = 562$ sind für beide Obermodelle und zwei Typ-1-Fehlerwahrscheinlichkeiten in Abbildung 10.1 graphisch veranschaulicht. Man sieht, daß bei $\alpha = .05$ erst Parameterwerte ab $d = s = .80$ mit überzeugender Teststärke entdeckt werden, wenn die getestete Nullhypothese $d = s = 0.00$ behauptet. Selbst bei Verwendung von $\alpha = .10$ muß d bzw. s mindestens .30 betragen, damit der statistische Test die Verletzung der Antworttendenzhypothese eher aufdecken kann als ein Münzwurf, der immer mit Wahrscheinlichkeit .5 zugunsten der Alternativhypothese entscheidet (sic!). Unterstellt man, daß die Wahrscheinlichkeit von Gedächtnisbeeinträchtigungen (d) oder Abrufbeeinträchtigungen (s) im Untersuchungsparadigma von Wagenaar und Boer unter 0.50 liegt, so kann es schon aus rein statistischen Gründen nicht überraschen, daß Wagenaar und Boer (1987) das Antworttendenzmodell nicht verwerfen konnten.

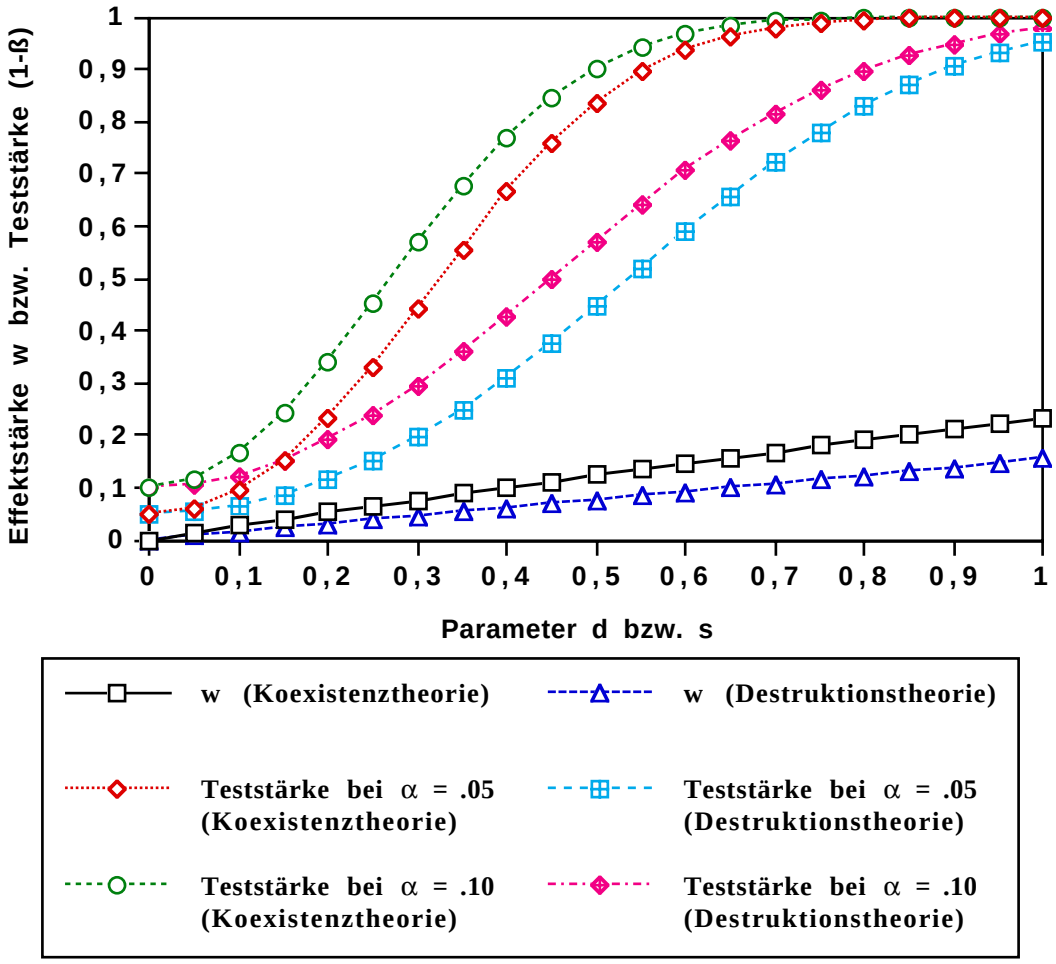

Abbildung 10.1:
Effektstärken w sowie Teststärken für $\alpha = .05$ und $\alpha = .10$ beim Differenzen-χ^2-Test der Antworttendenzhypothese (Nullhypothese $d = s = 0$), wenn tatsächlich die Koexistenztheorie (mit $s > 0$) bzw. die „destructive-updating theory" (mit $d > 0$) für die Untersuchung von Wagenaar und Boer (1987) zutreffen sollte.

Schon die Teststärkeproblematik macht verständlich, warum die Arbeit von Wagenaar und Boer (1987) nur wenig Wirkung entfalten konnte: Die Theorien von Loftus sowie Bekerian und Bowers hatten in dieser Untersuchung nur eine geringe Chance, sich empirisch zu bewähren, selbst wenn sie zutreffen sollten. Hinzu kommt, daß die Autoren keine Validierungsexperimente zu ihren Modellen berichteten, so daß die psychologische Interpretation der Modellparameter fragwürdig ist. Dies wiegt umso schwerer, als einige vereinfachende und z.T. auch unplausible Annahmen (z.B. über Ratewahrscheinlichkeiten) in die Modelle eingehen, denen man ohne eine empirisch fundierte Evaluation der Modelle kaum zustimmen kann.

Im Prinzip sind die bislang genannten Kritikpunkte zwar für die Untersuchung von Wagenaar und Boer (1987) von Bedeutung, nicht aber für die von den Autoren vorgeschlagenen Modelle, denn man könnte in Nachfolgeuntersuchungen die fehlenden Validierungsexperimente nachholen und auch deutlich größere Stichprobenumfänge verwenden, so daß die Teststärkeproblematik beim Modellvergleich entschärft wird. Daß dies bislang noch nicht geschehen ist, dürfte auf der wohl entscheidenden Schwäche des Ansatzes beruhen: Die Modelle beziehen sich auf keinen der bei Untersuchungen zum Falschinformationseffekt üblichen Gedächtnistests, sondern auf ein sehr spezielles Untersuchungsparadigma — Globalrekognition der OI (Ampel) mit anschließender Detailrekognition (Farbe des aufleuchtenden Ampellichts) — und zudem auf sehr spezielles Itemmaterial. Man muß also bei Folgeuntersuchungen die Originaluntersuchung mehr oder minder genau replizieren — was den Raum für Validierungsexperimente und andere Modellanwendungen stark einschränkt — oder aber die Modelle auf andere Gedächtnistests adaptieren, was praktisch der Konstruktion neuer Modelle gleichkommt.

Ich werde in den folgenden Abschnitten zwei Versuche schildern, die genannten Probleme der Ansätze von Batchelder et al. (1990) und Wagenaar und Boer (1987) zu überwinden, wobei selbstverständlich einige Gedanken der erwähnten Modelle erneut aufgegriffen werden. Hauptziel der folgenden Abschnitte ist es, Modelle vorzustellen, die sich auf Gedächtnistests beziehen, welche in Untersuchungen zum Falschinformationseffekt häufig Verwendung finden, und für die darüber hinaus zumindest erste Schritte zu einer erfolgreichen Modellvalidierung bereits absolviert wurden. In Abschnitt 10.1 wird ein multinomiales Modell zum 2AFC-Rekognitionstest von Loftus et al. (1978) vorgestellt und evaluiert, in Abschnitt 10.2 ein Modell zum Quellendiskriminationstest, wie er zuerst von Lindsay und Johnson (1989) eingesetzt wurde. Eine zusammenfassende Diskussion der Ergebnisse zu beiden Modellen schließt das Kapitel ab (Abschnitt 10.3).

10.1 Ein Modell zum 2AFC-Rekognitionstest

Loftus et al. (1978) haben argumentiert, daß ein 2AFC-Rekognitionstest, bei dem die Vp zwischen OI und falscher NI wählen muß, am besten geeignet ist, Gedächtnisbeeinträchtigungseffekte als Folge nachträglicher Falschinformation nachzuweisen. Schon dies allein läßt es ratsam scheinen, ein Modell zum 2AFC-Rekognitionstest zu konstruieren, das geeignet ist, wenigstens die wichtigsten kognitiven Einflußfaktoren auf das Antwortverhalten voneinander zu separieren. Ein solches Modell wurde in Abschnitt 2.6 bereits vorgestellt und soll hier erneut aufgegriffen werden. Dieses Modell zielt nicht darauf ab, alle bislang vorgeschlagenen Theorien zum Falschinformationseffekt gegeneinander zu testen. Vielmehr geht es ausschließlich darum, die Antworttendenzhypothese von McCloskey und Zaragoza (1985) gegen solche Theorien zu testen, die Gedächtnisprozesse für den Falschinformationseffekt verantwortlich machen. Von untergeordneter Bedeutung ist dabei, ob Gedächtnisbeeinträchtigungen nun als Modifikationen der OI-Gedächtnisspur, als Abrufbeeinträchtigungen oder als Quellenkonfusionen zu interpretieren sind. Die Frage ist lediglich, ob überhaupt *irgendwelche* Gedächtnisbeeinträchtigungen vorliegen.

Trotz des vergleichsweise bescheidenen Zieles der Modellkonstruktion sind wir in Abschnitt 2.6 auf Schwierigkeiten gestoßen, die mit dem Problem der Identifizierbarkeit der Modellparameter zusammenhängen: Drei Modellparameter — die OI-Abrufwahrscheinlichkeit *r*, der globale Gedächtnisbeeinträchtigungseffekt *m* und der Antwortbias *b* zugunsten der Falschinformation — können nicht auf der Basis von nur zwei Modellgleichungen bestimmt werden.

In den Abschnitten 4.4 bis 4.6 wurden mehrere Wege vorgestellt, wie man mit diesem Identifizierbarkeitsproblem umgehen kann. Da im gegebenen Falle Parameterrestriktionen einerseits und Erweiterungen der Beobachtungskategorien andererseits als Abkehr vom Ziel der Modellkonstruktion gewertet werden müssen, bleibt eigentlich nur die Erweiterung des Modells auf ein Mehr-Gruppen-Modell, in dem dann bestimmte gruppenübergreifende Parameterrestriktionen möglich werden.

Erfreulicherweise haben schon Loftus et al. (1978) ein Experiment berichtet, das geeignete Daten für ein Mehr-Gruppen-Modell zum 2AFC-Test liefert. In ihrem Experiment 3 wurden insgesamt 648 Vpn in 27 verschiedenen Versuchsbedingungen mit einem 2AFC-Rekognitionstest untersucht, wobei die Vpn u.a. entscheiden mußten, ob in der OI ein Stop-Schild oder ein Vorfahrt-Achten-Schild zu sehen war. Eine Gruppe von 72 Vpn absolvierte das oben geschilderte Experiment derart, daß das Betrachten der Diaserie mit dem roten Datsun-Sportwagen (Phase 1), die Bearbeitung des Fragebogens mit falscher, konsistenter oder keiner Nachinformation zum Verkehrsschild (Phase 2) und der abschließende Rekognitionstest (Phase 3) zeitlich unmittelbar aufeinander folgten. Für vier weitere Gruppen mit

jeweils 72 Vpn folgte Phase 2 des Experiments unmittelbar auf Phase 1, der 2AFC-Test dann jedoch erst nach einem Retentionsintervall von 20 Minuten, einem Tag, zwei Tagen oder einer Woche. Die gleichen Retentionsintervalle wurden in vier weiteren Gruppen mit jeweils 72 Vpn benutzt, in denen allerdings der Nachinformationsfragebogen nicht unmittelbar nach Phase 1, sondern erst kurz vor dem Rekognitionstest dargeboten wurde. Jede dieser neun Versuchsgruppen wurde noch einmal in jeweils drei gleich große Gruppen mit 24 Vpn aufgeteilt, die entweder falsche, konsistente oder keine Nachinformation zur OI bekamen.

Die Ergebnisse zum kritischen Itempaar im Rekognitionstest sind für alle Versuchsgruppen in Abbildung 2 bei Loftus et al. (1978, S. 24) dargestellt. Ich habe die relativen Häufigkeiten korrekter Urteile mit Hilfe von Millimeterpapierfolie aus dieser Abbildung abgelesen. Dies führte zu den Schätzungen, die hier in Tabelle 10.1 wiedergegeben sind. Ablesevergleiche mit einer weiteren, unabhängig ablesenden Person zeigten, daß bei den Schätzungen eine Ungenauigkeit von ± .01 einkalkuliert werden muß. Dies dürfte für die zu berichtende Reanalyse der Daten jedoch unerheblich sein, da sich die Ungenauigkeiten beim Umrechnen der relativen Häufigkeiten in absolute Häufigkeiten wegen der Rundung auf natürliche Zahlen wieder weitgehend aufheben.

Tabelle 10.1:
Relative Häufigkeiten korrekter Rekognitionsurteile, geschätzt aus Abbildung 2 bei Loftus et al. (1978, S. 24).

	Unmittelbare NI			Verzögerte NI		
Intervall	wahr	keine	falsch	wahr	keine	falsch
0 Min.	.84	.89	.39			
20 Min.	.76	.80	.43	.72	.59	.43
1 Tag	.75	.72	.47	.76	.64	.39
2 Tage	.71	.57	.43	.71	.48	.29
1 Woche	.63	.51	.55	.68	.52	.18

Da die Vpn-Stichprobe pro Zelle der Ergebnismatrix n = 24 Vpn umfaßt, lassen sich die relativen Häufigkeiten leicht in absolute Häufigkeiten von korrekten und inkorrekten Urteilen umrechnen. Die entsprechenden Ergebnisse sind für die Versuchsbedingungen mit falscher und keiner Nachinformation in Abschnitt 21 des Anhangs wiedergegeben. Nur auf diese Versuchsbedingungen bezieht sich das in Abschnitt 2.6 vorgestellte Modell, und deshalb bleiben die Daten zur Versuchsbedingung mit konsistenter Nachinformation hier außer Betracht.

Das Modell gemäß Abbildung 2.1 wurde auf neun Versuchsbedingungen erweitert. Bei dieser Erweiterung sind folgende Annahmen zugrunde gelegt worden:

1) Da die NI in der Kontrollbedingung nicht mit der OI interferieren kann, hängt die OI-Abrufwahrscheinlichkeit in der KB (Parameter *r*) nicht vom Zeitpunkt der NI-Darbietung ab, sondern ausschließlich von der Länge des Retentionsintervalls. Folglich gibt es nur fünf *r*-Parameter, für jedes der fünf Retentionsintervalle einen. Wenn die Parameter valide Maße der Gedächtnisleistung sind, dann sollte gelten: $r_{(0\ \text{Min.})} \geq r_{(20\ \text{Min.})} \geq r_{(1\ \text{Tag})} \geq r_{(2\ \text{Tage})} \geq r_{(1\ \text{Woche})}$. Hierbei stehen die Indizes für die Länge des Retentionsintervalls.
2) Der Ratebias *b* zugunsten der falschen NI hängt nicht vom Retentionsintervall, sondern ausschließlich vom zeitlichen Abstand zwischen NI und Rekognitionstest ab, denn je größer dieser Abstand ist, um so größer ist auch die Wahrscheinlichkeit, daß die NI zum Zeitpunkt des Gedächtnistests gar nicht mehr abrufbar ist. Folglich können die *b*-Parameter für die fünf Bedingungen, bei denen der Test unmittelbar auf die NI folgt, gleichgesetzt werden. Dies betrifft die Bedingung mit 0 Minuten Retentionsintervall einerseits und die vier Bedingungen mit „verzögerter NI" andererseits. Analog zum obigen Argument für die *r*-Parameter sollte $b_{(0\ \text{Min.})} \geq b_{(20\ \text{Min.})} \geq b_{(1\ \text{Tag})} \geq b_{(2\ \text{Tage})} \geq b_{(1\ \text{Woche})} \geq 0.5$ gelten; in diesem Fall stehen die Indizes allerdings für die Länge des Zeitintervalls zwischen NI und Rekognitionstest.
3) Ein etwaiger Gedächtnisbeeinträchtigungseffekt hängt ausschließlich vom Zeitintervall zwischen OI und NI ab; folglich können die fünf *m*-Parameter der Bedingungen mit „unmittelbarer NI" gleichgesetzt werden. Es gibt demnach nur fünf *m*-Parameter, nämlich $m_{(0\ \text{Min.})}$, $m_{(20\ \text{Min.})}$, $m_{(1\ \text{Tag})}$, $m_{(2\ \text{Tage})}$ und $m_{(1\ \text{Woche})}$, wobei der Index für das Zeitintervall zwischen OI und NI steht.

Das aus diesen Annahmen resultierende Modell ist gedächtnispsychologisch plausibel und hat statt der 9·3 = 27 Parameter des unrestringierten Modells nur noch 3·5 = 15 freie Parameter. Das restringierte Modell ist im Gegensatz zum unrestringierten Modell identifizierbar, und der entsprechende Modellgeltungstest hat $df = 18 - 15 = 3$ Freiheitsgrade.

Dieses Neun-Gruppen-Modell wurde nach der ML-Methode mittels AppleTree (Rothkegel, 1998; 1999) an die aus Tabelle 10.1 geschätzten Daten angepaßt. Ungleichheitsrestriktionen blieben hierbei zunächst unberücksichtigt.[101] Da sich herausstellte, daß einer der *b*-Parameter ($b_{(1\ \text{Woche})}$) auf einen Wert kleiner als .5 geschätzt wurde, wurde mittels eines Hilfsparameters h_5 die Ungleichheitsrestriktion $b_{(1\ \text{Woche})} \geq 0.5$ zusätzlich realisiert.[102] Aus dem Hilfsparameter erhält man den interessierenden Parameter wie folgt: $b_{(1\ \text{Woche})} = .5 + .5 \cdot h_5$.

[101] Die entsprechende Modellbeschreibungsdatei LMBOUR9G.EQN ist in Abschnitt 19 des Anhangs angegeben, die zugehörige Dateneingabedatei LMB789G.MDT in Abschnitt 21.

[102] Die entsprechende Modellbeschreibungsdatei LMBMUR9G.EQN ist in Abschnitt 20 des Anhangs wiedergegeben.

Das aus dieser zusätzlichen Ungleichheitsrestriktion resultierende Modell paßt gut zu den Daten ($G^2_{(3)} = 4.04$).[103] Fügt man die Annahme hinzu, daß alle Gedächtnisbeeinträchtigungsparameter m_j, $j = 1, ..., 5$, gleich sind, ergibt sich keine signifikante Verschlechterung der Anpassungsgüte ($\Delta G^2_{(4)} = 1.74$). Der globale, über alle Versuchsbedingungen hinweg gleichgesetzte m-Parameter wird auf immerhin $\hat{m} = .62$ geschätzt und ist signifikant von null verschieden, sofern $\alpha = .10$ zugrunde gelegt wird ($\Delta G^2_{(1)} = 3.78$). Abbildung 10.2 zeigt, daß die Parameterschätzungen für die Abrufwahrscheinlichkeiten r_j, und die Antworttendenzen b_j, deskriptiv dem prognostizierten Muster folgen. Die *r_j-Parameterschätzungen* für die fünf Retentionsintervalle unterscheiden sich signifikant ($\Delta G^2(4) = 16.41$), nicht jedoch die b_j,-Schätzungen, deren Anpassungsstatistik den kritischen Wert zu $\alpha = .10$ knapp verfehlt ($\Delta G^2(4) = 7.41$). Die weitergehende Nullhypothese, daß keine Antworttendenz zugunsten der falschen NI vorliegt (H_0: $b_{(0\ \text{Min.})} = b_{(20\ \text{Min.})} = b_{(1\ \text{Tag})} = b_{(2\ \text{Tage})} = b_{(1\ \text{Woche})} = 0.5$), muß demgegenüber klar verworfen werden ($\Delta G^2_{(1)} = 11.97$).

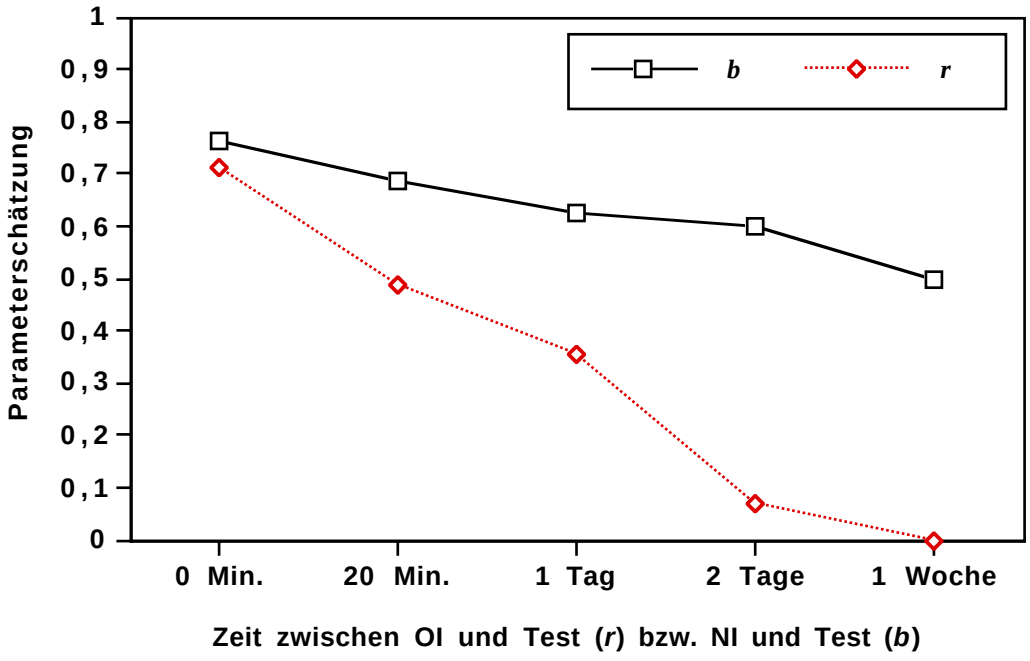

Abbildung 10.2:
ML-Parameterschätzungen für die Antworttendenzparameter b_j in Abhängigkeit vom Zeitintervall zwischen Nachinformation und Rekognitionstests und für die Erinnerungsparameter r_j in Abhängigkeit vom Retentionsintervall. Zugrundegelegt wurde das im Text beschriebene multinomiale Modell und die Daten von Experiment 3 aus Loftus, Miller und Burns (1978).

[103] Wegen des relativ kleinen Stichprobenumfanges von $N = 12 \cdot 2 \cdot 9 = 432$ sollte man bei der inferenzstatistischen Bewertung der Modellanpassungsstatistiken keinesfalls ein strengeres Typ-1-Fehlerrisiko als $\alpha = .05$ wählen. Liegen kleine Effekte vor ($w = .1$), ist bei $\alpha = .05$ und den hier gegebenen Freiheitsgraden $(1-\beta) > .34$; liegen moderate Effekte vor ($w = .2$), ist $(1-\beta) > .93$. Etwas günstiger sieht es aus, wenn man $\alpha = .10$ wählt: Für $w = .1$ ist dann $(1-\beta) > .47$, für $w = .2$ immerhin $(1-\beta) > .97$.

Ist dieses Ergebnis als Fortschritt gegenüber dem Modellierungsansatz von Wagenaar und Boer (1987) zu werten? Zunächst einmal muß kritisch eingestanden werden, daß die berichtete Reanalyse unter dem Aspekt der Teststärke eher weniger als mehr überzeugt. Wie Abbildung 10.3 deutlich macht, hat der für das Antworttendenzmodell kritische Test der Hypothese, daß keinerlei Gedächtnisbeeinträchtigungseffekte vorliegen (alle $m_j = 0$), selbst bei sehr großen Gedächtnisbeeinträchtigungseffekten noch keine voll überzeugende Teststärke. Dies liegt im wesentlichen an dem im Vergleich zur Wagenaar-Boer-Studie noch kleineren Gesamtstichprobenumfang, nicht so sehr an der Beziehung zwischen *m* und dem Cohenschen Effektstärkemaß *w*.

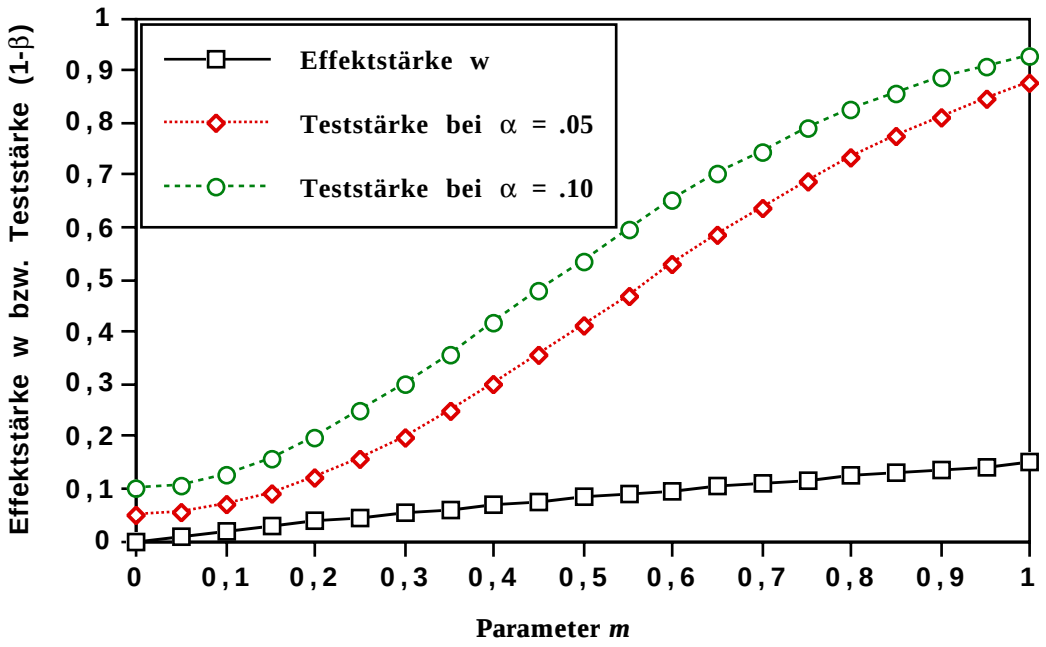

Abbildung 10.3:
Effektstärken *w* sowie Teststärken für $\alpha = .05$ und $\alpha = .10$ beim Differenzen-χ^2-Test des Antworttendenzmodells (Nullhypothese $m = 0$), wenn tatsächlich ein Gedächtnisbeeinträchtigungseffekt (mit $m > 0$) in Experiment 3 von Loftus et al. (1978) vorgelegen haben sollte.

Trotz des offenkundigen Teststärkeproblems wird der kritische Test im Falle von Experiment 3 aus Loftus et al. (1978) signifikant, jedenfalls wenn man das 10%-Signifikanzniveau zugrunde legt. Dies muß als Bestätigung für die Annahme von Elizabeth Loftus und Mitarbeitern gewertet werden, daß der 2AFC-Rekognitionstest besonders sensitiv für echte Gedächtnisbeeinträchtigungen (oder Abrufbeeinträchtigungen) durch nachträgliche Falschinformationen ist, d.h. sehr hohe *m*-Parameter produziert. Dies gilt auch dann, wenn diese Effekte — wie offenbar im hier betrachteten Fall — zusätzlich noch von Antworttendenzen im Sinne von McCloskey und Zaragoza (1985) überlagert werden.

Berücksichtigt man darüber hinaus, daß das vorgestellte Modell nicht nur gut auf die Daten von Loftus et al. (1978) paßt, sondern auch Parameterschätzungen liefert, die nahtlos zu den oben genannten Prognosen passen, so erscheint der vorgestellte Ansatz trotz des Teststärkeproblems doch so vielversprechend, daß er weiterverfolgt werden sollte. Bei Nachfolgeuntersuchungen muß jedoch unbedingt darauf geachtet werden, daß der Stichprobenumfang deutlich größer als in der hier reanalysierten Arbeit gewählt wird, damit die kritischen Modelltests bei kleineren Typ-1- und Typ-2-Fehlerwahrscheinlichkeiten durchgeführt werden können. Praktisch läßt sich dies z.B. durch Verwendung mehrerer kritischer Items pro Vp und anschließendes Aggregieren über Items und Personen erreichen. Auf die Vor- und Nachteile dieses Vorgehens wurde in Kapitel 5 bereits eingegangen.

Halten wir fest: Man kann ein einfaches MVB-Modell für den klassischen 2AFC-Rekognitionstest formulieren, das (a) identifizierbar ist, wenn es zu einem Mehr-Gruppen-Modell mit gruppenübergreifenden Parameterrestriktionen erweitert, (b) gut zu den Ergebnissen von Loftus et al. (1978, Experiment 3) paßt, (c) gedächtnispsychologisch sinnvolle Parameterschätzungen liefert und (d) tendenziell (bei relativ hoher statistischer Unsicherheit) Evidenz für einen Gedächtnis- oder Abrufbeeinträchtigungseffekt durch nachträgliche Falschinformation erbringt.

Läßt sich dieses Ergebnismuster auch für andere häufig benutzte Gedächtnistests zum Falschinformationseffekt erzielen? Im nächsten Abschnitt soll dies exemplarisch anhand des Quellendiskriminationstests untersucht werden.

10.2 Ein Modell zum Quellendiskriminationstest

Lindsay und Johnson (1989) haben erstmals einen Quellendiskriminationstests zur Untersuchung des Falschinformationseffekts eingesetzt. Personen einer Kontrollbedingung ohne nachträgliche Falschinformation und einer Versuchsbedingung mit nachträglicher Falschinformation zu einem Dia mußten für einer Reihe von Items angegeben, ob ein Item (a) nur in dem Dia enthalten war (OI-Urteil), (b) nur in der Nacherzählung erwähnt wurde (NI-Urteil), (c) sowohl im Dia als auch in der Nacherzählung erwähnt wurde (OI&NI-Urteil) oder (d) weder im Dia noch in der Nacherzählung erwähnt wurde (Neu-Urteil). Anders als bei Verwendung eines einfachen Ja-Nein-Rekognitionstests („War diese Information im Dia enthalten oder nicht?”) fanden die Autoren im Quellendiskriminationstest keine Hinweise auf Falschinformationseffekte. Falsche Quellenattribution kamen zwar durchaus vor, wurden in ihrer Häufigkeit jedoch nicht signifikant dadurch beeinflußt, ob in der anschließenden Nacherzählung falsche Informationen enthalten waren oder nicht. Auch Items, die nur in der Nacherzählung erwähnt wurden, im Bild jedoch gar nicht vorkamen, wurden in der Experimentalgruppe nicht häufiger auf das Bild attribuiert als in der Kontrollgruppe, für die diese Items reine

Distraktoritems waren. Die Autoren interpretierten dieses Ergebnis als Hinweis darauf, daß Falschinformationseffekte zum Verschwinden gebracht werden können, wenn man die Vpn durch Instruktionen und durch die Art des verwendeten Gedächtnistests anregt, die verschiedenen möglichen Informationsquellen sorgfältig auseinanderzuhalten: „Our results indicate that mislead subjects are capable of identifying the actual source of their memories of the misleading suggestions. (...) Thus one might conclude that eyewitness suggestibility effects reflect nothing other than demand characteristics and do not involve genuine memory-source confusions" (Lindsay & Johnson, 1989, S. 356).

Nachfolgeuntersuchungen zur Pilotstudie von Lindsay und Johnson (1989) haben gemischte Resultate erbracht. Viele Untersuchungen replizierten das Verschwinden des Falschinformationseffekts bei Verwendung von Quellendiskriminationstests an unterschiedlichen Stichproben (Kikuno, 1995; Multhaup, De Leonardis & Johnson, 1999; Schumacher-Bittner, 1993; Zaragoza & Koshmider, 1989; Zaragoza & Lane, 1994), andere Untersuchungen fanden signifikante Falschinformationseffekte auch bei Verwendung unterschiedlicher Quellendiskriminationstests (Belli, Lindsay, Gales & McCarthy, 1994; Higham, 1998; Lindsay, 1990; Manning & Loftus, 1996; Zaragoza & Lane, 1994; Zaragoza & Mitchell, 1996; Zaragoza & Muench, 1989). Falschinformationseffekte in Quellengedächtnistests scheinen durch folgende Faktoren begünstigt zu werden:

(a) ein langes Retentionsintervall zwischen Falschinformation und Gedächtnistest (Higham, 1998; Zaragoza & Muench, 1989; siehe jedoch Belli et al., 1994);
(b) oftmaliges Darbieten der Falschinformation (Zaragoza & Mitchell, 1996);
(c) Visualisierung der verbalen Falschinformation (Zaragoza & Muench, 1989);
(d) Induktion nachträglicher Falschinformation durch suggestive Fragen statt durch eine Nacherzählung mit falschen Details (Zaragoza & Lane, 1994, Experimente 1, 3 und 5);
(e) Induktion von Falschinformation durch eine zufällig angeordnete Nacherzählung, deren Inhalte zunächst in die richtige Reihenfolge zu bringen sind (Zaragoza & Lane, 1994, Experiment 2);
(f) Falschinformation, welche die Originalinformation lediglich ergänzt und ihr nicht direkt widerspricht (Zaragoza & Lane, 1994);
(g) Originalinformationen, die eher periphere Details der betrachteten Szene betreffen (Belli et al., 1994; Sporer & aus dem Kahmen, 1993, S. 12/13) und schließlich
(h) Originalinformation und Falschinformation, die zeitlich oder perzeptuell nur schwer diskriminierbar sind (Lindsay, 1990).

Kurz: Je schwächer die Gedächtnisspuren von Original- und Nachinformation, je ähnlicher sie einander sind und je ausgeprägter die gemeinsame Verarbeitung von

Original und Nachinformation, desto wahrscheinlicher werden Falschinformationseffekte auch bei Verwendung von Quellendiskriminationstests.

Kann aus dem berichteten Ergebnismuster der Schluß gezogen werden, daß zumindest unter den genannten förderlichen Bedingungen nachträgliche Falschinformationen das Gedächtnis für visuelle Originalinformationen beeinträchtigen? Dies ist aus zwei Gründen nicht der Fall. Der erste Grund ist die Unklarheit über die Bedeutung des Begriffs „Falschinformationseffekt im Quellendiskriminationstest". Dieser Begriff wird von verschiedenen Autoren in unterschiedlicher Weise verwendet. Während beispielsweise Lindsay und Johnson (1989), Kikuno (1995), Multhaup et al. (1999) sowie Zaragoza und Lane (1994) von einem Falschinformationseffekt sprechen, wenn nachträglich suggerierte Informationen im Gedächtnistest häufiger auf die Originalquelle attribuiert werden als das in einer Kontrollgruppe der Fall ist, die die betreffenden Informationen nur als Distraktoritems im Test bekamen, ohne sie zuvor in der Nacherzählung gelesen zu haben, schließen Manning und Loftus (1996) auf einen Falschinformationseffekt, wenn Fehlattributionen auf die Originalquelle bei nachträglicher verbaler Falschinformation häufiger vorkommen als bei Darbietung eines zweiten Bildes, das sich in einigen Details vom Originalbild unterscheidet. Lindsay (1990) wiederum verwendete zwei Kriterien für einen Falschinformationseffekt in Reproduktionstests: Experimentalitem-Kontrollitem-Unterschiede der Gedächtnisurteile zu nachträglichen Falschinformationen (falsche Attributionen auf die Originalquelle) und zu Originalinformationen (korrekte Reproduktionen der Originalquelle). Offensichtlich kann je nach verwendetem Kriterium ein anderer Schluß auf das Vorliegen oder Nichtvorliegen eines Falschinformationseffekts resultieren. Lindsay (1990) hat das auch ganz explizit gezeigt: Gute Diskriminierbarkeit von Nachinformation und Originalinformation eliminiert Experimentalitem-Kontrollitem-Unterschiede bezüglich falscher Attributionen auf die Originalquelle, nicht jedoch bezüglich korrekter Reproduktionen der Originalquelle. Die oben berichteten förderlichen und hinderlichen Einflüsse auf Falschinformationseffekte in Quellendiskriminationstests müssen also keine generell beobachtbaren Phänomene sein, sondern können wesentlich von dem jeweils verwendeten Kriterium für einen Falschinformationseffekt abhängen.

Ein zweites Problem bei der Interpretation der bislang vorliegenden Befunde ist die Nichtberücksichtigung von Antwort- bzw. Ratetendenzen. Es wird inzwischen allgemein akzeptiert, daß Urteile in Quellengedächtnistests nicht nur vom Item- und Quellengedächtnis, sondern auch von Ratetendenzen bei Item- oder Quellengedächtnisunsicherheit abhängen (vgl. Abschnitt 7.1.2 und Kapitel 8 der vorliegenden Arbeit). Diese Annahmen müssen natürlich auch für Quellengedächtnistests zum Effekt nachträglicher Falschinformation gelten (siehe etwa Lindsay, 1990). Selbst wenn also zweifelsfrei feststünde, daß unter bestimmten Bedin-

gungen nach einem wohldefinierten Kriterium ein Falschinformationseffekt in Quellengedächtnistests vorliegt, so hieße das noch nicht, daß Item- oder Quellengedächtnisbeeinträchtigungen durch nachträgliche Falschinformation hierfür die Ursache sein müssen. Der Effekt kann unter Umständen ausschließlich auf Unterschieden in den Ratetendenzen beruhen. Umgekehrt kann auch das Ausbleiben eines Falschinformationseffekts nach einem der o.g. Kriterien nicht als zwingende Evidenz gegen einen Gedächtnisbeeinträchtigungseffekt interpretiert werden, weil (a) das gewählte Kriterium für einen Falschinformationseffekt möglicherweise gar nicht sensitiv für genuine Gedächtnisbeeinträchtigungseffekte ist oder (b) Gedächtnisbeeinträchtigungseffekte und Rateeffekte eventuell antagonistisch wirken und sich in ihrer Wirkung auf korrekte bzw. falsche Quellenurteile genau aufheben.

Wie immer bei Dekomponierungsproblemen diese Art gibt es zwei mögliche Auswege (vgl. Abschnitt 2.6): (a) den designbasierten Ansatz, der im gegebenen Fall auf modifizierte Quellentests hinausläuft, welche Einflüsse von Ratetendenzen minimieren (vgl. Lindsay, 1990; Zaragoza & Lane, 1994, Experiment 5, Zaragoza & Mitchell, 1996; Higham, 1998), oder (b) den modellbasierten Ansatz, welcher Gedächtnis- und Rateeinflüsse auf Quellenurteile explizit berücksichtigt, so daß Gedächtnis- und Rateeinflüsse separat schätzbar werden. In diesem Abschnitt soll ein solcher modellbasierter Ansatz vorgestellt und in den folgenden Abschnitten dann evaluiert werden. Er ist im Kern eine Modifikation und zugleich eine Erweiterung des bereits geschilderten Ansatzes von Batchelder et al. (1990) für ein komplexeres und empirisch ergiebigeres Versuchsdesign.

Ausgangspunkt bei der Entwicklung des Modells war die Suche nach einem Versuchsdesign, das diverse Falschinformationseffekte nach möglichst allen oben genannten Kriterien erfassen und abbilden kann. Das in Tabelle 10.2 dargestellte Design, welches insgesamt neun Versuchsbedingungen umfaßt — fünf Klassen alter Items und vier Klassen neuer Items —, kommt diesem Ziel sehr nahe. Jede Versuchsbedingung entspricht einer bestimmten Klasse von Testitems im Quellendiskriminationstest, die jeweils durch eine bestimmte Beziehung zur Originalinformation (OI) und zur Nachinformation (NI) ausgezeichnet ist. Betrachten wir ein konkretes Beispiel. Die Vpn sehen zunächst ein bestimmtes Dia, das ein Studentenzimmer mit bestimmten Objekten darin darstellt (z.B. mit einer Coca-Cola-Dose, einer Zigarettenschachtel etc.). Alle für die Testphase wesentlichen Objekte sind vollständig abgebildet, deutlich erkennbar und im Zentrum des Bildes lokalisiert. Im Anschluß an eine Distraktortätigkeit bekommen die Vpn einen verbale Schilderung des Zimmers, die teilweise mit den Dia dargebotenen Objekten konsistent ist, teilweise widersprüchliche Objektbeschreibungen liefert (z.B. eine Fanta-Dose anstelle einer Cola-Dose), teilweise Objekte der Originalszene ausläßt und

teilweise neue Objekte hinzuerfindet. In der Testphase können dann neun Klassen von Testitems unterschieden werden, die den Zeilen der Tabelle 10.2 entsprechen.

Tabelle 10.2:
Versuchsdesign zur Untersuchung verschiedener Falschinformationseffekte mit dem Quellendiskriminationstest.

Versuchsbedingung	Das Testitem ist	Die OI hierzu war	Die NI hierzu war
A1	alt (Quelle: OI & NI)	konsistent	konsistent
A2	alt (Quelle: OI)	konsistent	neutral
A3	alt (Quelle: OI)	konsistent	widersprüchlich
A4	alt (Quelle: NI)	neutral	konsistent
A5	alt (Quelle: NI)	widersprüchlich	konsistent
N1	neu (Distraktor)	neutral	neutral
N2	neu (Distraktor)	neutral	widersprüchlich
N3	neu (Distraktor)	widersprüchlich	neutral
N4	neu (Distraktor)	widersprüchlich	widersprüchlich

Itemklasse A1 entspräche einem Item (z.B. der Cola-Dose), das in der OI vorhanden war und auch in der NI explizit benannt wurde. Dies ist der einfache Fall konsistenter Nachinformationen. Die Itemklassen A2 und A3 betreffen ebenfalls Testitems, die in der OI vorhanden waren, die jedoch in der NI unerwähnt blieben (A2) bzw. in widersprüchlicher Form erwähnt wurden (A3). Wenn also beispielsweise in der OI eine Coca-Cola-Dose vorne rechts auf dem Schreibtisch zu sehen war, dann blieb in der NI der Itemklasse A2 unerwähnt, was sich vorne rechts auf dem Schreibtisch befand, während in der NI zur Itemklasse A3 z.B. fälschlich behauptet wurde, vorne rechts auf dem Schreibtisch habe sich eine Fanta-Dose befunden. Itemklasse A3 entspricht also dem klassischen Fall nachträglicher Falschinformation, wie er bereits in den Experimenten von Loftus et al. (1978), McCloskey und Zaragoza (1985) und vielen anderen untersucht wurde. Itemklasse A2 ist eine mögliche Kontrollbedingung hierfür, ebenso Klasse A1 mit konsistenter NI.

Die Itemklassen A4 und A5 entsprechen Testitems, die zwar in der NI explizit erwähnt wurden, in der OI aber entweder gar nicht vorkamen (A4) oder in widersprüchlicher Form vorkamen. Beispielsweise könnte nach einer Fanta-Dose gefragt werden, die in der NI erwähnt wurde, während in der OI keine Getränkedose vorhanden war (A4) oder aber lediglich eine Coca-Cola-Dose zu sehen war (A5). Auch die Itemklasse A5 entspricht einem Widerspruch zwischen OI und NI, wobei hier jedoch das NI-Item dem Testitem entspricht. Man kann hier also von „vorangestellter Falschinformation" sprechen. Die Versuchsbedingungen A4 und A1 wären mögliche Kontrollbedingungen für einen „umgekehrten Falschinforma-

tionseffekt", bei dem es um Beeinträchtigungen der Gedächtnisurteile zur NI durch inkonsistente OI-Informationen geht.

Die Itemklassen N1 bis N4 betreffen ausschließlich Distraktoritems, die weder in der OI noch in der NI vorgekommen sind. Die Itemklasse N4 entspricht neuen Items, zu denen in der OI und in der NI ähnliche Informationen vorhanden waren. Das wäre z.B. der Fall, wenn in der OI eine Cola-Dose vorhanden war, die in der NI korrekt erwähnt wurde. Gefragt wird jedoch nach dem Distraktoritem „Fanta-Dose", das weder in der OI noch in der NI vorkam. Die Itemklassen K2 und K3 entsprechen ebenfalls Distraktoritems, zu denen nur in der OI (K2) bzw. nur in der NI (K3) ähnliche Informationen enthalten waren. Itemklasse K1 schließlich umfaßt Distraktoritems, die weder einen Bezug zur OI noch zur NI aufweisen. Wäre „Fanta-Dose" das entsprechende Testitem, dann dürfte also in Bedingung K1 eine Getränkedose weder in der OI noch in der NI enthalten oder erwähnt sein.

Man kann leicht erkennen, das sich sehr viele unterschiedliche „Falschinformationseffekte" anhand des geschilderten Designs definieren lassen. Die wohl am häufigsten benutzte Definition geht auf Lindsay und Johnson (1989) zurück und entspricht dem Unterschied falscher OI-Attributionen zwischen einer der beiden Itemklassen A4 oder A5 auf der einen und einer der Distraktorklassen N1 bis N4 auf der anderen Seite. Welcher Kontrast im Einzelfall genau gemeint ist, kann erst nach genauer Inspektion des Versuchsmaterials entschieden werden. In der Originalstudie von Lindsay und Johnson (1989) wurde eine A4 analoge Bedingung gegen eine N1 analoge Bedingung kontrastiert. Genauer: Wenn bei A4-Items wesentlich häufiger fälschliche Quellenattributionen auf die OI erfolgen als bei der Distraktorklasse N1, würden Lindsay und Johnson (1989) von einem Falschinformationseffekt oder „Suggestibilitätseffekt" sprechen. Das von Lindsay (1990) zusätzlich benutzte Kriterium entspricht demgegenüber dem Vergleich korrekter Quellenattributionen in den Bedingungen A2 und A3: Mehr korrekte Quellenurteile in der Bedingung ohne inkonsistente Nachinformation (A2) im Vergleich zur Bedingung mit inkonsistenter Nachinformation (A3) entsprechen einem Falschinformationseffekt im Sinne von Lindsay (1990). Denkbar ist natürlich auch ein Kontrast zwischen A3 und A1, was man einen Falschinformationseffekt dritter Art nennen könnte. Auf analoge Weise lassen sich „Falschinformationseffekte" zur NI definieren, indem man A5 gegen A4 oder A5 gegen A1 kontrastiert (Falschinformationseffekte 4. und 5. Art). Hierbei geht es um die bereits erwähnten „umgekehrten Falschinformationseffekte". Schließlich kann man jede der drei Distraktorbedingungen N2, N3 oder N4 gegen die Distraktorbedingung N1 kontrastieren (Falschinformationseffekte 6. - 8. Art). Hierbei ist von Interesse, ob bzw. wie stark falsche Alarme durch die Ähnlichkeit eines Distraktors zu einer OI oder NI beeinflußt werden.

Sind die unterschiedlichen Falschinformationseffekte 1. bis 8. Art sensitiv für echte Gedächtnisbeeinträchtigungen durch OI-NI-Inkonsistenzen? Diese Frage läßt sich erst beantworten, wenn man sich auf ein Modell bezieht, aus dem hervorgeht, wie Gedächtnis- und Urteilsprozesse die verschiedenen Quellenurteile zu den neun Itemklassen beeinflussen.

Ein solches Modell, das in den folgenden Abschnitten diskutiert und evaluiert werden soll, ist in den Abbildungen 10.4a) bis 10.4d) in Form binärer Ereignisbaumdiagramme dargestellt. Das Modell beinhaltet insgesamt sechs Gedächtnisparameter — drei zur OI und drei zur NI — sowie sechs Urteils- bzw. Rateparameter. Die Parameter sollen zunächst kurz charakterisiert werden, ehe auf ihr Zusammenwirken in den verschiedenen Versuchsbedingungen näher eingegangen wird.

Die Gedächtnisparameter zur Originalinformation entsprechen folgenden drei Wahrscheinlichkeiten:

$r(\mathrm{OI})_k$: die Wahrscheinlichkeit, ein OI-Item als Teil der OI zu erinnern, wenn die NI-Information hierzu konsistent ist;

$r(\mathrm{OI})_n$: die Wahrscheinlichkeit, ein OI-Item als Teil der OI zu erinnern, wenn die NI hierzu keine Information liefert;

$r(\mathrm{OI})_w$: die Wahrscheinlichkeit, ein OI-Item als Teil der OI zu erinnern, wenn die NI hierzu widersprüchliche Information liefert.

Gedächtnisbeeinträchtigungseffekte durch nachträgliche Nachinformation können also in der Form $r(\mathrm{OI})_w < r(\mathrm{OI})_n$ oder in der Form $r(\mathrm{OI})_w < r(\mathrm{OI})_k$ in Erscheinung treten.

In analoger Weise kann man drei Gedächtnisparameter zur Nachinformation definieren:

$r(\mathrm{NI})_k$: die Wahrscheinlichkeit, ein NI-Item als Teil der NI zu erinnern, wenn die OI-Information hierzu konsistent ist;

$r(\mathrm{NI})_n$: die Wahrscheinlichkeit, ein NI-Item als Teil der NI zu erinnern, wenn die OI hierzu keine Information liefert;

$r(\mathrm{NI})_w$: die Wahrscheinlichkeit, ein NI-Item als Teil der NI zu erinnern, wenn die OI hierzu widersprüchliche Information liefert.

Gedächtnisbeeinträchtigungen bezüglich der Nachinformation können ebenfalls auf zweierlei Weise vorliegen, nämlich $r(\mathrm{NI})_w < r(\mathrm{NI})_n$ oder $r(\mathrm{NI})_w < r(\mathrm{NI})_k$.

Neben den sechs Gedächtnisparametern gibt es zwei Urteils- und Rateparameter zu OI-Quellenzuschreibungen:

$g(\mathrm{OI})_{nw}$: die Wahrscheinlichkeit, im Zustand der Gedächtnisunsicherheit ein Item der OI-Quelle zuzuschreiben, wenn das betreffende Testitem zur OI nicht im Widerspruch steht;

$g(\mathrm{OI})_w$: die Wahrscheinlichkeit, im Zustand der Gedächtnisunsicherheit ein Item der OI-Quelle zuzuschreiben, wenn das betreffende Testitem zur OI im Widerspruch steht.

Der zuletzt genannte Fall entspricht beispielsweise dem Testitem „Fanta-Dose", wenn in der OI eine Cola-Dose enthalten war. Der erstgenannte Fall liegt vor, wenn in der OI entweder gar keine Getränkedose oder aber eine Fanta-Dose enthalten war.

Prinzipiell lassen sich analoge Ratewahrscheinlichkeiten auch für NI-Quellenzuschreibungen definieren. Da jedoch denkbar ist, daß OI-Zuschreibungen und NI-Zuschreibungen korreliert sind — beispielsweise aufgrund einer generellen „Ja"-Sage-Bereitschaft —, sollen NI-Ratewahrscheinlichkeiten für den Fall einer OI-Zuschreibung und für den Fall keiner OI-Zuschreibung durch einen zusätzlichen Index (1 bzw. 2) unterschieden werden:

$g_1(\mathrm{NI})_{\mathrm{nw}}$: die Wahrscheinlichkeit, im Zustand der Gedächtnisunsicherheit ein Item der NI-Quelle zuzuschreiben, wenn (a) das betreffende Testitem zur NI nicht im Widerspruch steht und (b) bereits eine OI-Zuschreibung vorgenommen wurde;

$g_1(\mathrm{NI})_{\mathrm{w}}$: die Wahrscheinlichkeit, im Zustand der Gedächtnisunsicherheit ein Item der NI-Quelle zuzuschreiben, wenn (a) das betreffende Testitem zur NI im Widerspruch steht und (b) bereits eine OI-Zuschreibung vorgenommen wurde;

$g_2(\mathrm{NI})_{\mathrm{nw}}$: die Wahrscheinlichkeit, im Zustand der Gedächtnisunsicherheit ein Item der NI-Quelle zuzuschreiben, wenn (a) das betreffende Testitem zur NI nicht im Widerspruch steht und (b) zuvor keine OI-Zuschreibung vorgenommen wurde;

$g_2(\mathrm{NI})_{\mathrm{w}}$: die Wahrscheinlichkeit, im Zustand der Gedächtnisunsicherheit ein Item der NI-Quelle zuzuschreiben, wenn (a) das betreffende Testitem zur NI im Widerspruch steht und (b) zuvor keine OI-Zuschreibung vorgenommen wurde.

Unkorrelierte Ratewahrscheinlichkeiten bezüglich beider Quellen kann man durch die Restriktionen $g_1(\mathrm{NI})_{\mathrm{nw}} = g_2(\mathrm{NI})_{\mathrm{nw}}$ und $g_1(\mathrm{NI})_{\mathrm{w}} = g_2(\mathrm{NI})_{\mathrm{w}}$ aus dem Rahmenmodell generieren. Die weitergehende Restriktion $g_1(\mathrm{NI})_{\mathrm{nw}} = g_2(\mathrm{NI})_{\mathrm{nw}} = g_1(\mathrm{NI})_{\mathrm{w}} = g_2(\mathrm{NI})_{\mathrm{w}}$ entspricht der Hypothese, daß Rate- bzw. Urteilsprozesse durch die Beziehung eines Testitems zur faktischen NI nicht beeinflußt wird. Analoges gilt im Falle der Gültigkeit der Restriktion $g(\mathrm{OI})_{\mathrm{nw}} = g(\mathrm{OI})_{\mathrm{w}}$ in bezug auf die OI.

Abbildung 10.4a) illustriert die kognitiven Prozesse, die zu den vier möglichen Quellenurteilen führen können, wenn ein altes Item vorliegt, das in OI und NI konsistent dargeboten wurde (Bedingung A1). Das Item kann zunächst als Bestandteil der OI erinnert werden (mit Wahrscheinlichkeit $r(\mathrm{OI})_{\mathrm{k}}$) oder nicht erinnert werden (mit der komplementären Wahrscheinlichkeit $1 - r(\mathrm{OI})_{\mathrm{k}}$). Dies entspricht der ersten Aufgabelung im Ereignisbaumdiagramm. Unabhängig vom Ergebnis des OI-Erinnerungsprozesses kann das Item zusätzlich als Bestandteil der NI erinnert werden (mit Wahrscheinlichkeit $r(\mathrm{NI})_{\mathrm{k}}$) oder nicht erinnert werden (mit

Wahrscheinlichkeit $1 - r(\mathrm{NI})_k$). Dies entspricht der zweiten Ebene des Ereignisbaumdiagramms. Es wird angenommen, daß der Abruf der NI und der OI stochastisch unabhängig erfolgt, d.h. die Erinnerungswahrscheinlichkeit der NI ist immer gleich groß, egal ob die OI bereits erinnert wurde oder nicht. Diese Unabhängigkeitsannahme ist natürlich nicht unproblematisch. Da sich ganz ähnliche Annahmen im Kontext des Dual-Code-Markoff-Modells verbalen Lernens (vgl. Annahme (5) in Abschnitt 6.5.1) und im Kontext verschiedener Speicher-Abruf-Modelle bereits bewährt haben (vgl. Abschnitt 7.1.4), erscheint die Unabhängigkeitsannahme jedoch vertretbar.

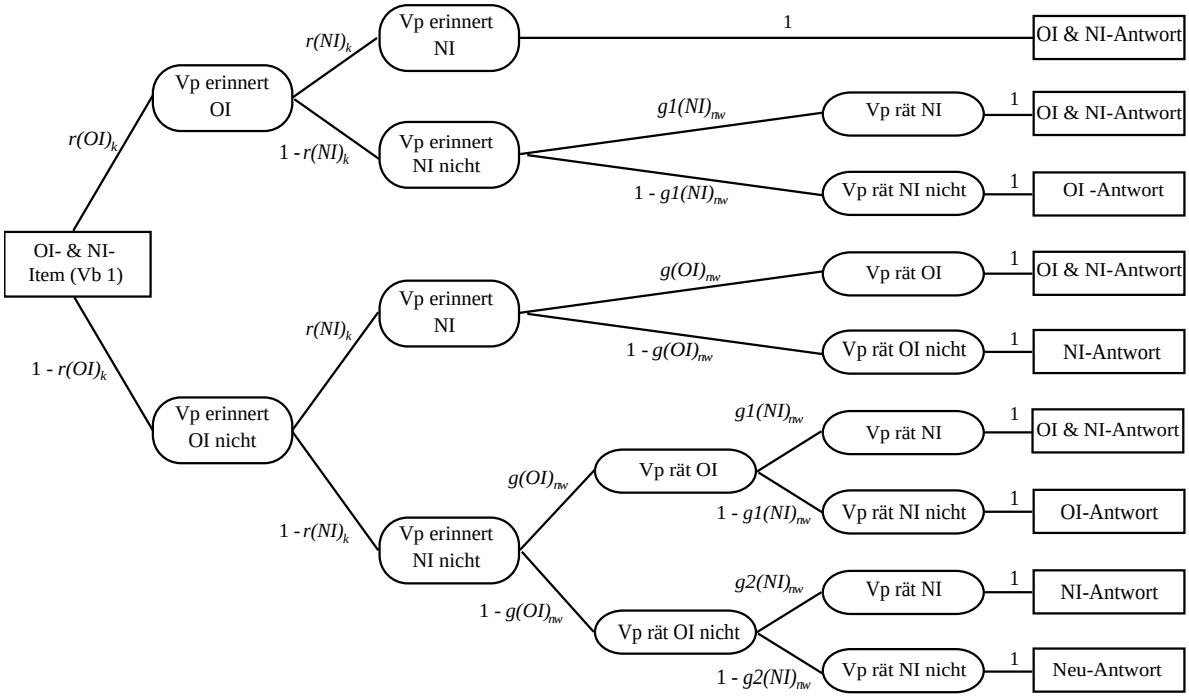

Abbildung 10.4a):
Binäres Ereignisbaumdiagramm zum Falschinformationsmodell für Quellendiskriminationstests für Testitems, die sowohl in der Originalinformation als auch in er Nachinformation konsistent dargeboten wurden. Die Erläuterung der Parameterbedeutungen erfolgt im Text.

Wenn beide Gedächtnisprozesse erfolgreich verlaufen, wird konsequenterweise geantwortet, daß das Item sowohl in der OI als auch in der NI vorhanden war. Scheitert einer der beiden Gedächtnisprozesse oder gar beide, so wird die entsprechenden Quelleninformation mit Wahrscheinlichkeit $g(\mathrm{OI})_{\mathrm{nw}}$ (bei Scheitern des OI-Abrufprozesses) bzw. $g(\mathrm{NI})_{\mathrm{nw}}$ (bei Scheitern des NI-Abrufprozesses) geraten. Hierbei kann ein korrektes oder ein falsches Quellenurteil resultieren — je nach Ergebnis des Rateprozesses (siehe dazu Abbildung 10.4a).

Die Modellgleichungen für den Ereignisbaum in Abbildung 10.4a) lassen sich in der üblichen Weise durch Multiplikation der Parameter pro Ast des Baumes und Aufsummieren aller Äste, die in dieselbe Beobachtungskategorie ein-

münden, ganz einfach herleiten. Um Platz zu sparen, wird auf die explizite Aufführung der Modellgleichungen hier verzichtet. Das System der Modellgleichungen für alle Ereignisbäume des Gesamtmodells kann im übrigen aus der Modellbeschreibungsdatei „FI-MODEL.EQN" in Abschnitt 22) des Anhangs entnommen werden.

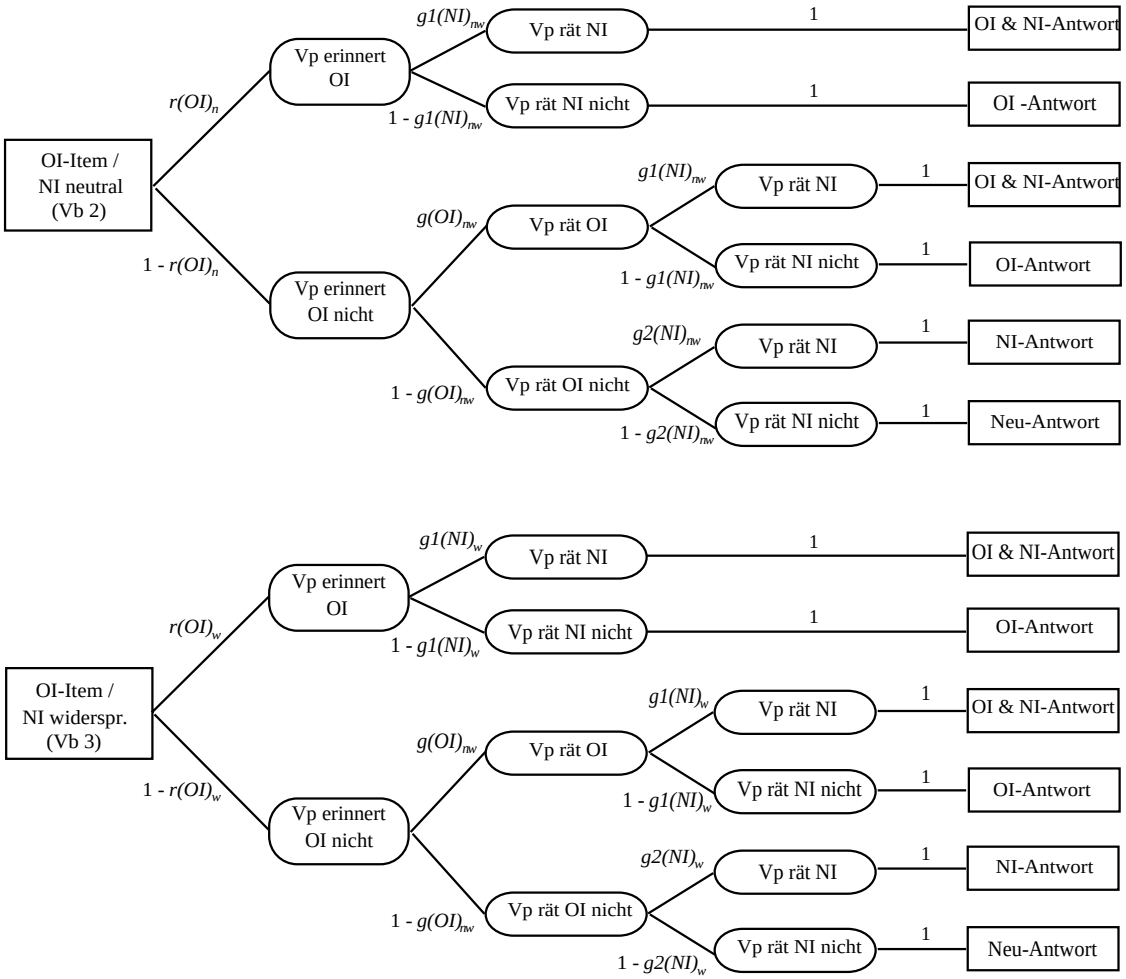

Abbildung 10.4b):
Binäres Ereignisbaumdiagramm zum Falschinformationsmodell für Quellendiskriminationstests für Testitems, die in der Originalinformation enthalten waren, während in der Nachinformation diesbezüglich keine Information (obere Abbildung) bzw. widersprüchliche Information enthalten war (untere Abbildung). Die Erläuterung der Parameterbedeutungen erfolgt im Text.

Die Ereignisbäume für die anderen acht Itemklassen, auf die sich das Falschinformationsmodell bezieht, folgen immer dem gleichen Prinzip, so daß die Beschreibung hier sehr knapp erfolgen kann. Grundsätzlich wird bei alten Items immer erst ein erfolgreicher oder nicht erfolgreicher Gedächtnisprozeß postuliert, auf den Rateprozesse bezüglich derjenigen Quelle bzw. den Quellen folgen, für die der Gedächtnisprozeß keine Ergebnisinformation liefert. Abbildung 10.4b) illustriert dies für Testitems, die nur in der OI enthalten waren, Abbildung 10.4c) für Testitems, die nur in der NI enthalten waren. In allen vier Ereignisbäumen kommt anders als in Abbildung 10.4a) immer nur ein Gedächtnisprozeß vor, da

das betreffende Testitem in der jeweils anderen Quelle nicht dargeboten wurde und folglich auch nicht erinnert werden kann. Zu beachten ist darüber hinaus, daß die Gedächtnis- und Urteilsparameter nun anders als in Abbildung 10.4a) indiziert sind, um möglichen Informationsinkonsistenz-Einflüssen auf Gedächtnis- und Urteilsprozesse Rechnung zu tragen.

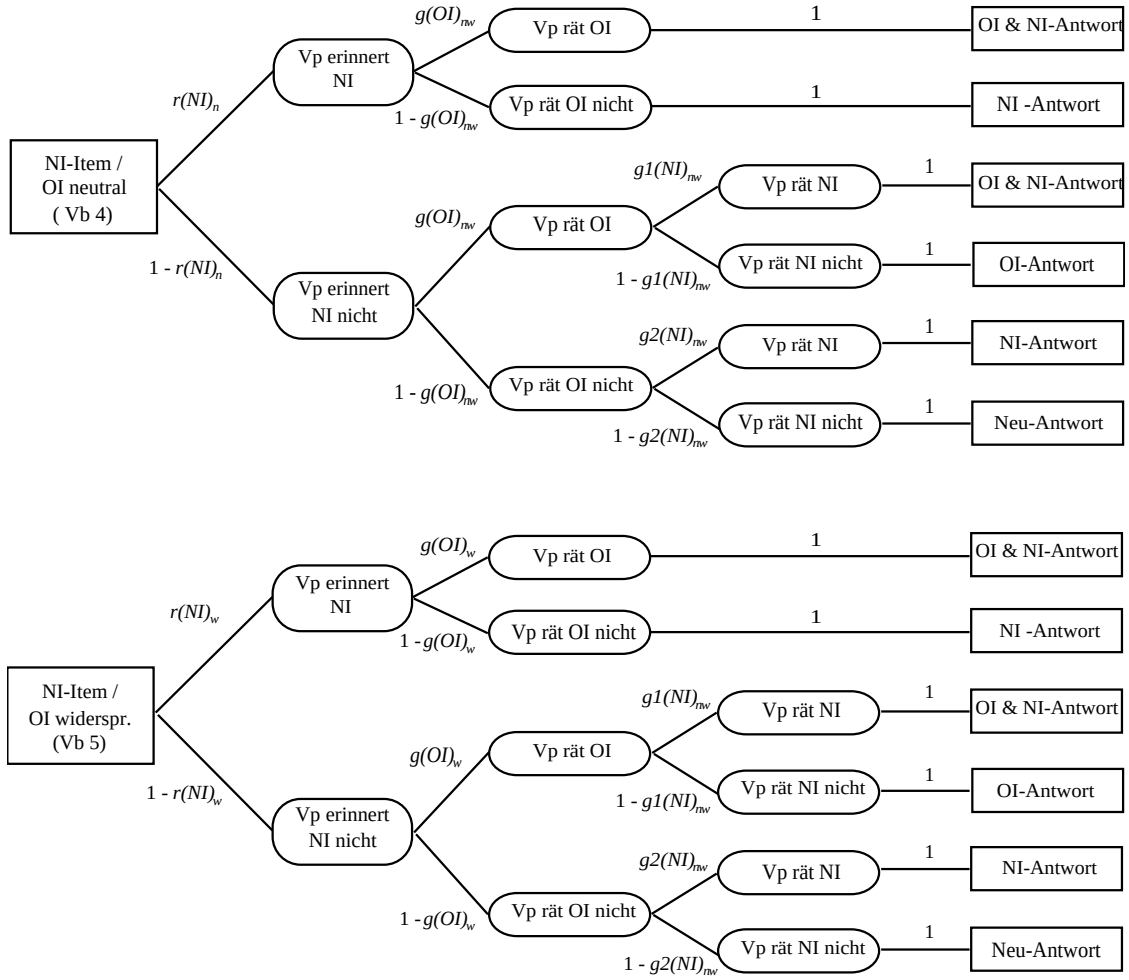

Abbildung 10.4c):
Binäres Ereignisbaumdiagramm zum Falschinformationsmodell für Quellendiskriminationstests für Testitems, die in der Nachinformation enthalten waren, während in der Originalinformation diesbezüglich keine Information (obere Abbildung) bzw. widersprüchliche Information enthalten war (untere Abbildung). Die Erläuterung der Parameterbedeutungen erfolgt im Text.

Abbildung 10.4d) beinhaltet die Ereignisbaumdiagramme für die vier Klassen von Distraktoritems. Anders als in 2HT-Modellen wird hier kein Distraktorentdeckungsmechanismus angenommen. Alle Parameter sind folglich Urteils- bzw. Rateparameter. Grundsätzlich müssen zwei Rateprozesse bezüglich der OI und der NI vorgesehen werden, die aber anders als die Gedächtnisprozesse zur OI und NI nicht von vornherein als stochastisch unabhängig angenommen werden. Deshalb werden vier Rateparameter zur NI vorgesehen. Die Indizierung der Parameter richtet sich danach, ob das betreffende Testitem zu einer Information aus der OI, der NI oder zu beidem im Widerspruch steht oder nicht.

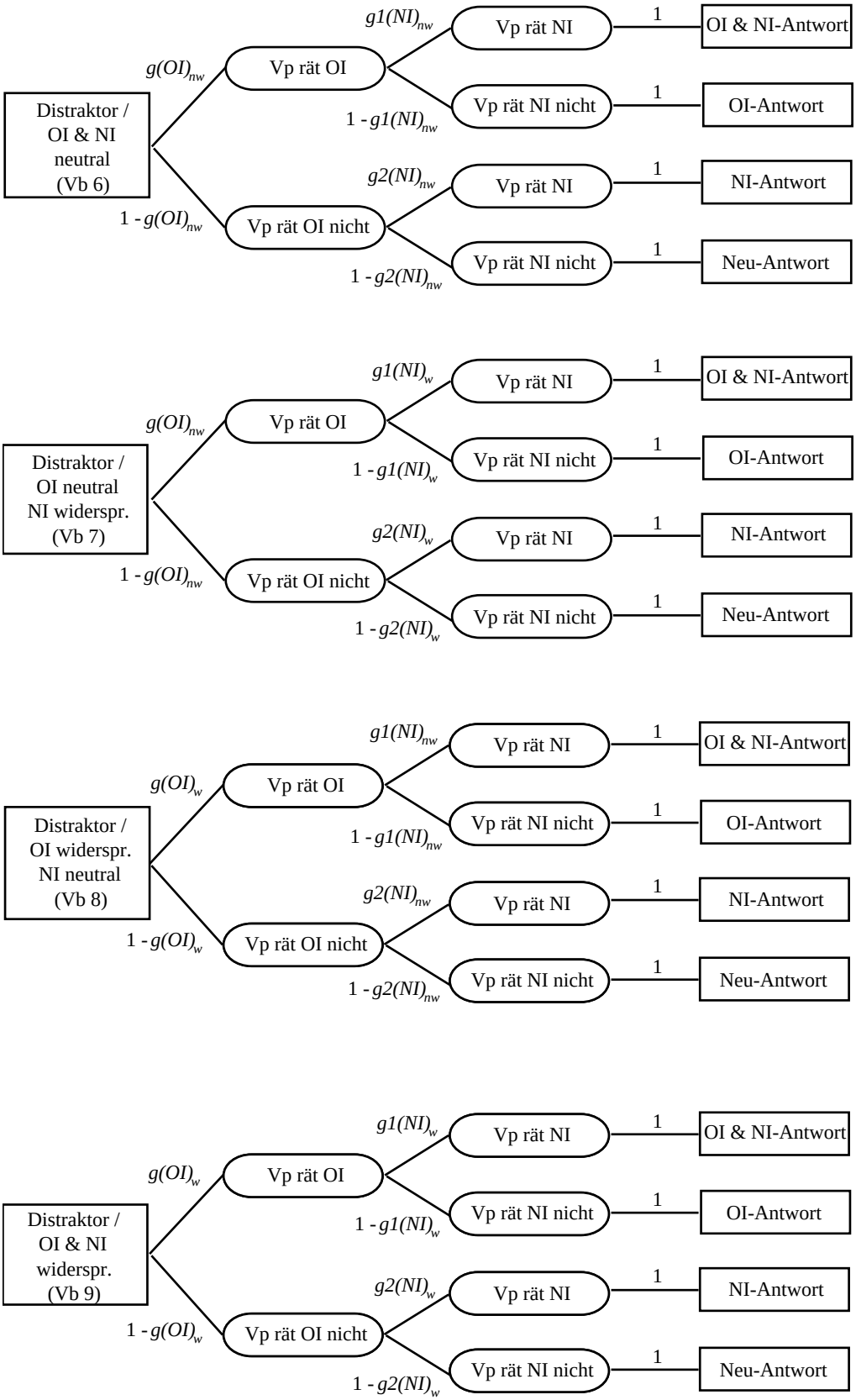

Abbildung 10.4d):
Binäres Ereignisbaumdiagramm zum Falschinformationsmodell für verschiedene Klassen von Distraktoritems. In der oberen Abbildung sind Distraktoritems dargestellt, die keinen Bezug zu OI- oder NI-Informationen aufweisen. Die folgenden Abbildungen betreffen Distraktoritems, zu denen in der NI, in der OI bzw. in OI und NI widersprüchliche Informationen enthalten waren. Die Erläuterung der Parameterbedeutungen erfolgt im Text.

Auf dem Hintergrund des dargestellten Modells kann nun die Frage beantwortet werden, ob die verschiedenen Typen von Falschinformationseffekten, die sich für das komplexe Versuchsdesign definieren lassen, für Gedächtnisbeeinträchtigungseffekte sensitiv sind oder nicht. Betrachten wir zunächst das wohl am häufigsten benutzte Kriterium falscher Quellenattributionen auf die OI in der Suggestionsbedingung A4 verglichen mit der Distraktorbedingung N1 (vgl. Lindsay & Johnson, 1989). Die Wahrscheinlichkeit, in Bedingung A4 entweder ein OI-Urteil oder ein OI&NI-Urteil zu geben, entspricht der Summe der Wahrscheinlichkeiten des ersten, des dritten und des viertobersten Astes im oberen Ereignisbaum der Abbildung 10.4c). Diese Summe ist gerade $r(\mathrm{NI})_\mathrm{n} \cdot g(\mathrm{OI})_\mathrm{nw} + (1 - r(\mathrm{NI})_\mathrm{n}) \cdot g(\mathrm{OI})_\mathrm{nw} = g(\mathrm{OI})_\mathrm{nw}$, d.h. identisch mit der Wahrscheinlichkeit, im Zustand der Unsicherheit die OI-Quelle zu raten, sofern dort keine widersprüchliche Information präsentiert wurde. Ein Blick auf den obersten Ereignisbaum in Abbildung 10.4d) zeigt sofort, daß die Wahrscheinlichkeit eines OI- oder OI&NI-Urteils für die Itemklasse N1 ebenfalls genau $g(\mathrm{OI})_\mathrm{nw}$ ist. Mit anderen Worten: Der Kontrast falscher OI-Attribution zwischen A4 und N1 ist *weder* für Gedächtnisbeeinträchtigungseffekte *noch* für Rateeffekte sensitiv. Gedächtnisparameter gehen in den Kontrast gar nicht ein, und die involvierten Ratewahrscheinlichkeiten müssen gleich sein, wenn das Modell aus Abbildung 10.4 gilt. Genau diese Prognose haben Lindsay und Johnson (1989) und die meisten Replikationsuntersuchungen zu ihrer Studie auch bestätigt.

Auf analoge Weise kann man zeigen, daß auch die Kontraste falscher OI-Attribution zwischen A4 und den Distraktorklassen N2 bis N4 sowie zwischen A5 und den vier Distraktorklassen nicht sensitiv für Gedächtnisbeeinträchtigungseffekte sind, d.h. ausschließlich von Rateparametern abhängen. Allerdings müssen die Ratewahrscheinlichkeiten — je nach betrachtetem Kontrast — nicht mehr notwendig gleich sein. Die Wahrscheinlichkeit falscher OI-Attributionen in A5 ist beispielsweise $g(\mathrm{OI})_\mathrm{w}$, die in N1 oder N2 ist dagegen $g(\mathrm{OI})_\mathrm{nw}$. Wenn nun die Ratewahrscheinlichkeit $g(\mathrm{OI})_\mathrm{w}$ etwas größer als $g(\mathrm{OI})_\mathrm{nw}$ ist — eine plausible Annahme, wie die folgenden Abschnitte zeigen werden —, dann ist ein Falschinformations- bzw. Suggestibilitätseffekt die zwingende Konsequenz. Auch dieses Ergebnismuster wurde von einigen Autoren gefunden (s.o.). Unsere modellbasierte Betrachtung zeigt, daß dieser Effekt gar nichts mit Gedächtnisbeeinträchtigungen zu tun haben muß, sondern ausschließlich aufgrund bedingungsspezifisch unterschiedlicher Ratewahrscheinlichkeiten resultieren kann.

Wie sieht es mit dem zweiten Kriterium von Lindsay (1990) aus, d.h. mit dem Vergleich korrekter OI-Urteile zwischen den Bedingungen A2 und A3? Aus Abbildung 10.4b) kann man die Wahrscheinlichkeiten $p_{\mathrm{A2}}(\text{„OI”})$ für ein OI-Urteil in A2 und $p_{\mathrm{A3}}(\text{„OI”})$ für ein OI-Urteil in A3 wie folgt herleiten:

$$p_{A2}(\text{„OI”}) = \left(r(\text{OI})_{\text{n}} + \left(1 - r(\text{OI})_{\text{n}}\right) \cdot g(\text{OI})_{\text{nw}}\right) \cdot \left(1 - g_1(\text{NI})_{\text{nw}}\right)$$
$$p_{A3}(\text{„OI”}) = \left(r(\text{OI})_{\text{w}} + \left(1 - r(\text{OI})_{\text{w}}\right) \cdot g(\text{OI})_{\text{nw}}\right) \cdot \left(1 - g_1(\text{NI})_{\text{w}}\right) \qquad (10.2\text{-}1)$$

Offensichtlich *kann* ein Unterschied zwischen beiden Bedingungen hier wirklich auf einem Gedächtnisbeeinträchtigungseffekt beruhen, z.B. dann, wenn alle Rateparameter sich zwischen A2 und A3 nicht unterscheiden, jedoch $r(\text{OI})_{\text{n}} > r(\text{OI})_{\text{w}}$ gilt. Auf der anderen Seite *muß* ein Unterschied nicht zwangsläufig auf Gedächtnisbeeinträchtigungseffekten beruhen, er kann auch allein durch unterschiedliche Ratetendenzen bei widersprüchlicher ($g_1(\text{NI})_{\text{w}}$) und nicht widersprüchlicher Information ($g_1(\text{NI})_{\text{nw}}$) in der Nachinformation bedingt sein. Schließlich — und das scheint mir fast noch bedeutsamer — kann auch aus der Gleichheit $p_{\text{A2}}(\text{„OI”}) = p_{\text{A3}}(\text{„OI”})$ kein psychologisch informativer Schluß gezogen werden. Auch wenn ein Gedächtnisbeeinträchtigungseffekt $r(\text{OI})_{\text{n}} > r(\text{OI})_{\text{w}}$ faktisch vorliegt, kann dennoch ein Nulleffekt $p_{\text{A2}}(\text{„OI”}) = p_{\text{A3}}(\text{„OI”})$ resultieren, wenn sich nämlich die Ratewahrscheinlichkeiten bezüglich der NI gerade gegenläufig verhalten, d.h. $g_1(\text{NI})_{\text{nw}} > g_1(\text{NI})_{\text{w}}$ gilt. Obwohl unwahrscheinlich erscheinen mag, daß sich beide Effekte gegenseitig exakt aufheben, ist dies doch immerhin möglich. Dies beweist, daß auch ein Nulleffekt der Art $p_{\text{A2}}(\text{„OI”}) = p_{\text{A3}}(\text{„OI”})$ nicht eindeutig interpretierbar ist, d.h. er schließt Gedächtnisbeeinträchtigungseffekte nicht zwingend aus.

Ohne weitere mögliche Definitionen für einen Falschinformationseffekt näher zu betrachten, kann bereits jetzt festgestellt werden, daß derartige Effekte wie auch deren Ausbleiben im allgemeinen psychologisch mehrdeutig sind. Natürlich steht diese Bewertung unter dem Vorbehalt, daß sie aus einem speziellen Modell abgeleitet wurde, für das vorerst nicht mehr als Plausibilitätserwägungen sprechen. Ein alternatives Modell, das für die eindeutige Interpretierbarkeit bestimmter „Falschinformationseffekte” spricht und darüber hinaus als empirisch bewährt gelten kann, ist allerdings bislang nicht in Sicht.

Der Schluß aus der gegebenen Situation kann nur lauten, daß man am besten das hier vorgestellte multinomiale Modell zunächst einmal möglichst rigorosen Tests und Validitätsüberprüfungen unterzieht. Übersteht das Modell diese Tests, dann begründet dies auch die oben geschilderte Kritik an der Forschung zu Falschinformationseffekten in Quellendiskriminationstests. Gleichzeitig eröffnet sich die Möglichkeit, die Frage nach möglichen Gedächtnisbeeinträchtigungen durch nachträgliche Falschinformation über Parameterschätzungen und Hypothesentests zu entscheiden.

Die folgenden drei Abschnitte schildern drei Experimente, die der Überprüfung des geschilderten multinomialen Modells dienen. Die wichtigsten Schlußfolgerungen werden im abschließenden Abschnitt dieses Kapitels zusammengefaßt.

10.2.1 Experiment 1: Gibt es einen Gedächtnisbeeinträchtigungseffekt?

Das erste zu schildernde Experiment war von der Konzeption her eine Pilotuntersuchung, mit der eruiert werden sollte, ob das Falschinformationsmodell (FI-Modell) aus Abbildung 10.4 empirische Daten zum Versuchsdesign aus Tabelle 10.2 potentiell adäquat abbilden kann. Darüber hinaus sollte ein erster Test der Gedächtnisbeeinträchtigungshypothese durchgeführt werden. Praktisch läßt sich dies über einen χ^2-Differenzentest der Nullhypothese $r(OI)_k = r(OI)_n = r(OI)_k$ realisieren, nach der die Abrufbarkeit der Originalinformation nicht von der Qualität der Nachinformation abhängt. Validierungshypothesen zum FI-Modell werden erst in den folgenden beiden Abschnitten 10.2.2 und 10.2.3 aufgegriffen und bleiben in diesem Abschnitt noch ausgeklammert.

Insgesamt 43 männliche und weibliche Studierende des zweiten Semesters im Diplomstudiengang Psychologie nahmen an einem Gruppenexperiment teil, dessen Versuchsdesign Tabelle 10.2 entsprach. Die Daten zweier Vpn gingen allerdings in die Auswertung nicht ein, da sie die Untersuchung nicht bis zum Ende absolvierten bzw. unvollständige Antwortbögen abgaben. Das Experiment wurde als Untersuchung zum Einfluß visueller und verbaler Kontextinformationen auf die Personenwahrnehmung ausgegeben. Der Versuchsablauf umfaßte fünf Phasen:

1) Zunächst sahen die Vpn für 20 Sekunden ein Dia, in welchem ein Studentenzimmer mit einem Schreibtisch im Mittelpunkt dargestellt wurde. Auf dem Schreibtisch befanden sich allerlei Objekte, darunter die kritischen Objekte für den späteren unangekündigten Gedächtnistest. Den Vpn wurde gesagt, daß sie das Dia sorgfältig betrachten und sich einen Eindruck von der Person machen sollten, die dieses Zimmer bewohnt.
2) Im Anschluß an die Diabetrachtung mußten die Vpn ihren Eindruck vom Zimmerbewohner über ein semantisches Differential (Ertel, 1965) beurteilen. Sie hatten hierfür fünf Minuten Zeit. In Wirklichkeit handelte es sich lediglich um eine Distraktortätigkeit, die bei der Auswertung unberücksichtigt bliebt.
3) Anschließend bekamen die Vpn eine schriftliche Schilderung des Zimmers und der darin enthaltenen Objekte, die — so wurde den Vpn gesagt — nicht vom Photographen des Dias, sondern von einer anderen Person zu einem späteren Zeitpunkt angefertigt wurde. Die Vpn wurden instruiert, sich aufgrund der Zimmerbeschreibung erneut einen Eindruck vom Zimmerbewohner zu machen. Hierfür hatten sie fünf Minuten Zeit.
4) Danach sollten die Vpn ein zweites Mal ihren gegenüber Phase 2 möglicherweise veränderten Eindruck vom Zimmerbewohner mittels semantischem Differential beurteilen. Auch diese Beurteilung war eine reine Distraktortätigkeit; sie dauerte fünf Minuten.

5) Das Ende des Experiments bildete eine überraschender Quellendiskriminationstest, bei dem die Vpn zu 18 kritischen Items und 16 unkritischen Items (filler items) jeweils ankreuzen sollten, ob sie das betreffende Objekt entweder
   - nur im Dia gesehen haben (OI-Antwort),
   - nur in der Geschichte gelesen haben (NI-Antwort),
   - im Dia gesehen und in der Geschichte gelesen haben (OI&NI-Antwort) oder
   - weder im Dia gesehen noch in der Geschichte gelesen haben (Neu-Antwort).

Das gesamte Experiment dauerte etwa 30 Minuten.

Vor Beginn des Experiments wurden die Vpn per Zufall auf drei annähernd gleich große Gruppen aufgeteilt, die jeweils unterschiedliche Versionen des Dias aus Phase 1 sahen. Von den 18 kritischen Items des späteren Gedächtnistests wurden pro Dia sechs Items in Version A, sechs Items in Version B und sechs Items gar nicht dargeboten. Die beiden Versionen eines Items waren unterschiedliche Instanzen einer gemeinsamen Oberkategorie. Zur Oberkategorie „Getränkedose vorne rechts auf dem Schreibtisch" konnte Version A etwa „Cola-Dose" und Version B „Fanta-Dose" lauten, zur Oberkategorie „Jacke auf dem Stuhl vor dem Schreibtisch" analog „Lederjacke" (Version A) versus „Jeansjacke" (Version B) usw. Welche sechs Items jeweils in Version A, in Version B oder gar nicht dargeboten wurden, wurde über die drei Dias und die damit assoziierten Vpn-Gruppen ausbalanciert, so daß jedes Item über alle Vpn hinweg annähernd gleich oft in jeder der drei Versionen (Version A, Version B, Distraktorbedingung) vorkam.

Analog zu den drei Dias wurden drei Versionen der Raumbeschreibung erstellt, wobei von den 18 kritischen Items jeweils sechs in Version A, sechs in Version B und sechs gar nicht geschildert wurden. Ungefähr ein Drittel der Vpn jeder der drei Bild-Bedingungen bekam jeweils eine der drei Versionen der Raumbeschreibungen, so daß $3 \cdot 3 = 9$ verschiedene Bild-Text-Kombinationen für die kritischen Items resultierten: A - A (konsistente Information), A - B (widersprüchliche Information), A - keine Information (NI neutral), B - A (widersprüchliche Information), B - B (konsistente Information), B - keine Information (NI neutral), keine Information - A (OI neutral), keine Information - B (OI neutral), keine Information - keine Information (reine Distraktorbedingung). Jede dieser neun Gruppen von Vpn wurde nun noch einmal auf zwei Subgruppen aufgeteilt, die unterschiedliche Versionen des Quellendiskriminationstests bekamen. In Version A des Tests wurde grundsätzlich die A-Form jedes kritischen Items präsentiert, in Version B analog die B-Form. Zusätzlich enthielten die beiden Versionen des Quellendiskriminationstests 16 weitere Testitems (filler items), die bei der Auswertung unberücksichtigt bleiben.

Insgesamt resultierten durch die Aufteilungsprozedur $3 \cdot 3 \cdot 2 = 18$ (Versionen der OI × Versionen der NI × Versionen des Gedächtnistests) verschiedene Bedingungskombinationen, die sich problemlos zu den neun Bedingungen des Versuchsdesigns aus Tabelle 10.2 zusammenfassen lassen, wenn man die Inhalte der Items, die in konsistenter, neutraler oder widersprüchlicher Version zueinander stehen, unberücksichtigt läßt. Jede Vp liefert 18 Urteile zu den kritischen Items, und zwar zwei Urteile für jede der neun Versuchsbedingungen des Designs. Der Gesamtstichprobenumfang für die modellbasierten Analysen beläuft sich also auf 41 Vpn × 18 Items, d.h. $N = 774$. Für jede der neun Versuchsbedingungen ergibt dies ein $N_k = 82$.

Tabelle 10.3:
Absolute Antworthäufigkeiten (ohne Klammern) und relative Antworthäufigkeiten (mit Klammern) in Experiment 1 zum FI-Modell.

		Vp-Antwort im Quellendiskriminationstest			
Versuchsbedingung	Beschreibung	„OI“	„NI“	„OI&NI“	„neu“
A1	OI & NI-Item	10 (.12)	22 (.27)	41 (.50)	9 (.11)
A2	OI-Item, keine NI	46 (.56)	2 (.02)	3 (.04)	31 (.38)
A3	OI-Item, widerspr. NI	36 (.44)	6 (.07)	6 (.07)	34 (.41)
A4	NI-Item, keine OI	4 (.05)	59 (.72)	4 (.05)	15 (.18)
A5	NI-Item, widerspr. OI	6 (.07)	43 (.52)	17 (.21)	16 (.20)
N1	keine OI, keine NI	4 (.05)	4 (.05)	0 (.00)	74 (.90)
N2	keine OI, widers. NI	4 (.05)	14 (.17)	0 (.00)	64 (.78)
N3	widers. OI, keine NI	16 (.20)	2 (.02)	4 (.05)	60 (.73)
N4	widers. OI und NI	7 (.09)	10 (.12)	7 (.09)	58 (.70)

Tabelle 10.3 stellt die absoluten und relativen Antworthäufigkeiten für jede der neun Versuchsbedingungen zusammen. Ein Falschinformationseffekt im Sinne von Lindsay und Johnson (1989) ist offenbar auch in unserer Untersuchung nicht aufgetreten, weil sowohl in Bedingung A4 als auch in Bedingung N1 jeweils 8 von 82 möglichen Fehlattributionen auf die OI-Quelle vorgenommen wurden. Deskriptiv deutlich erkennbar ist jedoch ein Falschinformationseffekt zweiter Art im Sinne von Lindsay (1990), weil die OI bei neutraler NI 46 von 82 mal korrekt wiedererkannt wird, bei widersprüchlicher Information dagegen nur 36 von 82 mal. Allerdings erreicht dieser Effekt bei dem gegebenen geringen Stichprobenumfang die 5%-Signifikanzschranke nicht ($\chi^2_{(1)} = 2.44$). Signifikant auf dem Niveau $\alpha = .05$ wird dagegen der umgekehrte Falschinformationseffekt: bei neutraler OI wird ein NI-Item 59 mal korrekt auf die NI attribuiert, bei widersprüchlicher OI dagegen nur 43 mal ($\chi^2_{(1)} = 6.78$). Besonders auffällig sind die Unterschiede in den OI- und NI-Antworthäufigkeiten bei Distraktoritems der Bedingungen N2 und N3 ($\chi^2_{(1)} =$

16.20) : Ein Distraktoritem wird offenbar eher fälschlich einer bestimmten Quelle zugeordnet, wenn in dieser Quelle ein anderes Item der gleichen Oberkategorie vorhanden war.

Insgesamt ist somit überzeugende Evidenz für eine Beeinflussung von Quellengedächtnisurteilen durch die jeweilige irrelevante Kontextquelle gegeben, und zwar sowohl für alte Items aus beiden Quellen, als auch für Distraktoritems, die in keiner der beiden Quellen vorkamen. Wir können uns nunmehr der modellbasierten Analyse dieser Effekte zuwenden.

Die psychologischen Ursachen für die auf der Ebene der Antworthäufigkeiten beobachtbaren Effekte lassen sich durch eine modellbasierte Analyse der Daten erhellen, sofern vorausgesetzt werden kann, daß das FI-Modell aus Abbildung 10.4 die zugrundeliegenden Gedächtnis- und Urteilsprozesse adäquat beschreibt. Bei den im folgenden zu berichtenden statistischen Analysen wurde generell ein α-Niveau von .05 zugrunde gelegt, damit die Stärke der Tests aufgrund des relativ geringen Stichprobenumfangs nicht zu gering ausfällt. Mittelstarke Effekte im Sinne von Cohen (w = .30) werden bei dem gegeben N und α = .05 mit einer Teststärke von $1-\beta > .98$ entdeckt, kleine Effekte (w = .10) dagegen nur mit $1-\beta = .55$ (bei df = 5) bzw. $1-\beta = .79$ (bei df = 1). Der G^2-Modellanpassungstest liefert für das Gesamtmodell (angewandt auf alle neun Itemklassen) die Anpassungsstatistik G^2 = 19.70, was bei $df = 9 \cdot (4-1) - 12 = 15$ nahezu dem Erwartungswert der χ^2-Verteilung entspricht und somit sehr zufriedenstellend ist.

Die ML-Parameterschätzungen und 90%-Konfidenzintervalle sind in Tabelle 10.4 zusammengefaßt. Auf den ersten Blick sind generell bessere Gedächtnisleistungen für Nachinformationsitems verglichen mit Originalinformationsitems zu erkennen. Dies kann drei potentielle Ursachen haben: (a) das kürzere Retentionsintervall für NI-Items, (b) generell bessere Enkodierungsbedingungen für verbal statt visuell dargebotene Informationen und (c) die bessere Übereinstimmung zwischen Enkodierungs- und Abrufsituation im Falle verbal dargebotener Informationen (vgl. z.B. Baddeley, 1997, Kapitel 11). Die Klärung der Frage nach den Ursachen kann mit den Daten des ersten Experiments nicht geleistet werden. Wir werden in Abschnitt 10.2.3 auf diesen Punkt zurückkommen.

Die ML-Parameterschätzungen und 90%-Konfidenzintervalle sind in Tabelle 10.4 zusammengefaßt. Auf den ersten Blick sind generell bessere Gedächtnisleistungen für Nachinformationsitems verglichen mit Originalinformationsitems zu erkennen. Dies kann drei potentielle Ursachen haben: (a) das kürzere Retentionsintervall für NI-Items, (b) generell bessere Enkodierungsbedingungen für verbal statt visuell dargebotene Informationen und (c) die bessere Übereinstimmung zwischen Enkodierungs- und Abrufsituation im Falle verbal dargebotener Informationen (vgl. z.B. Baddeley, 1997, Kapitel 11). Die Klärung der Frage nach den Ursa-

chen kann mit den Daten des ersten Experiments nicht geleistet werden. Wir werden in Abschnitt 10.2.3 auf diesen Punkt zurückkommen.

Tabelle 10.4:
ML-Parameterschätzungen und 90%-Konfidenzintervalle bei Anwendung des FI-Modells auf die Daten zu Experiment 1 (vgl. Tabelle 10.3).

		90%-Konfidenzintervall	
Parameter	ML-Schätzwert	untere Grenze	obere Grenze
$r(\mathrm{OI})_\mathrm{k}$	.58	.49	.68
$r(\mathrm{OI})_\mathrm{n}$	.56	.46	.65
$r(\mathrm{OI})_\mathrm{w}$	.47	.37	.56
$r(\mathrm{NI})_\mathrm{k}$	.73	.64	.81
$r(\mathrm{NI})_\mathrm{n}$	.74	.66	.82
$r(\mathrm{NI})_\mathrm{w}$	.70	.61	.78
$g(\mathrm{OI})_\mathrm{nw}$	.09	.06	.11
$g(\mathrm{OI})_\mathrm{w}$	.24	.20	.29
$g_1(\mathrm{NI})_\mathrm{nw}$	.12	.06	.18
$g_2(\mathrm{NI})_\mathrm{nw}$	.06	.03	.09
$g_1(\mathrm{NI})_\mathrm{w}$	.24	.16	.33
$g_2(\mathrm{NI})_\mathrm{w}$	.17	.13	.22

Im Zentrum des ersten Experiments stand die Frage nach möglichen Gedächtnisbeeinträchtigungen durch inkonsistente Informationen. Schlechtere Gedächtnisparameterschätzwerte für die Bedingungen mit widersprüchlicher Information (d.h. $r(\mathrm{OI})_\mathrm{w}$ und $r(\mathrm{NI})_\mathrm{w}$) sind zwar erkennbar, aber schon die überlappenden Konfidenzintervalle deuten an, daß dieser Effekt statistisch nicht signifikant ist. Tatsächlich erreicht sowohl die Differenzen-G^2-Statistik zur Nullhypothese $r(\mathrm{OI})_\mathrm{k} = r(\mathrm{OI})_\mathrm{n} = r(\mathrm{OI})_\mathrm{w}$ ($\Delta G^2 = 2.12$, $df = 2$) als auch die zur Nullhypothese $r(\mathrm{NI})_\mathrm{k} = r(\mathrm{NI})_\mathrm{n} = r(\mathrm{NI})_\mathrm{w}$ ($\Delta G^2 = 0.36$, $df = 2$) die Signifikanzgrenze bei weitem nicht. Trotz der auf Ebene der Antworthäufigkeiten erkennbaren Falschinformationseffekte ist also kein Grund für die Annahme von Gedächtnisbeeinträchtigungen durch nachträgliche oder vorausgehende Falschinformationen gegeben. Allerdings ist bei diesem Urteil die Teststärke der Hypothesentests im Blick zu behalten. Nimmt man an, daß die wahre Situation in der zugrundeliegenden Population durch die Parameter $r(\mathrm{OI})_\mathrm{k} = .60$, $r(\mathrm{OI})_\mathrm{n} = .50$ und $r(\mathrm{OI})_\mathrm{w} = .40$ gekennzeichnet ist, wobei die anderen Parameter die Werte der ML-Schätzer aus Tabelle 10.4 annehmen, so ist bei dem gegebenen Stichprobenumfang und $\alpha = .05$ die approximative Teststärke des G^2-Differenzentests zur Nullhypothese $r(\mathrm{OI})_\mathrm{k} = r(\mathrm{OI})_\mathrm{n} = r(\mathrm{OI})_\mathrm{w}$ lediglich $1-\beta = .57$, bei den Parameterwerten $r(\mathrm{OI})_\mathrm{k} = .70$, $r(\mathrm{OI})_\mathrm{n} = .50$ und $r(\mathrm{OI})_\mathrm{w} = .30$ aber immerhin

schon $1-\beta = .99$.[104] Für kleine Abweichungen von der Nullhypothese ist die Teststärke also nicht überzeugend, aber doch deutlich größer als in der zu Beginn von Kapitel 10 besprochenen Untersuchung von Wagenaar und Boer (1987).

Erwartungsgemäß deuten sich positiv korrelierte Ratetendenzen bezüglich der OI- und NI-Quellen auf der Ebene der Parameterschätzwerte an. Der Test der stochastischen Unabhängigkeit der Ratetendenzen ($g_1(\mathrm{NI})_{\mathrm{nw}} = g_2(\mathrm{NI})_{\mathrm{nw}} = g(\mathrm{NI})_{\mathrm{nw}}$ sowie $g_1(\mathrm{NI})_{\mathrm{w}} = g_2(\mathrm{NI})_{\mathrm{w}} = g(\mathrm{NI})_{\mathrm{w}}$) zeigt jedoch, daß die Nullhypothese dennoch beibehalten werden kann ($\Delta G^2 = 4.24$, $df = 2$). Massive signifikante Effekte liegen dagegen bezüglich der Ratewahrscheinlichkeiten vor: $g(\mathrm{OI})_{\mathrm{w}}$ ist signifikant größer als $g(\mathrm{OI})_{\mathrm{nw}}$ ($\Delta G^2 = 24.53$, $df = 1$) und die beiden $g(\mathrm{NI})_{\mathrm{w}}$-Parameter sind signifikant größer als die beiden $g(\mathrm{NI})_{\mathrm{nw}}$-Parameter ($\Delta G^2 = 13.28$, $df = 2$). Neben den besseren Gedächtnisparametern für NI-Items verglichen mit OI-Items (s.o.) ist der Unterschied zwischen den beiden g-Parameterklassen ganz offensichtlich die stärkste Effektquelle in Experiment 1.

Zusammenfassend kann festgehalten werden, daß das FI-Modell Daten zum Quellendiskriminationstest — erhoben im Neun-Bedingungen-Design aus Tabelle 10.2 — adäquat repräsentieren kann. Experiment 1 liefert keine überzeugenden Hinweise auf Gedächtnisbeeinträchtigungseffekte durch nachträgliche oder vorausgehende Falschinformation, schließt diese aber auch nicht völlig aus, weil deskriptiv ein Beeinträchtigungseffekt auf der Ebene der Parameterschätzwerte vorliegt, wenn auch nur schwach. Unverkennbar ist dagegen ein sehr starker Effekt der Informationsinkonsistenz auf Urteilsprozesse: Im Zustand der Gedächtnisunsicherheit sind Personen offenbar viel eher bereit, ein Item einer bestimmten Quelle zuzuordnen, wenn sich in dieser Quelle ein ähnliches Item aus der gleichen Oberkategorie befand, als wenn dies nicht der Fall ist. Die auf der Ebene der Antworthäufigkeiten beobachtbaren Unterschiede zwischen den verschiedenen Versuchsbedingungen lassen sich im wesentlichen auf diesen differentiellen Rateeffekt zurückführen, der psychologisch sehr plausibel ist.

10.2.2 Experiment 2: Effekte der Falschinformationswarnung

Nach der erfolgreichen Erprobung des FI-Modells in Experiment 1 sollte mit Experiment 2 zweierlei erreicht werden: Zum einen die Replikation der Befunde des ersten Experiments, insbesondere im Hinblick auf das Ausbleiben eines signifikanten Effekts der Informationskonsistenz- versus -inkonsistenz auf die Gedächtnisparameter des Modells. Zum zweiten sollte der Effekt einer Warnung exploriert werden, der die Vpn kurz vor dem Gedächtnistests auf die Möglichkeit widersprüchlicher Informationen in Original- und Nachinformation explizit hinweist.

---

[104] Die Teststärke wurde nach dem in Abschnitt 4.3.3.5 empfohlenen Verfahren approximiert, wobei der Nonzentralitätsparameter $\gamma^{(\lambda=0)}$ verwendet wurde.

Bisherige Untersuchungen zur Wirksamkeit derartiger Warnungen haben unterschiedliche Ergebnisse erbracht, was möglicherweise mit den in diesen Untersuchungen jeweils verwendeten Gedächtnistests zusammenhängt (Christiaansen & Ochalek, 1983; Green, Flynn & Loftus, 1982; Sporer, 1993; Sporer & aus dem Kahmen, 1993).

Das Design, das Material und der Ablauf des zweiten Experiments entsprachen dem von Experiment 1, mit dem einzigen Unterschied, daß nunmehr zwei Vpn-Gruppen untersucht wurden, eine unter den Bedingungen des ersten Experiments (Standardbedingung) und eine mit explizitem Warnhinweis vor dem Quellengedächtnistest (Warnbedingung). Insgesamt 69 Studierende des zweiten Semesters im Diplomstudiengang Psychologie nahmen an dem Experiment teil, davon 40 in der Standardbedingung und 29 in der Warnbedingung. Jede der beiden Gruppen wurde analog zum ersten Experiment auf 18 Untergruppen aufgeteilt, die zwecks Ausbalancierung von Item- und Materialeffekten mit unterschiedlichen Kombinationen von a) OI-Items, b) NI-Items und c) Items im Quellengedächtnistest konfrontiert wurden. Insgesamt gingen pro Vp Antworten auf 18 kritische Items in die Auswertung ein, genau wie im ersten Experiment. Da einige wenige Items von einigen Vpn nicht beantwortet wurden, wurde die Maximalzahl möglicher Antworten in einigen Bedingungen des Designs nicht erreicht.

Erwartet wurden folgende Resultate:

1) Das Modell sollte wie in Experiment 1 gut auf die Daten beider experimentellen Bedingungen passen;
2) Die Standardbedingung von Experiment 2 sollte die Ergebnisse des ersten Experiments nahezu perfekt replizieren, da sich beide Untersuchungen in allen relevanten Details entsprachen;
3) Effekte der Warnbedingung sind insofern zu erwarten, als die Durchführung eines Quellendiskriminationstests in der Standardbedingung auf die Möglichkeit von OI-NI-Widersprüchen nicht ausdrücklich hinweist. Allerdings sollten die Gedächtnisparameter des Modells hiervon nicht betroffen sein, sondern nur die Urteilsparameter. Die Bereitschaft, im Zustand der Gedächtnisunsicherheit ein Item einer bestimmten Quelle zuzuordnen, sollte in der Warnbedingung geringer sein, und zwar um so mehr, je höher die entsprechende Ratebereitschaft in der Standardbedingung ausfällt. Denn ist die Ratebereitschaft schon in der Standardbedingung sehr gering, dann verhindert ein „Bodeneffekt", daß die Warnung einen zusätzlichen Reduktionseffekt bewirken kann. Experiment 1 hatte gezeigt, daß die Ratebereitschaft in bezug auf die NI-Quelle etwas größer ist als in bezug auf die OI-Quelle, und daß die Ratebereitschaft größer ist, wenn sich in der entsprechenden Quelle ein ähnliches Item der gleichen Oberkategorie befindet. Wenn also ein Effekt der Warnbedingung auftritt, dann sollte er sich besonders in den

NI-Rateparametern und den Rateparametern mit dem Index „w“ bemerkbar machen.

Tabelle 10.5:
Absolute Antworthäufigkeiten (ohne Klammern) und relative Antworthäufigkeiten (mit Klammern) in den beiden Versuchsbedingungen des Experiments 2 zum FI-Modell.

		Standardbedingung				Falschinformationswarnung			
Vb	Quelle	„OI“	„NI“	„OI&NI“	„neu“	„OI“	„NI“	„OI&NI“	„neu“
A1	OI&NI	16 (.20)	21 (.26)	35 (.44)	8 (.10)	13 (.22)	14 (.24)	27 (.47)	4 (.07)
A2	OI (n.)	37 (.46)	4 (.05)	8 (.10)	31 (.39)	29 (.50)	3 (.05)	1 (.02)	25 (.43)
A3	OI (w.)	41 (.53)	3 (.04)	12 (.15)	22 (.28)	25 (.43)	1 (.02)	5 (.09)	27 (.47)
A4	NI (n.)	2 (.03)	54 (.68)	4 (.05)	19 (.24)	1 (.02)	45 (.79)	2 (.04)	9 (.16)
A5	NI (w.)	5 (.06)	54 (.69)	8 (.10)	11 (.14)	3 (.05)	36 (.62)	9 (.16)	10 (.17)
N1	nw./nw.	3 (.04)	7 (.09)	1 (.01)	69 (.86)	1 (.02)	2 (.04)	1 (.02)	53 (.92)
N2	nw./w.	7(.09)	15 (.19)	1 (.01)	57 (.71)	2 (.04)	9 (.16)	2 (.04)	44 (.77)
N3	w. /nw.	8(.10)	7 (.09)	3 (.04)	62 (.78)	5 (.09)	1 (.02)	1 (.02)	51 (.87)
N4	w./w.	10 (.13)	10 (.13)	1 (.01)	58 (.73)	8 (.14)	6 (.11)	3 (.05)	40 (.70)

Die Antworthäufigkeiten des Experiments 2 sind in Tabelle 10.5 zusammengefaßt. Der Vergleich der relativen Häufigkeiten in der Standardbedingung mit denen des Experiments 1 zeigt sehr viele Ähnlichkeiten. Interessanterweise ist auf der Ebene der (relativen) Antworthäufigkeiten in Experiment 2 eher weniger denn mehr Evidenz für Falschinformationseffekte gegeben. Der Vergleich der Bedingungen A2 und A3 sowie A4 und A5 deutet sogar leichte negative Falschinformationseffekte an, d.h. bessere Gedächtnisurteile für OI- oder NI-Items, falls in der jeweils anderen Quelle eine widersprüchliche Information enthalten ist. Die Antworten auf die Distraktoren der Bedingungen N2 (ähnliches Item in der NI) und N3 (ähnliches Item in der OI) zeigen zwar deskriptiv ein ähnliches Muster wie in Experiment 1, jedoch werden die Unterschiede der Falschattributionen auf die OI- und die NI-Quelle nicht signifikant ($\chi^2_{(1)} = 1.71$).

In der Warnbedingung sind demgegenüber leichte positive Falschinformationseffekte zu beobachten, ähnlich wie in Experiment 1. Die 5%-Signifikanzschranke wird für den Effekt der NI auf korrekte Erinnerungen an die OI nicht erreicht ($\chi^2_{(1)} = 0.69$), wohl aber für den Effekt inkonsistenter OI auf die Erinnerungen an die NI ($\chi^2_{(1)} = 3.93$). Der Hauptunterschied der Warnbedingung zur Standardbedingung ist jedoch die geringere Rate falscher Attributionen eines Items auf die Nachinformation. Anders als in der Standardbedingung werden in der Warnbedingung auch die unterschiedlichen Falschattributionsraten für die Distraktorklassen N2 und N3 in der erwarteten Richtung signifikant ($\chi^2_{(1)} = 6.81$).

Tabelle 10.6
ML-Parameterschätzungen und 90%-Konfidenzintervalle bei Anwendung des FI-Modells auf die Daten beider Versuchsbedingungen des Experiments 2 (vgl. Tabelle 10.5).

Parameter	Standardbedingung			Falschinformationswarnung		
		90%-Konfidenzintervall			90%-Konfidenzintervall	
	ML-Schätzwert	untere Grenze	obere Grenze	ML-Schätzwert	untere Grenze	obere Grenze
$r(\mathrm{OI})_k$	.61	.51	.70	.67	.57	.78
$r(\mathrm{OI})_n$	.53	.43	.63	.49	.38	.61
$r(\mathrm{OI})_w$	.65	.56	.75	.49	.38	.61
$r(\mathrm{NI})_k$	.65	.54	.75	.69	.58	.79
$r(\mathrm{NI})_n$	.70	.61	.79	.82	.73	.90
$r(\mathrm{NI})_w$	.77	.68	.85	.76	.67	.86
$g(\mathrm{OI})_{nw}$	.08	.05	.10	.05	.02	.08
$g(\mathrm{OI})_w$	.15	.11	.19	.17	.12	.21
$g_1(\mathrm{NI})_{nw}$	.18	.11	.26	.07	.01	.14
$g_2(\mathrm{NI})_{nw}$	.10	.07	.14	.05	.02	.08
$g_1(\mathrm{NI})_w$	.19	.12	.27	.22	.12	.32
$g_2(\mathrm{NI})_w$	.17	.12	.22	.13	.08	.17

Tabelle 10.6 faßt die Ergebnisse der modellbasierten Analyse der Ergebnisse zusammen. Alle statistischen Tests wurden wie im ersten Experiment unter Zugrundelegung von $\alpha = .05$ durchgeführt. Relativ zu diesem Kriterium paßt das FI-Modell auf die Daten der Standardbedingung ($G^2 = 9.58$, $df = 15$) und auch auf die Daten der Warnbedingung ($G^2 = 17.72$, $df = 15$) sehr gut.

Schlechtere Gedächtnisparameterschätzwerte für die Bedingungen mit widersprüchlicher Information (d.h. $r(\mathrm{OI})_w$ und $r(\mathrm{NI})_w$) sind weder in der Standardbedingung noch in der Warnbedingung erkennbar. Auch gegenteilige Effekte liegen nicht vor: Die Differenzen-G^2-Statistik zur Nullhypothese $r(\mathrm{OI})_k = r(\mathrm{OI})_n = r(\mathrm{OI})_w$ ist für die Standardbedingung ($\Delta G^2 = 2.37$, $df = 2$) und für die Warnbedingung ($\Delta G^2 = 4.80$, $df = 2$) insignifikant, ebenso die Statistiken zur Nullhypothese $r(\mathrm{NI})_k = r(\mathrm{NI})_n = r(\mathrm{NI})_w$ (Standardbedingung: $\Delta G^2 = 2.36$, $df = 2$; Warnbedingung: $\Delta G^2 = 2.41$, $df = 2$). Die Evidenz für Gedächtnisbeeinträchtigungseffekte durch widersprüchliche Informationen ist in Experiment 2 also noch schwächer als in Experiment 1. Erneut zeigt sich auch die bessere Gedächtnisleistung für NI-Items im Vergleich zu OI-Items. Dieser Effekt ist sowohl für die Standardbedingung als auch für die Warnbedingung zu beobachten, wenn auch nicht ganz so prägnant wie in Experiment 1.

Positiv korrelierte Ratetendenzen bezüglich der OI- und NI-Quellen deuten sich wie schon in Experiment 1 auf deskriptiver Ebene an. Der Test der stochastischen Unabhängigkeit der Ratetendenzen zeigt jedoch erneut, daß die Nullhypothese für die Standardbedingung ($\Delta G^2 = 2.95$, $df = 2$) und für die Warnbedingung ($\Delta G^2 = 2.71$, $df = 2$) beibehalten werden kann. Starke Effekte liegen dagegen bezüglich der Ratewahrscheinlichkeiten vor: $g(\mathrm{OI})_{\mathrm{w}}$ ist signifikant größer als $g(\mathrm{OI})_{\mathrm{nw}}$, und zwar für die Standardbedingung ($\Delta G^2 = 6.48$, $df = 1$) und auch für die Warnbedingung ($\Delta G^2 = 17.89$, $df = 1$). Ein ähnliches Muster ist für die $g(\mathrm{NI})_{\mathrm{w}}$- und die $g(\mathrm{NI})_{\mathrm{nw}}$-Parameter zu beobachten, wobei der Effekt allerdings nur für die Warnbedingung signifikant wird ($\Delta G^2 = 9.66$, $df = 2$), während er in der Standardbedingung die Signifikanzgrenze nicht erreicht ($\Delta G^2 = 2.61$, $df = 2$). Insgesamt zeigt dieses Ergebnismuster nur geringfügige Unterschiede zu dem von Experiment 1. Die ersten beiden Hypothesen zu Experiment 2 können also als bestätigt gelten.

Zu klären ist nun, ob sich die Warnbedingung im Vergleich zur Standardbedingung auf die Modellparameter so auswirkt, wie es in Hypothese 3 vorhergesagt wurde. Erwartungsgemäß unterscheiden sich die sechs Gedächtnisparameter beider Bedingungen nicht signifikant ($\Delta G^2 = 6.25$, $df = 6$). Signifikant wird demgegenüber die Reduktion der Ratewahrscheinlichkeiten $g_1(\mathrm{NI})_{\mathrm{nw}}$ und $g_2(\mathrm{NI})_{\mathrm{nw}}$ durch die Warnbedingung ($\Delta G^2 = 6.36$, $df = 2$), was ebenfalls den Erwartungen entspricht. Nicht signifikant wird dagegen der Effekt der Warnbedingung auf $g_1(\mathrm{NI})_{\mathrm{w}}$ und $g_2(\mathrm{NI})_{\mathrm{w}}$. Dieses Ergebnis ist überraschend, da die entsprechenden Rateparameter in der Standardbedingung die höchsten Werte annehmen. Auch die Rateparameter zur OI-Quelle werden durch die Warnbedingung nicht wesentlich affiziert.

Wie lassen sich diese Ergebnisse interpretieren? Offenbar fruchtet die Warnung nichts, wenn die Vpn beispielsweise ein gelbes Langenscheidt-Wörterbuch als Distraktoritem zu beurteilen haben, während in der NI ein grünes Pons-Wörterbuch erwähnt wurde. Die Bereitschaft, das Distraktoritem fälschlich der betreffenden Quelle zuzuordnen, läßt sich durch eine Warnung kaum oder gar nicht beeinflussen. Eine Erklärungsmöglichkeit besteht darin, daß die Vpn das ursprüngliche NI-Item (grünes Pons-Wörterbuch) gar nicht auf der Detail-Ebene enkodiert haben, sondern sich lediglich eingeprägt haben, daß dort ein Wörterbuch erwähnt wurde. Unter diesen Umständen erscheint „Langenscheidt-Wörterbuch" subjektiv gar nicht mehr als Distraktoritem, sondern als ein altes Item, das in der betreffenden Quelle wiedererkannt wird. Es besteht also aus Vp-Sicht kein Zustand der Gedächtnisunsicherheit, so daß selbst in der Warnbedingung kein besonderer Grund zur Vorsicht gegeben ist. Anders sieht es aus, wenn ein Distraktoritem zu beurteilen ist, das keine Ähnlichkeiten zu einem erinnerten OI- oder NI-Item aufweist. Der Zustand der Gedächtnisunsicherheit wird sich in Verbindung mit der Warnung in einer konservativen Ratestrategie niederschlagen, was auch beobachtet wurde.

Zusammenfassend kann festgehalten werden, daß die Hypothesen 1 und 2 bestätigt wurden, die Hypothese 3 über Effekte einer Warnbedingung aber nur teilweise. Die Warnung affiziert die Gedächtnisparameter nicht, wohl aber die Rateparameter, die in der Standardbedingung hohe Werte annehmen, d.h. insbesondere die *g*(NI)-Parameter. Hiervon auszunehmen sind Rateeffekte im Falle ähnlicher (kategoriegleicher) Items in der betreffenden Zielquelle. Eine mögliche Ursache für diesen überraschenden Effekt ist die Enkodierung der ursprünglich dargebotenen Information auf Oberkategorieebene.

10.2.3 Experiment 3: Effekte der Nachinformationsmodalität

Das dritte hier zu berichtende Experiment diente der Suche nach Bedingungen, die Falschinformationseffekte auf das Gedächtnis — wenn es sie denn geben sollte — wahrscheinlicher bzw. durchschlagskräftiger machen. Darüber hinaus sollte der Effekt der Modalität der Nachinformation (Textdarbietung versus Bilddarbietung) auf die Parameter des FI-Modells untersucht werden. Zwei Erwartungen standen dahinter: Einerseits könnten Gedächtnisbeeinträchtigungseffekte eher zum Ausdruck kommen, wenn OI und NI der gleichen Modalität angehören, weil dann ein perzeptuelles Attribut fehlt, das die Diskriminierung zwischen Original- und Nachinformation in den Experimenten 1 und 2 möglicherweise erleichtert hat (visuelle versus verbale Darbietung einer Information). Zum zweiten erlaubt die Darbietung der Nachinformation in Bildform eine Überprüfung möglicher Hypothesen zur Erklärung des Gedächtnisvorteils der Nachinformation: Wenn das kürzere Retentionsintervall für die NI entscheidend ist, sollte der Gedächtnisvorteil der Nachinformation unabhängig davon auftreten, ob die Nachinformation als Text oder als Bild dargeboten wird. Erleichtert dagegen die Textform das Enkodieren des Zielmaterials oder die Übereinstimmung zwischen Enkodierungskontext und Abrufkontext den Abruf der Zielinformation, dann sollte der Nachinformationsvorteil verschwinden, wenn die Nachinformation genau wie die Originalinformation in bildlicher Form dargeboten wird.

Durchsucht man die Literatur zu Falschinformationseffekten nach Hinweisen auf Bedingungen, die Falschinformationseffekte wahrscheinlicher machen, dann wird oft die periphere Darbietung einer Information als förderlicher Einflußfaktor genannt (Sporer, 1993; Sporer & aus dem Kahmen, 1993). In Experiment 1 und 2 wurden die Zielitems jedoch zentral auf dem Schreibtisch des abgebildeten Studentenzimmers dargeboten, was möglicherweise Suggestiveffekte nachträglicher Falschinformation verhindert hat. Experiment 3 verwendete aus diesem Grunde neues Material, dessen kritische Items ausschließlich peripheren Informationen entsprachen. Dies sollte einerseits die Gedächtnisparameter gegenüber Experiment 1 und 2 verringern, andererseits auch Falschinformationseffekte fördern.

Die Instruktion der Vpn in Experiment 3 — erneut 62 Studierende im zweiten Semester des Diplomstudienganges Psychologie, die natürlich an den vorherigen Untersuchungen nicht teilgenommen hatten — zielte darauf ab, daß es um die Überprüfung einer neuen Bilderserie zum Thematischen Apperzeptionstest (TAT) geht. Die Vpn waren mit der Original-TAT-Bilderserie nach Murray vertraut und somit darüber informiert, daß die Originalbilder des Tests veraltet wirken. Das vorgebliche Bemühen, neue TAT-Bilder entwickeln zu wollen, erschien deshalb in den Augen der Vpn plausibel.

Die Vpn sahen — wie sie es vom TAT gewohnt waren — für 1 Minute ein Bild, und mußten dann kurz beschreiben, was in dem Bild vorgeht, wie es zu der betreffenden Situation wohl gekommen ist, woran die Personen gerade denken und wie es wohl weitergehen wird. Das Bild stellte eine junge Frau und einen jungen Mann in einer Wohnküche dar. Sowohl ihr Gesichtsausdruck als auch ihre Körperhaltung waren mehrdeutig, so daß die Verwendung als TAT-Bild plausibel erschien. Eigentlicher Zweck der Bildpräsentation war natürlich die Darbietung peripherer kritischer Objekte, die in der Wohnküche plaziert waren. Hierzu zählt beispielsweise eine rote versus schwarze Thermoskanne, ein grünes versus rotbraunes Handtuch usw. Drei Bilder mit verschiedenen Kombinationen von 18 kritischen Items wurden erzeugt, analog zu den Experimenten 1 und 2.

Die nächste Phase des Experiments unterschied sich für Vpn der Bild-Bild-Bedingung und für Vpn der Bild-Text-Bedingung. In der Bild-Bild-Bedingung erhielten die Vpn ein weiteres vermeintliches TAT-Bild, das sich bei oberflächlicher Betrachtung nur dadurch vom ersten Bild unterschied, daß die junge Frau nunmehr einige Schritte auf den Mann zugegangen war. Im Detail waren jedoch einige Objekte der Wohnküche verändert, fortgelassen oder neu hinzugefügt, wie es für die Nacherzählungen in Falschinformationseffekten üblich ist. Anschließend bekamen die Vpn weitere Fragen zu den beiden Protagonisten, bevor der überraschende Quellendiskriminationstest für 18 kritische Items und weitere 18 Items folgte, die nicht in die Auswertung eingingen.

Für die Probanden der Bild-Text-Bedingung war der Ablauf identisch, nur daß statt des zweiten Bildes eine Nacherzählung präsentiert wurde, analog zum Ablauf von Experiment 1 und Experiment 2. Insgesamt 39 Vpn wurden der Bild-Bild-Bedingung zugewiesen, 23 Vpn der Bild-Text-Bedingung. Beide Gruppen wurden zufällig auf 18 Ausbalancierungsbedingungen aufgeteilt, analog zu Experiment 1 und Experiment 2.

Tabelle 10.7 faßt die absoluten und relativen Antworthäufigkeiten für Experiment 3 zusammen. Offensichtlich sind — wie intendiert — die Trefferraten für alte Items durchweg niedriger als in den vorherigen beiden Experimenten. Die Trefferraten für die Nachinformation sind in der Bild-Bild-Bedingung niedrig und in der Bild-Text-Bedingung vergleichsweise hoch, was die Vermutung nahelegt,

daß die NI nur dann besser erinnert wird als die OI, wenn erstere in Textform dargeboten wird. Trefferraten für die OI und die NI werden durch die Art der Kontextquelle offensichtlich nicht im Sinne eines Falschinformationseffekts beeinflußt. Auf deskriptiver Ebene erkennbar ist jedoch ein Falschinformationseffekt im Sinne von Lindsay und Johnson (1989): Nachinformationsitems der Bedingung A4 werden häufiger fälschlich auf die OI attribuiert als reine Distraktoritems der Bedingung N1, und zwar sowohl in der Bild-Bild-Bedingung (17% versus 12%) als auch in der Bild-Text-Bedingung (14% versus 4%). In beiden Bedingungen findet sich auch das bekannte Muster unterschiedlicher Raten falschen Alarms bei den Distraktorklassen N2 und N3, das sowohl für die Bild-Bild-Bedingung ($\chi^2_{(1)}$ = 19.06) als auch für die Bild-Text-Bedingung ($\chi^2_{(1)}$ = 10.83) signifikant wird.

Tabelle 10.7:
Absolute Antworthäufigkeiten (ohne Klammern) und relative Antworthäufigkeiten (mit Klammern) in den beiden Versuchsbedingungen des Experiments 3 zum FI-Modell.

		OI: Bild, NI: Bild				OI: Bild, NI: Text			
Vb	Quelle	„OI“	„NI“	„OI&NI“	„neu“	„OI“	„NI“	„OI&NI“	„neu“
A1	OI&NI	7 (.09)	9 (.12)	19 (.25)	42 (.55)	5 (.12)	15 (.35)	10 (.23)	13 (.30)
A2	OI (n.)	18 (.23)	5(.06)	5 (.06)	49 (.64)	11 (.24)	3 (.07)	2 (.04)	29 (.64)
A3	OI (w.)	19 (.26)	5 (.07)	7 (.09)	45 (.59)	7 (.16)	3 (.07)	5 (.12)	28 (.65)
A4	NI (n.)	2 (.03)	16 (.22)	10 (.14)	45 (.62)	2 (.05)	17 (.39)	4 (.09)	21 (.48)
A5	NI (w.)	6 (.08)	21 (.29)	8 (.11)	38 (.52)	5 (.12)	20 (.47)	4 (.09)	14 (.33)
N1	nw./nw.	5 (.07)	1 (.01)	4 (.05)	63 (.86)	1 (.02)	4 (.09)	1 (.02)	37 (.86)
N2	nw./w.	5 (.06)	12 (.15)	7 (.09)	54 (.69)	1 (.02)	10 (.23)	2 (.05)	30 (.70)
N3	w. /nw.	9 (.15)	3 (.05)	4 (.07)	43 (.73)	5 (.12)	3 (.07)	0 (.00)	35 (.81)
N4	w./w.	8 (.11)	8 (.11)	6 (.08)	51 (.70)	6 (.14)	9 (.21)	4 (.09)	24 (.56)

Die Anpassung des FI-Modells an die Daten aus Tabelle 10.7 ergibt für die Bild-Bild-Bedingung (G^2 = 24.33, *df* = 15) und auch für die Bild-Text-Bedingung (G^2 = 10.85, *df* = 15) zufriedenstellende Modellanpassungsstatistiken, die auf dem 5%-Signifikanzniveau nicht bedeutsam sind. Die ML-Parameterschätzungen und 90%-Konfidenzintervalle sind in Tabelle 10.8 zusammengefaßt. Ganz offensichtlich stellt sich die erwartete Abschwächung der Gedächtnisparameterwerte gegenüber den Experimenten 1 und 2 ein. Dies ist offenbar eine Konsequenz der peripheren Darbietung der Zielitems. Ein Falschinformationseffekt auf die *r*(OI)-Gedächtnisparameter deutet sich nur in der Bild-Text-Bedingung ganz schwach an, wird jedoch weder dort (ΔG^2 = 0.52, *df* = 2) noch in der Bild-Bild-Bedingung signifikant (ΔG^2 = 0.40, *df* = 2). Für die *r*(NI)-Gedächtnisparameter gibt es deskriptiv keinerlei Hinweise auf Falschinformationseffekte. Die Unterschiede zwischen den

Parameterschätzwerten sind insignifikant ($\Delta G^2 = 2.40$, $df = 2$, für die Bild-Bild-Bedingung und $\Delta G^2 = 0.81$, $df = 2$, für die Bild-Text-Bedingung).

Tabelle 10.8
ML-Parameterschätzungen und 90%-Konfidenzintervalle bei Anwendung des FI-Modells auf die Daten beider Versuchsbedingungen des Experiments 3 (vgl. Tabelle 10.7).

Parameter	OI: Bild, NI: Bild			OI: Bild, NI: Text		
		90%-Konfidenzintervall			90%-Konfidenzintervall	
	ML-Schätzwert	untere Grenze	obere Grenze	ML-Schätzwert	untere Grenze	obere Grenze
$r(\mathrm{OI})_k$	.22	.11	.33	.27	.14	.41
$r(\mathrm{OI})_n$	.18	.07	.28	.21	.08	.34
$r(\mathrm{OI})_w$	.23	.12	.34	.20	.07	.33
$r(\mathrm{NI})_k$	.18	.07	.28	.50	.35	.64
$r(\mathrm{NI})_n$	.25	.15	.35	.40	.26	.54
$r(\mathrm{NI})_w$	.32	.21	.42	.49	.35	.63
$g(\mathrm{OI})_{nw}$	.15	.11	.19	.12	.07	.16
$g(\mathrm{OI})_w$	.20	.15	.25	.21	.16	.27
$g_1(\mathrm{NI})_{nw}$	.43	.33	.53	.23	.11	.35
$g_2(\mathrm{NI})_{nw}$	.05	.02	.08	.11	.06	.16
$g_1(\mathrm{NI})_w$	.39	.27	.50	.45	.30	.60
$g_2(\mathrm{NI})_w$	.14	.10	.19	.23	.16	.29

Positiv korrelierte Ratetendenzen bezüglich der OI- und NI-Quellen deuten sich erheblich stärker als in Experiment 1 und Experiment 2 an. Der Test der stochastischen Unabhängigkeit der Ratetendenzen zeigt hier, daß die Nullhypothese für die Bild-Bild-Bedingung ($\Delta G^2 = 56.54$, $df = 2$) und auch für die Bild-Text-Bedingung ($\Delta G^2 = 8.34$, $df = 2$) klar verworfen werden muß.

Die bereits aus den ersten beiden Experimenten bekannten Unterschiede zwischen den Rateparametern bei Darbietung widersprüchlicher bzw. nichtwidersprüchlicher Information in der Lernphase liegen auch in Experiment 3 vor. Erneut ist $g(\mathrm{OI})_w$ größer als $g(\mathrm{OI})_{nw}$, wobei der Effekt in der Bild-Bild-Bedingung nicht signifikant ($\Delta G^2 = 2.08$, $df = 1$), in der Bild-Text-Bedingung ($\Delta G^2 = 4.47$, $df = 1$) dagegen signifikant wird. Der Unterschied zwischen den $g(\mathrm{NI})_w$- und $g(\mathrm{NI})_{nw}$-Parametern wird dagegen sowohl für die Bild-Bild-Bedingung ($\Delta G^2 = 9.45$, $df = 2$) als auch für die Bild-Text-Bedingung signifikant ($\Delta G^2 = 8.89$, $df = 2$).

Offenbar entspricht das Ergebnismuster des dritten Experiments qualitativ dem der ersten beiden Experimente, und zwar für beide Versuchsgruppen. Damit läßt sich die Hypothese, daß Gedächtnisbeeinträchtigungen durch periphere

Darbietung der kritischen Items oder durch visuelle Darbietung der nachträglichen Falschinformation wahrscheinlicher werden, klar zurückweisen.

Der Effekt der Nachinformationsmodalität — Text- versus Bild — ist eindeutig: Die *r*(NI)-Parameter sind signifikant größer, wenn die Nachinformation in Textform statt als Bild dargeboten wird ($\Delta G^2 = 11.37$, $df = 3$). Die *r*(OI)-Parameter unterscheiden sich erwartungsgemäß nicht in Abhängigkeit von der Art der Nachinformation ($\Delta G^2 = 0.49$, $df = 3$) und auch der Effekt auf die Rateparameter ist nicht signifikant ($\Delta G^2 = 10.79$, $df = 6$). Damit ist klar gezeigt, daß der Gedächtnisvorteil der Nachinformationsitems, der in den ersten beiden Experimenten beobachtet wurde, wesentlich mit der Darbietungsmodalität zu tun hat und nicht auf das kürzere Retentionsintervall im Vergleich zur Originalinformation zurückzuführen ist. Signifikant wird der Unterschied zwischen korrespondierenden *r*(OI)- und *r*(NI)-Parametern nur, wenn die Nachinformation als Text dargeboten wird ($\Delta G^2 = 10.20$, $df = 3$). Bei bildlicher Darbietung der Nachinformation verschwindet der Effekt ($\Delta G^2 = 1.75$, $df = 3$).

Zusammenfassend bleibt festzuhalten, daß Experiment 3, in dem zwei Maßnahmen zur Erhöhung der Durchschlagskraft von Falschinformationseffekten ergriffen wurden, ebenfalls keine Hinweise auf Gedächtnisbeeinträchtigungen durch nachträgliche Falschinformation erbracht hat. Lediglich Ratetendenzen werden davon beeinflußt, ob in der betreffenden Quelle eine ähnliche Information enthalten ist, die der gleichen semantischen Kategorie wie das Testitem angehört, oder ob dies nicht der Fall ist. Wie gut Original- und Nachinformationsitems erinnert werden, scheint nach den vorliegenden Befunden nur davon abzuhängen, in welcher Modalität die betreffende Information dargeboten wird: In Textform dargebotenen Informationen werden besser erinnert als in Bildform dargebotene Informationen. Offen bleibt vorerst, ob dies auf die bessere Enkodierbarkeit verbal dargebotener Informationen zurückzuführen ist, oder ob lediglich die Abrufbarkeit verbal dargebotener Information im Kontext eines verbalen Quellendiskriminationstests besser ist.

10.3 Zusammenfassung

In diesem Kapitel wurden die von Batchelder et al. (1990) und Wagenaar und Boer (1987) vorgeschlagenen MVB-Modelle zum Falschinformationseffekt diskutiert und zwei neue Modelle als Alternative vorgeschlagen. Das Modell von Batchelder et al. (1990) ist konzeptuell einfach und bezieht sich auf den relativ häufig verwendeten Quellendiskriminationstest, krankt aber an schlechten Anpassungsstatistiken. Die Modelle von Wagenaar und Boer (1987) passen zwar auf die Daten, beziehen sich aber auf ein relativ unübliches Testformat (Rekognitionstest mit anschließendem 2AFC-Test für Objektdetails). Die positive Evidenz für das reine

Antworttendenzmodell sensu McCloskey und Zaragoza (1985) wird durch die Teststärkeproblematik in dieser Untersuchung eingeschränkt: Selbst wenn Gedächtnisbeeinträchtigungseffekte in der Untersuchung von Wagenaar und Boer (1987) vorgelegen haben sollten, hatte der durchgeführte Test ihres Antworttendenzmodells (no conflict theory) kaum eine Chance, diese Effekte aufzudecken.

Die beiden neu vorgeschlagenen Modelle beziehen sich auf häufig verwendete Gedächtnistests (2AFC-Test und Quellendiskriminationstest), verlangen aber recht komplexe Versuchsdesigns, um identifizierbar und testbar zu sein. Sowohl das Modell zum Loftus-2AFC-Test (Abschnitt 10.1) als auch das FI-Modell zum Quellendiskriminationstest (Abschnitt 10.2) passen gut auf die Daten. Die Bedeutung von Ratetendenzen für Falschinformationseffekte spiegelt sich in den Parametern dieser beiden Modellen ebenso wider wie in den Modellen von Wagenaar und Boer (1987). Unterschiedliche Ergebnisse ergeben sich jedoch in bezug auf die Bedeutung von Gedächtnisbeeinträchtigungen durch nachträgliche Falschinformation: Während die Reanalyse der Daten von Experiment 3 aus Loftus et al. (1978) Evidenz für Gedächtnisbeeinträchtigungen erbrachte (Abschnitt 10.1), gilt das für fünf Datensätze aus drei Experimenten zum FI-Modell nicht (Abschnitte 10.2.1 - 10.2.3). Die Teststärke zur Aufdeckung kleiner Gedächtnisbeeinträchtigungseffekte war in den Experimenten zum FI-Modell zwar nicht voll befriedigend, aber doch deutlich besser als bei den Untersuchungen von Wagenaar und Boer (1987) und Loftus et al. (1978). Außerdem weisen die Gedächtnisparameterschätzungen für die Mehrzahl der Datensätze zum FI-Modell nicht einmal deskriptiv in die richtige Richtung: Ein Falschinformationseffekt deutete sich lediglich in zwei von fünf Datensätzen, ein umgekehrter Falschinformationseffekt (schlechtere Erinnerung der Nachinformation bei widersprüchlicher Originalinformation) nur in einem von fünf Datensätzen an (vgl. die Tabellen 10.4, 10.6 und 10.8). Wenn wirklich ein reliabler Gedächtnisbeeinträchtigungseffekt vorliegen würde, ist ein solches Ergebnismuster recht unwahrscheinlich.

Wie ist die Diskrepanz zwischen den Ergebnissen der Abschnitte 10.1 und 10.2 zu erklären? Das Retentionsintervall wäre ein denkbarer Erklärungsansatz, weil es in Experiment 3 von Loftus et al. (1978) bis hin zu einer Woche verlängert wurde, während es in den Experimenten des Abschnittes 10.2 konstant auf ca. 15-20 Minuten gehalten wurde. Gegen die Bedeutung des Retentionsintervalls spricht jedoch, daß der Gedächtnisbeeinträchtigungsparameter des Modells zum 2AFC-Test über alle Retentionsintervalle hinweg gleichgesetzt werden konnte (vgl. Abschnitt 10.1). Die Stärke des Effekts scheint also nicht von Zeitparametern der Untersuchung abzuhängen.

Eine alternative Erklärung lautet, daß Gedächtnisbeeinträchtigungseffekte durch den 2AFC-Rekognitionstest von Loftus et al. (1978) eher erfaßt werden können als durch andere Gedächtnistests, beispielsweise den Quellendiskriminations-

test. Wenn der Gedächtnisbeeinträchtigungseffekt im wesentlichen ein Abrufbeeinträchtigungseffekt ist — hervorgerufen durch Konkurrenz zwischen dem Abruf der OI und dem Abruf der NI —, dann ist es durchaus verständlich, daß der Effekt im 2AFC-Test mit OI und NI als *gleichzeitig* dargebotenen Antwortalternativen auftritt, im Quellendiskriminationstest mit *isoliert* dargebotenen Testitems dagegen nicht. Diese Interpretation wäre mit der sogenannten Blocking-Hypothese zum Falschinformationseffekt sehr gut vereinbar (Belli, 1993; Chandler, 1991; Roediger, Jacoby & McDermott, 1996; Schooler, Foster & Loftus, 1988; Schreiber & Sergant, 1998).

Eine weitere Erklärungsmöglichkeit zielt auf die exakte Definition der Gedächtnisparameter in den Modellen zum 2AFC-Rekognitionstest und zum Quellendiskriminationstest ab. Im 2AFC-Modell ist der Gedächtnisparameter die Fähigkeit, zwischen zwei zur gleichen Oberkategorie gehörenden Details gedächtnisbasiert diskriminieren zu können. Im FI-Modell zum Quellendiskriminationstest sind die Gedächtnisparameter Wahrscheinlichkeiten des Abrufs von *Oberkategorie* und *Detail* eines Items (z.B. roter Aktenordner) in einer bestimmten Quelle, wobei es für die korrekte Quellenzuordnung in vielen Fällen ausreicht, nur die Oberkategorie zu erinnern. Nimmt man an, daß Gedächtnisbeeinträchtigungseffekte in erster Linie Details eines bestimmten Objekts betreffen, nicht aber die Erinnerung der Objektkategorie selbst, dann wird verständlich, warum der Gedächtnisparameter des 2AFC-Modells sensitiver hierfür ist als die Gedächtnisparameter des FI-Modells.

Die beiden zuletzt genannten Erklärungsmöglichkeiten sind plausibel. Ein Weg, zwischen beiden Erklärungsmöglichkeiten zu unterscheiden, könnte darin bestehen, ein alternatives Modell zum Falschinformationseffekt zu entwickeln, das Kategoriegedächtnis und Detailgedächtnis als distinkte Parameter vorsieht. Wenn ein solches Modell auf die Daten paßt und sich Effekte nachträglicher Falschinformationen zwar auf das Detailgedächtnis, nicht aber auf das Kategoriegedächtnis einstellen, würde das klar die zuletzt genannte Erklärung stützen. Bleiben dagegen Falschinformationseffekte auf beide Klassen von Gedächtnisparametern aus, dann stützt dies eher die Blocking-Hypothese.

Die Entwicklung eines FI-Modells mit unterschiedlichen Gedächtnisparametern für das Kategorie- und das Detailgedächtnis wäre ein lohnenswerter Ansatz für die zukünftige Forschung zum Falschinformationseffekt.

11. Multinomiale Modellierung: Gegenwart und Zukunft

„Psychology can be separated into two camps, a substantive camp that is primarily interested in understanding important aspects of human behavior and thought and a methodological camp that is primarily interested in developing tools that will be used by the substantive researchers to answer difficult questions" (Zickar, 1998, S. 116). Auch mehr als 100 Jahre nach G. T. Fechner und fast 80 Jahre nach Publikation der Arbeiten von L. L. Thurstone ist die Grenze zwischen Substanzwissenschaft und quantitativer Modellbildung in der Psychologie — anders als in den klassischen Naturwissenschaften — immer noch deutlich erkennbar; in einigen Bereichen scheint sie geradezu undurchdringlich. Dies mag seine Ursache teilweise darin haben, daß stochastische Modellbildung — gerade im Bereich der Psychometrie — in der Vergangenheit oftmals ohne Bezugnahme auf substanzwissenschaftliche fundierte Fragestellungen und Theorien betrieben wurde (Roskam, 1985). Umgekehrt hat sich die substanzwissenschaftlich orientierte Psychologie zu lange mit der Forschungsstrategie der Nullhypothesenverwerfung im Rahmen varianzanalytischer Versuchspläne zufriedengegeben, ein Weg, der progressive Forschungsprogramme im Sinne von Lakatos (1970) eher behindert als fördert (Bredenkamp, 1972; Gigerenzer, 1998; Gigerenzer et al., 1999; Meehl, 1967). Die vorstehenden Kapitel sollten an vielen Beispielen deutlich gemacht haben, daß die Grenze zwischen Substanzwissenschaft und quantitativer Modellbildung prinzipiell überwindbar ist: Stochastische Modellbildung muß in substanzwissenschaftlicher Hinsicht nicht zwangsläufig neutral oder „atheoretisch" sein und substanzwissenschaftliche Forschung muß nicht immer in eine Varianzanalyse oder einen *t*-Test mit einem Ad-hoc-Index als abhängiger Variable einmünden. Gerade die Familie multinomialer Verarbeitungsbaummodelle ist so flexibel, daß sie eine Vielzahl psychologisch interessanter Fragestellungen und Hypothesen abdecken kann. Insbesondere die kognitionspsychologische Forschung kann hiervon profitieren, weil sich in diesem Bereich das Problem der adäquaten Erfassung latenter psychologischer Zustände und Prozesse besonders massiv stellt (vgl. Kapitel 2).

Im Vorwort der vorliegenden Arbeit wurde die These aufgestellt, daß substanzwissenschaftlich begründete multinomiale Modelle die Überwindung des „Jochs der precision-importance trade-off function" (Schwarz, 1991) in der kognitiven Psychologie immerhin denkbar erscheinen lassen. Kann diese These auf dem Hintergrund der in den vorstehenden Kapiteln berichteten Ergebnisse zu den Grundlagen und zur Anwendung multinomialer Modellierungen in der kognitiven Psychologie aufrechterhalten werden? Wenn wir uns zunächst dem Präzisionsanspruch zuwenden, so muß man eingestehen, daß multinomiale Modelle deutlich hinter dem Formalisierungsgrad von Meßmodellen zurückbleiben, die aus der sogenannten Repräsentationstheorie des Messens (Krantz, Luce, Suppes &

Tversky, 1971; Luce, Krantz, Suppes & Tversky, 1990; Suppes, Krantz, Luce & Tversky, 1989; Suppes & Zinnes, 1963) erwachsen sind. Axiomatisierungen multinomialer Modelle bzw. Repräsentationssätze, die notwendige und hinreichende Bedingungen für die Gültigkeit bestimmter Modelle benennen, existieren nicht. Zwar dürfte es in einer Reihe von Fällen leicht möglich sein, Repräsentationssätze auf der Grundlage bestimmter Relationen zwischen den Kategoriewahrscheinlichkeiten zu formulieren, die die Gültigkeit eines bestimmten Modells garantieren. So ist es etwa notwendig und hinreichend für die Gültigkeit des 1HT-Modells der Rekognition, wenn Treffer- und Falsche-Alarm-Wahrscheinlichkeiten in einer linearen Beziehung zueinander stehen, derart, daß die Gerade im ROC-Diagramm durch den Punkt (1, 1) geht. Aber das ist nicht die Art der Axiomatisierung, auf die die abstrakte Meßtheorie abzielt. Vielmehr interessieren Axiomatisierung in Termini qualitativer Axiome, deren empirische Primitiva direkt bestimmten Verhaltensdaten korrespondieren. Entsprechende Ansätze sind nach meiner Kenntnis für multinomiale Modelle nicht in Sicht.

Auf der anderen Seite hat die abstrakte Meßtheorie auch Kritik auf sich gezogen, weil sie in der psychologischen Forschungspraxis bislang weitgehend konsequenzenlos geblieben ist (z.B. Cliff, 1992). Dies wiederum dürfte daran liegen, daß sie bislang eine rein theoretische Disziplin war, welche die Beziehung zwischen Theorie und empirischen Daten stark vernachlässigt hat (Estes, 1975). Wie z.B. Schönemann (1994) moniert, wurden meßtheoretische Axiomensysteme teilweise komplett losgelöst von Beobachtungen bzw. vorliegenden empirischen Befunden entwickelt. Tatsächlich haben sich die wenigen Fälle, bei denen eine Überbrückung zwischen Meßtheorie und empirischen Daten auf der Grundlage stochastischer Axiomensysteme immerhin versucht wurde (z.B. Falmagne, Iverson & Marcovici, 1979), letztlich als empirische Fehlschläge erwiesen (Gigerenzer & Strube, 1983).

Multinomiale Modelle haben einen wesentlich pragmatischeren Anspruch als die Modelle der abstrakten Meßtheorie. Nicht die Gültigkeit verschiedener qualitativer Axiome, die zusammengenommen notwendig und hinreichend für die Modellgültigkeit sind, ist die *conditio sine qua non* einer erfolgreichen Modellanwendung, sondern die approximative Gültigkeit der Modellgeltungshypothese (vgl. Abschnitt 4.3). Dieses schwächere Kriterium der empirischen Adäquatheit läßt sich — wie in den Kapiteln 6 bis 10 an vielen Beispielen gezeigt wurde — ganz offensichtlich nicht nur auf zahlreiche kognitionspsychologische Forschungsfelder erfolgreich anwenden, sondern auch auf Forschungsfelder aus benachbarten Bereichen wie der Wahrnehmungspsychologie, der kognitiven Sozial- und Persönlichkeitspsychologie, der Entwicklungspsychologie, Psychometrie usw. Ein Verlust an Präzision ist mit diesem Gewinn an Anwendbarkeit nur insofern verknüpft, als im Falle des Scheiterns einer Modellgeltungshypothese keine Aussage darüber möglich ist, welche Annahme oder welche Annahmen des Modells für das empi-

rische Scheitern verantwortlich sind. Wir wissen dann lediglich, daß mindestens eine Annahme des Modells falsch sein muß. Lediglich Residuenbetrachtungen — d.h. informelle Analysen der Beziehungen zwischen beobachteten und bei Modellgültigkeit zu erwartenden Kategoriehäufigkeiten — können Vermutungen darüber nahelegen, welche Annahmen des Modells für die Anpassungsprobleme verantwortlich sein könnten. Insofern kommen multinomiale Modelle dem Ziel der Befreiung vom „Joch der precision-importance trade-off function" (Schwarz, 1991) nahe, wenn sie es vielleicht auch nicht voll erreichen.

In den folgenden Abschnitten sollen Aspekte des multinomialen Modellierens in der kognitiven Psychologie besprochen werden, die die konkrete Handhabung dieses Ansatzes (Abschnitt 11.1), seinen Nutzen (Abschnitt 11.2), seine Grenzen (Abschnitt 11.3) und vorhersehbare zukünftige Entwicklungen (Abschnitt 11.4) betreffen. Technische Details und konkrete Anwendungsbeispiele bleiben hierbei ausgeklammert, denn diese waren ja bereits Gegenstand der vorstehenden Kapitel. Eine Zusammenfassung schließt das Kapitel ab (Abschnitt 11.5).

11.1 Die Kunst des multinomialen Modellierens

Gerade weil die breite Anwendbarkeit multinomialer Modelle in der kognitiven Psychologie in Anbetracht der vorliegenden Ergebnisse und Veröffentlichungen kaum bezweifelt werden kann (vgl. vor allem Kapitel 7), mag es naheliegend erscheinen, sie einfach als alternative Datenauswertungsmethode zu betrachten, die ebenso routinemäßig auf kognitionspsychologische Daten angewandt werden kann, wie beispielsweise varianzanalytische Modelle oder log-lineare Modelle. In meinen Augen wäre dies eine ganz und gar verkehrte Sicht der Sachlage. Multinomiale Modellbildung ist immer ein kreativer Akt — eine Kunst, wenn man so will — ganz einfach weil die Modellgleichungen und oftmals auch das Kategoriensystem, auf das sie sich beziehen, nicht a priori feststehen, sondern erst theorie- oder hypothesenbezogen konstruiert werden müssen. Die multinomiale Modellierung ist Bestandteil der Theorieexplikation und gehört daher sachlogisch zum Theorieteil empirischer Arbeiten, nicht zum Auswertungsteil.

Obwohl kreative Akte dem Prinzip nach nicht normierbar sind, lassen sich doch einige Leitlinien angeben, deren Beachtung nach meinen Erfahrungen erfolgreiches multinomiales Modellieren wahrscheinlicher macht. Die erste Leitlinie betrifft den Ausgangspunkt beim multinomialen Modellieren. Der Ausgangspunkt der Modellbildung sollte nicht die Explikation theoretischer Annahmen in Form eines Ereignisbaumes sein, sondern die geeignete Auswahl eines Systems von Beobachtungskategorien (vgl. auch Batchelder & Riefer, 1999). In Abhängigkeit von einem (vorläufigen) System von Beobachtungskategorien ist dann zu überlegen, welche der interessierenden kognitiven Prozesse in diese Kategorien

einmünden können. Sind damit alle interessierenden Prozesse abgedeckt, kann man die Konstruktion eines multinomialen Modells versuchen. Spiegeln sich vermutlich nicht alle interessierenden Prozesse in den betreffenden Beobachtungskategorien wider oder sind mehr relevante Prozeßkomponenten vorhanden als Beobachtungskategorien, so muß das vorläufige System der Beobachtungskategorien modifiziert werden. Danach ist erneut die Frage zu stellen, ob sich alle interessierenden Prozesse mit dem System identifizieren lassen usw. Durch diesen spiralförmigen Prozeß sollte sich letztlich ein System von Beobachtungskategorien und ein dazugehöriges Modell ergeben, welches der substanzwissenschaftlichen Fragestellung oder Theorie sehr nahe kommt.

11.1.1 Heuristiken der Auswahl von Beobachtungskategorien

Bei der Auswahl von Beobachtungskategorien ist natürlich in erster Linie zu berücksichtigen, daß sie die interessierenden theoretischen Konstrukte reflektieren müssen, wenn auch nicht zwangsläufig im Sinne „konstruktreiner" Indikatoren. Man muß in das Kategoriensystem daher zunächst solche Beobachtungskategorien aufnehmen, die bestimmte Konstrukte prototypisch repräsentieren. Ist man z.B. an der Entwicklung eines Meßmodells zur Cluster-Speicherung und zum Cluster-Abruf für freie Reproduktionsdaten interessiert, so liegt es nahe, u.a. die Beobachtungskategorien „*x* Wörter eines Clusters werden unmittelbar nacheinander reproduziert" vorzusehen, wobei *x* von 2 bis zur maximalen Kategoriegröße der Lernliste laufen kann. Die intuitive Idee dahinter ist, daß sich Unterschiede in Cluster-Speicherungs- und Cluster-Abrufprozessen vermutlich in unterschiedlichen Häufigkeiten der genannten Beobachtungskategorien niederschlagen. Dies wäre beispielsweise kaum oder gar nicht der Fall, wenn man einfach die Anzahl der pro Kategorie reproduzierten Wörter (unabhängig von ihrer Reihenfolge im Reproduktionsprotokoll) als Kategoriesystem wählen wurde.

Oftmals ist man bei der Suche nach derartigen „Kernkategorien" gut beraten, wenn man zunächst nachschaut, auf welche Beobachtungskategorien die in der relevanten Literatur üblichen Ad-hoc-Maße zur Erfassung der interessierenden Konstrukte Bezug nehmen. Für die Cluster-Speicherung und den Cluster-Abruf wäre etwa das sogenannte „adjusted ratio of clustering" (ARC) relevant bzw. das Kategoriensystem, das der ARC-Wert implizit voraussetzt (zum ARC-Maß vgl. z.B. Puff, 1982). Tatsächlich sind sehr viele Kategoriensysteme, die im Bereich der gedächtnispsychologisch motivierten multinomialen Modellierung benutzt werden, mit jenen identisch, die auch die gängigen Ad-hoc-Maße psychologischer Konstrukte voraussetzen (vgl. etwa Batchelder & Riefer, 1980; 1990). Gelegentlich kann sogar der Fall eintreten, daß ein gebräuchliches Ad-hoc-Maß mit einem auf der Basis eines multinomialen Modells abgeleiteten ML-Parameterschätzer zum gleichen Ka-

tegoriesystem zufällig übereinstimmt bzw. durch eine einfache Transformation in einen solchen ML-Schätzer überführt werden kann. Dies ist z.B. bei dem als „Cohens Kappa" bekannten Ad-hoc-Maß für Beurteilerübereinstimmung der Fall, das einem Parameter des Übereinstimmungsmodells von Klauer und Batchelder (1996) bis auf eine einfache Quadratwurzeltransformation entspricht. In solchen Fällen liefert das betreffende multinomiale Modell — sei es nun intendiert oder auch nicht — nachträglich eine theoretische Begründung für die Verwendung der betreffenden Ad-hoc-Statistik.

Ein zweiter Aspekt, der bei der Konstruktion eines Kategoriensystems zu berücksichtigen ist, ist die Anzahl der Beobachtungskategorien. Liegen J Beobachtungskategorien vor, so kann ein identifizierbares Modell maximal J–1 Parameter umfassen. Ist von vornherein klar, daß mehr als J–1 Parameter benötigt werden, um die hypothetisch angenommen kognitiven Prozesse zu separieren, so reicht die Kategorieanzahl J offenbar nicht aus. Eine günstige Situation ist gegeben, wenn man durch Zerlegung der Menge möglicher Beobachtungen J so weit erhöhen kann, daß das Modell potentiell identifizierbar wird. Hierbei ist allerdings zu berücksichtigen, daß die bloße Erhöhung der Anzahl der Beobachtungskategorien dann unergiebig ist, wenn sie die Aufnahme neuer Parameter in das Modell erzwingt, die keinen zusätzlichen Erkenntnisgewinn bringen (beispielsweise Rateparameter für zusätzliche Antwortkategorien). Der Übergang von dichotomen Ja-Nein-Tests zu mehrstufigen Ratingskalen ist deshalb häufig unergiebig. Eine bessere Idee ist es oftmals, den Probanden sukzessive verschiedene Tests vorzulegen, die unterschiedlich sensitiv für die interessierenden Prozesse sind (vgl. z.B. Riefer & Rouder, 1992; Riefer & LaMay, 1998; Rouder & Batchelder, 1998). Bei der Modellkonstruktion sind dabei natürlich mögliche Einflüsse früherer Testantworten auf spätere angemessen zu berücksichtigen. Weitere Heuristiken zur Erhöhung von J wurden von Batchelder und Riefer (1999, S. 74f) vorgeschlagen und diskutiert.

Ergibt sich keine naheliegende Möglichkeit der Erhöhung von J, so ist der Übergang zu verbundenen multinomialen Modellen oft die beste Problemlösungsstrategie. Statt ein Beobachtungssystem mit J Kategorien nur in einer Versuchsbedingung zu untersuchen, untersucht man es z.B. unter zwei verschiedenen Versuchsbedingungen und erhält so $2 \cdot J$ Beobachtungskategorien. Dies erlaubt bereits identifizierbare Modelle mit potentiell $2 \cdot (J–1)$ Parametern. Besonders effektiv ist dieses Vorgehen, wenn die experimentelle Manipulation, welche die beiden Versuchsbedingungen voneinander unterscheidet, selektiv nur einen kognitiven Prozeß beeinflußt. Bei Modellen, die Rate- bzw. Antworttendenzprozesse annehmen — ein Standardfall im Bereich der multinomialen Modellierung kognitiver Prozesse —, kann der Übergang zu einem identifizierbaren verbundenen multinomialen Modell oftmals durch eine zweistufige Manipulation von Antworttendenzen erreicht werden. Dieses Vorgehen wurde u.a. in Kapitel 9 der vor-

liegenden Arbeit praktiziert und auch von Klauer et al. (1999) erfolgreich angewendet. Reicht der Übergang auf $K = 2$ Versuchsbedingungen nicht aus, so kann man K erhöhen oder alternativ auf mehrfaktorielle Versuchspläne mit $K_1 \cdot K_2$ Versuchsbedingungen übergehen, wobei idealerweise jeder Faktor selektiv einen kognitiven Prozeß beeinflussen sollte (vgl. z.B. Batchelder & Crowther, 1997; Gleißner & Erdfelder, 1994).

11.1.2 Heuristiken der Modellkonstruktion

Ist ein vermutlich adäquates System von Beobachtungskategorien verfügbar, kann man sich der eigentlichen Modellkonstruktion zuwenden. Zunächst ist auf theoretischer Ebene zu klären, wieviel und welche psychologischen Prozesse auf die Wahrscheinlichkeiten der Beobachtungskategorien Einfluß nehmen können. Hierbei ergeben sich in der Regel unterschiedliche Situationen in Abhängigkeit davon, ob das Modell zwecks Formalisierung einer psychologischen Theorie oder lediglich als Meßmodell für bestimmte theoretische Größen konstruiert werden soll[105]. Wenn das multinomiale Modell primär als Meßmodell für einige wenige Konstrukte dienen soll, muß man mit der Möglichkeit rechnen, daß im Modell mehr kognitive Prozesse als die eigentlich interessierenden zu berücksichtigen sind. Oftmals sind beispielsweise Rateprozesse psychologisch von untergeordneter Bedeutung, müssen aber dennoch in das Modell aufgenommen werden, um unkontaminierte Messungen der eigentlich interessierenden theoretischen Größen zu ermöglichen (vgl. z.B. Buchner et al., 1995; Erdfelder & Buchner, 1998a, 1998b). Wird die Modellkonstruktion dagegen als Formalisierung einer psychologischen Theorie verstanden, so muß oftmals konstatiert werden, daß nur ein Teil der relevanten theoretischen Prozesse im Modell berücksichtigt werden kann. Diese Restriktion ergibt sich aus den Grenzen, die das System der Beobachtungskategorien der Modellbildung auferlegt. Ich werde im folgenden Abschnitt hierauf zurückkommen.

Als zweites ist zu klären, ob Finite-State-Konzeptionen für alle kognitiven Prozesse des Modells adäquat sind, d.h. ob sie sinnvoll als diskrete latente Zufallsvariablen modelliert werden können, deren Ausprägungen latente Wahrscheinlichkeiten zugeordnet sind. Diese Frage läßt sich häufig auch dann bejahen, wenn die theoretisch angenommenen Prozesse eigentlich kontinuierlicher Natur sind. Drei Gründe können dafür sprechen, dennoch eine Finite-State-Konzeption anzuwenden: a) Threshold-Konzeptionen, b) die approximative Gültigkeit einer Fi-

[105] Wie ein Blick in Kapitel 7 der vorliegenden Arbeit zeigt, lassen sich fast alle Anwendungen multinomialer Modelle in der kognitiven Psychologie unter diese beiden Schwerpunkte subsumieren (vgl. auch Batchelder & Riefer, 1999).

nite-State-Konzeption oder c) die Äquivalenz eines kontinuierlichen Prozeßmodells mit einem Finite-State-Modell.

Ad a). Der Grundgedanke von Threshold-Konzeption ist, daß die eigentlich zugrundeliegenden Prozesse kontinuierlich sind, die verhaltenswirksamen Konsequenzen dieser Prozesse jedoch nur von der Überschreitung oder Nichtüberschreitung gewisser „Schwellenwerte" (thresholds) abhängen. Auf derartige Threshold-Konzeptionen wurde in mehreren Kapiteln dieser Arbeit zurückgegriffen. In einem gewissen Sinne transformieren Threshold-Konzeptionen stetige latente Variablen in diskrete, so daß Finite-State-Konzeptionen anwendbar werden.

Ad b). Eine zweite Möglichkeit, die diskrete Modellierung kontinuierlicher latenter Zufallsvariablen zu begründen, ist die *approximative* Gültigkeit der Finite-State-Konzeption. Selbst wenn eine bestimmte kognitive Fähigkeit in einer Population stetig verteilt ist, kann eine diskrete Variable, die nur zwischen hoher und niedriger Ausprägung der Fähigkeit unterscheidet, die stetige Variable doch immerhin approximieren. Möglicherweise gelingt dies im Rahmen einer Finite-State-Konzeption sogar besser als im Rahmen eines Modells mit stetigen Verteilungsfunktionen der latenten Variablen. Im Rahmen der Modellierung kontinuierlicher Prozesse ist man natürlich auch auf Verteilungsannahmen angewiesen — häufig Normalverteilungsannahmen und Varianzhomogenitätsannahmen —, und diese Annahmen sind im Regelfall auch nur als Approximationen an die wahre Verteilung aufzufassen, ebenso wie diskrete Verteilungsmodelle.

Ad c). Schließlich können diskrete und kontinuierliche Modelle trotz des konzeptuellen Unterschieds datenäquivalent sein, wenn man bestimmte Verteilungsannahmen mit dem kontinuierlichen Modell verbindet. Ein Beispiel hierfür wurde in Abschnitt 6.7 bereits besprochen (vgl. auch Macmillan & Creelman, 1991).

Obwohl immer wieder in der kognitiven Psychologie heiß diskutiert (z.B. Mulder, Sanders & van Galen, 1995), wird man sicher kein pauschales Urteil über die Angemessenheit oder Unangemessenheit von Finite-State-Konzeption fällen können. Die Frage nach der Angemessenheit ist in jedem Anwendungskontext neu zu stellen und zu beantworten, wobei gegebenenfalls auch empirische Vergleiche diskreter und kontinuierlicher Modellierungsansätze weiterhelfen können (vgl. z.B. Erdfelder & Buchner, 1998b; Garcia-Pèrez, 1990; Snodgrass & Corwin, 1988; Kapitel 9 der vorliegenden Arbeit).

Obwohl Finite-State-Konzeptionen einen MVB-Modellierungsansatz nicht unbedingt erzwingen — es gibt Finite-State-Modelle, die nicht der MVB-Modellgleichungsstruktur genügen —, so ist ein MVB-Modellierungsansatz doch fast immer lohnend, wenn theoretische Argumente nicht andere Modellierungsansätze nahelegen. Eine gute Heuristik zur Konstruktion eines Verarbeitungsbaummodells ist die *zeitliche Sequenz* möglicher kognitiver Prozesse, die zwischen Stimulusdarbietung und Erhebung der Vp-Reaktion ablaufen. Der Weg vom Stamm

bis hinein in eine der Astspitzen eines solchen Modells wäre dann jeweils eine zeitliche Prozeßsequenz, die denkbar ist. Sehr viele der in diese Arbeit besprochenen Modelle folgen dieser Struktur. Oftmals ergibt sich die zeitliche Sequenz auch ganz automatisch, wenn man etwa mehrere Beobachtungsvariablen modelliert, die in einer bestimmten Reihenfolge erhoben wurden. Das Verarbeitungsbaummodell, das sich auf diese Variablen bezieht, folgt dann sinnvollerweise der gleichen zeitlichen Sequenz wie die erhobenen Daten (z.B. Erdfelder & Buchner, 1998a, Riefer & LaMay, 1998; Riefer & Rouder, 1992; Rouder & Batchelder, 1998).

Andere Heuristiken sind denkbar. Beispielsweise kann man ein Verarbeitungsbaummodell mit den Zustandsunterscheidungen beginnen lassen, deren unbedingte Wahrscheinlichkeiten interessieren. Man fährt dann mit den Zuständen fort, deren bedingte Wahrscheinlichkeiten erster Ordnung interessieren usw. Im Prozeßdissoziationsmodell von Buchner et al. (1995, vgl. Abschnitt 7.1.3) wurde der Prozeß bewußter Erinnerung beispielsweise auf der ersten und der Vertrautheitsprozeß auf der zweiten Ebene des Verarbeitungsbaumes lokalisiert, weil die *unbedingte* Wahrscheinlichkeit bewußter Erinnerung und die *bedingte* Wahrscheinlichkeit von Vertrautheit ohne bewußte Erinnerung von Interesse war. Ein zeitliches Aufeinanderfolgen beider Prozesse sollte damit nicht ausgedrückt werden. Hält man also vertauschte bedingte und unbedingte Wahrscheinlichkeiten für die beiden Prozesse für geeigneter, so spricht überhaupt nichts dagegen, ihre Reihenfolge im Ereignisbaum umzudrehen. Die beiden Modellierungsansätze sind datenäquivalent (Buchner & Erdfelder, 1996).

Ein Standardproblem, das beim ersten Versuch der Konstruktion eines Verarbeitungsbaummodells häufig entsteht, ist die Inflation von Prozeßkomponenten und damit auch von Parametern, die mit jeder zusätzlichen Ebene im Modell fast zwangsläufig einhergeht. Kann man das System der Beobachtungskategorien nicht entsprechend erweitern, bleibt nur die Möglichkeit, Parametergleichsetzungsrestriktionen vorzunehmen oder bestimmte Prozeßkomponenten mit festen Wahrscheinlichkeiten zu belegen (typischerweise 0 oder 1), so daß keine zusätzlichen freien Parameter im Modell entstehen. Ist eine solche Vorgehensweise unabwendbar, sollten immer zwei Vorsichtsmaßnahmen ergriffen werden: a) Konzipierung von Experimenten, die den Annahmen des restringierten Modells möglichst gut entsprechen sowie b) Exploration der Auswirkung alternativer Parameterfixierungen oder Parametergleichsetzungsrestriktionen auf die anderen Parameter des Modells.

Ad a). Obwohl es prinzipiell wünschenswert ist, Modelle möglichst allgemein zu konzipieren, so daß sie auf möglichst viele experimentelle Paradigmen und experimentelle Randbedingungen anwendbar sind, sind diesem Ziel doch eigentlich immer mehr oder minder enge Grenzen gesetzt. Vor Anwendung eines Modells auf empirische Daten sollte man daher reflektieren, was die kritischen

Annahmen des betreffenden Modells sind. Man sollte Prüfexperimente dann gezielt so konzipieren, daß die kritischen Annahmen möglichst gut erfüllt sind (vgl. auch Ashby et al., 1996, S. 186). Unterläßt man dies, so kann die Verwerfung der Modellgeltungshypothese eigentlich nicht überraschen, denn das Modell wurde in einer Situation geprüft, die mit den Modellannahmen nicht kompatibel ist. Konkret kann die Anpassung eines Experiments auf die Annahmen eines Modells beispielsweise durch geeignete Selektion des Versuchsmaterials oder durch geeignete Versuchsinstruktionen geschehen. Wenn beispielsweise die Annahme erforderlich ist, daß bestimmte Ratewahrscheinlichkeiten null sind (vgl. z.B. Riefer & Batchelder, 1995), kann man die Vpn instruieren, im Zustand der Unsicherheit immer eine Nein- oder Neu-Antwort zu geben, statt zu raten. Ein ähnliches Beispiel für die Anpassung einer Vpn-Instruktion an die Annahmen eines Modells geben Kliegl und Lindenberger (1992).

Ad b). Eine Kontrollmaßnahme, die man auch dann nicht unterlassen sollte, wenn das Prüfexperiment auf die Annahmen des Modells hin konzipiert wurde, ist die Exploration der Auswirkungen alternativer Parameterfixierungen oder Parametergleichsetzungsrestriktionen auf die anderen Parameter des Modells. Es ist wichtig zu sehen, ob die substanzwissenschaftlichen Schlußfolgerungen, die man aus den Schätzungen der freien Parameter und den Hypothesentests ableitet, auch dann Bestand haben, wenn man die zwecks Herstellung von Identifizierbarkeit vorgenommenen Parameterrestriktionen ändert. Denkbar ist ja, daß die Maßnahmen zur Anpassung des Experiments an die Modellannahmen nicht oder nur begrenzt wirksam waren, so daß die Parameterrestriktionen nicht adäquat sind. Beispiele für die Exploration der Auswirkungen von Parameterfixierungen geben Riefer und Batchelder (1995) sowie Kohnert (1996). Kohnert (1996) setzte eine Ratewahrscheinlichkeit, die im Hindsight-Bias Modell von Erdfelder (1992; vgl. Abschnitt 7.1.9) auf $h = .50$ fixiert ist, versuchsweise auf andere Werte fest. Sowohl bei Riefer und Batchelder (1995) als auch bei Kohnert (1996) zeigte sich, daß Änderungen der Parameterfixierungen an den inhaltlichen Schlußfolgerungen aus den Daten wenig ändern.

Parameterrestriktionen vorzuziehen sind natürlich immer Erweiterungen des Systems der Beobachtungskategorien, soweit dies aus versuchsplanerischer und praktischer Sicht möglich ist. Grundsätzlich sollte die Modellentwicklung und Modellprüfung von vornherein auf ein Modell hinzielen, das nicht nur identifizierbar, sondern auch testbar ist. Die empirische Evaluation eines Modells sollte ebenfalls nicht auf ein isoliertes Prüfexperiment hin ausgerichtet sein, sondern auf eine komplette Experimentserie, die Evidenz für die psychologische Validität der Modellparameter erbringt (vgl. z.B. Buchner et al., 1995; Erdfelder & Buchner, 1998a; Klauer & Wegener, 1998; Riefer & Batchelder, 1995, vgl. Abschnitt 4.5). Erst wenn sich die zentralen Modellparameter im Einklang mit den Prognosen verhal-

ten und außerdem die Verteilung der Modellanpassungsstatistiken in mehreren unabhängigen Modellanwendungen mit der unter der Nullhypothese der Modellgültigkeit zu erwartenden Verteilung hinreichend kompatibel ist, kann von einer erfolgreichen Modellanwendung gesprochen werden. Schon allein wegen des Problems datenäquivalenter Modelle haben Anwendungen eines Modells auf nur einen isolierten Datensatz im Grunde genommen nur den Status einer Modellillustration, nicht aber den Status eines strengen Prüfversuchs (vgl. Abschnitt 4.4).

11.2 Der Nutzen des multinomialen Modellierens

Worin besteht genau der Nutzen des multinomialen Modellierens für die experimentelle kognitive Psychologie? Kann man das, was multinomiale Modelle liefern, nicht vielleicht doch mit traditionellen Mitteln und Methoden der kognitiven Psychologie auf einfachere Art und Weise erreichen? Zwei Situationen sind bei der Beantwortung dieser Frage erneut zu unterscheiden: a) multinomiale Modelle werden primär als Meßmodelle verwendet und b) multinomiale Modelle dienen der Theorie- bzw. Hypothesenformalisierung. Für beide Situationen soll die Ausgangsfrage im folgenden kurz beantwortet werden.

11.2.1 Multinomiale Modelle als Meßmodelle

Sofern multinomiale Modelle als Meßmodelle eingesetzt werden, besteht ihr primäres Ziel darin, unkontaminierte („prozeßreine") Maße für bestimmte kognitive Prozesse oder Zustände zu liefern (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). Die aus den Meßmodellen und gegebenen Daten abgeleiteten Maße sind statistisch gesehen Parameterschätzungen, typischerweise ML-Schätzwerte. Soweit explizite Lösungen der ML-Schätzgleichungen existieren (z.B. Batchelder & Riefer, 1986), erweisen sich die Schätzwerte häufig als relativ einfache Funktionen der Kategoriehäufigkeiten. Soweit explizite Lösungen nicht existieren, gibt es häufig doch recht einfache Wege, die ML-Schätzwerte über eine bestimmte Rechenformel zu approximieren. Wenn dem so ist, liegt die Frage nahe, ob man multinomiale Modelle als Meßmodelle überhaupt braucht. Der Zweck, für den sie eingesetzt werden, läßt sich oft mit weniger Aufwand erreichen, indem man einfach Indizes aus den Daten errechnet, die den ML-Schätzwerten entsprechen bzw. sie hinreichend genau approximieren.

Dieses Argument greift m.E. aus mehreren Gründen zu kurz. Multinomiale Modellierung ist für die experimentelle kognitionspsychologische Forschung nicht deshalb von Nutzen, weil sie Meßwerte für kognitive Prozesse liefert, die man auf anderem Wege (unter Umgehung des modellbasierten Parameterschätzverfahrens) nicht hätte gewinnen können. Ihr hauptsächlicher Nutzen besteht vielmehr darin, daß sie den Maßen eine Theorie gibt und somit explizit macht, welche An-

nahmen erforderlich sind, um die betreffenden Maße als korrekte Maße des betreffenden Prozesses ausweisen zu können. Dies ist die Voraussetzung dafür, daß die verwendeten Maße wissenschaftlich begründbar bzw. kritisierbar sind. Fehlt ein Modell im Hintergrund, kann man den Maßen vertrauen oder auch nicht. Es gibt keine Möglichkeit, einen der beiden Standpunkte fundiert zu begründen. Die Präferenz für bestimmte Maße wird zu einer Frage des persönlichen Geschmacks, der Zugehörigkeit zu bestimmten experimentalpsychologischen Schulen und der Bereitschaft, sich Mehrheitsmeinungen in der „scientific community" anzuschließen oder auch nicht. Eine wissenschaftliche Begründung für die Verwendung eines Maßes ist ohne zugehöriges Meßmodell nicht zu liefern.

Ein zweites gewichtiges Argument gegen die „modellfreie" Verwendung von Indizes ist die Tatsache, daß es hierbei grundsätzlich unentdeckt bleiben muß, wenn die Annahmen, die den Indizes implizit zugrunde liegen, im konkreten Anwendungsfall verletzt sein sollten. Damit lassen sich die Indizes psychologisch nicht mehr sinnvoll interpretieren (vgl. z.B. die Studie von Henkel et al., 1998; siehe Abschnitt 7.1.2), wovon jedoch ihr Anwender nichts erfährt. Dies ist im Rahmen eines sorgfältig kontrollierten modellbasierten Vorgehens weitgehend ausgeschlossen, weil der Interpretation der Parameterschätzwerte die Durchführung eines Modellanpassungstests vorausgehen wird. Parameterschätzwerte werden nur dann psychologisch interpretiert, wenn die Modellgeltungshypothese beibehalten werden kann (vgl. Abschnitt 7.1.2).

Ein dritter Vorteil modellbasierter Maße ist die quasi automatische Ratekorrektur der Parameterschätzwerte, von der immer dann ausgegangen werden kann, wenn alle relevanten Rateprozesse im Modell berücksichtigt werden. So sind z.B. Schätzwerte für die Wahrscheinlichkeiten r_C und r_E (Abruf des ursprünglichen Urteils bei Kontroll- bzw. Experimentalitems) im Hindsight-Bias-Modell von Erdfelder (1992; Erdfelder & Buchner, 1998a) automatisch ratekorrigiert, was für die Indizes „Prozent korrekter Erinnerungen bei Experimental- und Kontrollitems", die in der einschlägigen Forschung häufig zur Approximation von r_C und r_E herangezogen werden (z.B. Pohl, 1998; Pohl & Hell, 1996; Stahlberg & Maass, 1998), nicht zutrifft (vgl. Abschnitt 7.1.9). Selbst wenn der Effekt der Ratekorrektur bei Reproduktionsdaten im allgemeinen gering ist, kann er in bestimmten Anwendungsfällen doch beträchtlich sein (vgl. etwa Erdfelder & Buchner, 1998a, Experiment 4).

Ein vierter wichtiger Vorteil modellbasierter Maße ist die eindeutige Interpretation, die die Maße durch Einbettung in ein Modell erfahren. Während bei Ad-hoc-Indizes typischerweise kein Gedanke darauf verwendet wird, ob sie latente Mittelwerte oder latente Wahrscheinlichkeiten reflektieren, ob es sich bei latenten Wahrscheinlichkeiten um bedingte oder unbedingte Wahrscheinlichkeiten handelt usw., liegt die korrekte Interpretation der Parameter bei modellbasierten

Maßen unmittelbar auf der Hand. Daß dies kein bloß ästhetischer Aspekt ist, zeigt die Diskussion um die Unabhängigkeitsannahme in Jacobys (1991) Prozeßdissoziationsprozedur (Buchner et al., 1995; Buchner & Erdfelder, 1996; Erdfelder & Buchner, 1995; 1998b; Jacoby, 1991; Joordens & Merikle, 1993; Wainwright & Reingold, 1996; siehe auch Abschnitt 7.1.3). Jacoby (1991) hatte seine Indizes zur Erfassung von bewußter Erinnerung (*R*) und Vertrautheit (*F*) mit einer stochastischen Unabhängigkeitsannahme begründet, die nicht testbar ist. Buchner et al. (1995) haben aufgrund einer modellbasierten Analyse der Indizes gezeigt, daß diese unprüfbare Annahme vermieden werden kann, wenn *F* explizit als *bedingte* Wahrscheinlichkeit eines Vertrautheitsgefühls im Falle fehlender bewußter Erinnerung interpretiert wird. Alles, was über diese Interpretation hinausgeht, ist im Rahmen von Jacobys (1991) Prozeßdissoziationsprozedur empirisch nicht prüfbar. Umfangreichere Paradigmen sind erforderlich, um die Unabhängigkeitsannahme wirklich testen zu können (Erdfelder & Buchner, 1995; Krüger, 1995; Krüger & Vaterrodt-Plünnecke, 1997).

Ein fünftes Argument für modellbasierte Maße ist der heuristische Nutzen, der mit ihnen einhergeht. Zunächst einmal fallen beim modellbasierten Vorgehen nicht nur Schätzwerte für die primär interessierenden Prozesse an, sondern auch andere Schätzwerte, beispielsweise für Rateprozesse. Auch deren Ergebnismuster müssen über verschiedene experimentelle Bedingungen und Experimente hinweg psychologisch Sinn ergeben, sonst wäre das ein Grund, an der Validität des Modells zu zweifeln. Mit anderen Worten: Ein modellbasiertes Vorgehen erzwingt ein breitere, umfassendere Theorie des Antwortverhaltens als Ad-hoc-Indizes und regt damit die Theoriebildung an. Der Lohn hierfür ist die Einsicht in Gesetzmäßigkeiten, über die man ohne ein Meßmodell wahrscheinlich gar nicht nachgedacht hätte. Als Beispiel sei der Parameter *g* für den Quellenratebias in den Experimenten von Erdfelder und Bredenkamp (1998) genannt (vgl. Abschnitt 7.1.6). Obwohl hierzu a priori keine Hypothese formuliert wurde, wies sein Verhalten in zwei Experimenten doch eine erstaunliche Regelmäßigkeit auf: Die Wahrscheinlichkeit, im Zustand der Quellenunsicherheit eine bestimmte Quelle zu raten, entsprach fast exakt dem Anteil aller Testitems in dieser Quelle. Es scheint sich also um ein Beispiel für Wahrscheinlichkeitslernen zu handeln, ein Phänomen, das in der Lernpsychologie bereits intensiv untersucht wurde (zusammenfassend z.B. Jones, 1971).

Eine weitere heuristische Funktion modellbasierter Maße besteht darin, daß man auf bestimmte Maße oder spezielle Optimierungen von Indizes überhaupt erst durch den Versuch einer Modellbildung kommt. Wenn ein Modellversuch im ersten Ansatz scheitert, wird man Modellrevisionen versuchen, die mit modifizierten — möglicherweise verbesserten — Maßen der relevanten kognitiven Prozesse einhergehen (s. z.B. Ashby et al. 1996, S. 188). Bei der Verwendung von

Ad-hoc-Indizes, die ohne das Korrektiv des Modellgeltungstests angewandt werden, fehlt gewissermaßen der Motor für die Optimierung des Meßinstruments.

11.2.2 Multinomiale Modelle als Theorieformalisierungen

Dienen multinomiale Modelle der Formalisierung einer kognitionspsychologischen Theorie oder Hypothese, so ist ihr Nutzen offenkundig: Einerseits *expliziert das Modell die Theorie* und die ihr zugrundeliegenden Annahmen, andererseits ermöglicht es die *rigorose Überprüfung der Theorie* mit Hilfe von Parameterschätzungen und Hypothesentests.

Können konventionelle Methoden der kognitionspsychologischen Forschung ergänzend von Nutzen sein, wenn man erst einmal ein multinomiales Modell formuliert hat? Intuitiv scheint das Vertrauen in ein Resultat, das durch modellbasierte Analyse der Daten gewonnen wurde, gestärkt, wenn eine varianzanalytische Auswertung der gleichen Daten ein ähnlich zu interpretierendes Resultat erbringt. Beeinflussen beispielsweise zwei experimentell manipulierte Faktoren zwei unterschiedliche Parameter des Modells selektiv, so erscheint es plausibel, daß eine varianzanalytische Auswertung der gleichen Daten eine Nullinteraktion der beiden Faktoren aufzeigen sollte.

Bei derartigen Erwartungen ist jedoch große Vorsicht geboten. Ob sich eine „Korrespondenz" modellbasierter und varianzanalytischer Ergebnisse ergibt, hängt von Eigenschaften des multinomialen Modells einerseits und von der Definition der abhängigen Variablen in der Varianzanalyse andererseits ab. Der Regelfall ist eher der, daß Ergebnisse, die auf dem Hintergrund eines multinomialen Modells gewonnen wurden, mit varianzanalytischen Ergebnissen nichts zu tun haben müssen. Eine Korrespondenz der Ergebnisse beider Auswertungsmethoden wäre also eher „Zufall". Batchelder und Riefer (1986) haben beispielsweise gezeigt, daß zwei Faktoren, welche die beiden Parameter c und r des Enkodierung-Abruf-Modells selektiv beeinflussen, bei varianzanalytischer Auswertung eine Interaktion produzieren müssen, wenn die Gesamtzahl der reproduzierten Items als abhängige Variable gewählt wird (vgl. Abschnitt 3.5.1; Erdfelder & Bayen, 1991). Auch das umgekehrte Ergebnis ist prinzipiell denkbar: Eine Interaktion ist auf varianzanalytischer Ebene nicht gegeben, wohl aber bezüglich der Parameter eines multinomialen Modells.

Doppeltauswertungen von Daten im Rahmen eines multinomialen Modells und der Varianzanalyse tragen also nicht zur „Validierung" der Ergebnisse oder zur Erhöhung ihrer Glaubwürdigkeit bei. Die Glaubwürdigkeit modellbasierter Ergebnisse hängt allein von der Gültigkeit des multinomialen Modells ab und nicht davon, ob sich die Ergebnisse mittels Varianzanalysen replizieren lassen.

Auch wenn Varianzanalysen zur Modellvalidierung nichts beitragen können, muß auf sie im Rahmen eines modellbasierten Vorgehens nicht immer völlig verzichtet werden. Zwei Funktionen traditioneller Auswertungsmethoden bleiben auch im Rahmen eines modellbasierten Vorgehens bestehen:

1) Erbringen des Nachweises, daß ein Ergebnismuster, welches im Rahmen eines multinomialen Modells analysiert werden soll, mit Ergebnissen vergangener Untersuchungen zur gleichen Thematik hinreichend genau übereinstimmt (Replikationsargument);
2) ergänzende Analysen von Prognosen oder Hilfsannahmen, die im Rahmen des multinomialen Modells nicht prüfbar sind (Ergänzungsargument).

Konventionelle Analysen, die weder dem Nachweis der Replikation eines bestimmten Ergebnismusters dienen noch ergänzenden Charakter im Rahmen einer umfassenden Theorie- oder Hypothesenprüfung haben, sollten unterbleiben. Sie verwirren oft mehr als sie zur Klärung der Sachlage beitragen (Batchelder & Riefer, 1996; Erdfelder & Bayen, 1991).

11.3 Die Grenzen des multinomialen Modellierens

Multinomiale Modellbildung kann zwar viele, aber nicht alle Probleme lösen, die sich im Rahmen der kognitionspsychologischen Forschung stellen. Die offenkundigsten Grenzen multinomialer Modellbildung sollen in diesem Abschnitt kurz angesprochen werden.

Eine erste — eher technische — Beschränkung ergibt sich aus der diskreten Natur der Beobachtungsdaten, die multinomiale Modelle zwangsläufig voraussetzen (Kapitel 3). Kontinuierliche Beobachtungsdaten wie z.B. Reaktionszeiten können im Rahmen eines multinomialen Modells nur nach erfolgter künstlicher Diskretisierung (z.B. durch Mediansplit, Interquartilsplit usw.) behandelt werden (vgl. Abschnitt 6.8.1). Hiermit ist jedoch immer ein Informationsverlust verbunden, der nach Möglichkeit vermieden werden sollte. Außerdem kann die Gültigkeit bzw. Ungültigkeit eines multinomialen Modells für künstlich diskretisierte Beobachtungsvariablen von der Art der Diskretisierung abhängen. Dies wirft ein neues Bedeutsamkeitsproblem auf und ist deshalb unbefriedigend. Gerade für Reaktions- und Antwortzeiten gibt es andere stochastische Modellierungsansätze, die als Alternativen zu multinomialen Modellen unbedingt in Erwägung gezogen werden sollten (Luce, 1986; Townsend & Ashby, 1983).

Eine zweite Beschränkung ergibt sich aus den engen Grenzen, die praktisch handhabbare Systeme von Beobachtungskategorien für die Modellbildung setzen. Zwar kann man sich beliebig komplexe Beobachtungssysteme ausdenken. Hiermit ist aber solange wenig gewonnen, wie man nicht auf riesige Stichprobenumfänge zurückgreifen kann. Ein komplexes Beobachtungssystem ist dann nichts wert,

wenn die erwarteten Kategoriehäufigkeiten für einen gegeben Datensatz nahe bei null liegen. Praktisch bedeutet das, daß Beobachtungssysteme immer mehr oder minder „einfach" sein müssen. Mehr als fünf Beobachtungskategorien pro Versuchsbedingung dürften eher die Ausnahme sein. Dies setzt aber wiederum für die Größe eines Modells Grenzen, das für das gegebene Beobachtungssystem identifizierbar und testbar sein kann. Kurz: Multinomiale Modelle sind ebenso wie die zugehörigen Beobachtungssysteme fast immer „einfach", d.h. parameterarm. Dies hat zur Folge, daß sie ihre Rolle im Kontext der Überprüfung psychologischer Theorien nur begrenzt erfüllen können. Komplette Theorien lassen sich nur in Ausnahmefällen im Rahmen eines multinomialen Modells formalisieren. Fast immer handelt es sich nur um Theoriebausteine oder Theorieelemente, die im Rahmen multinomialer Modelle abbildbar und prüfbar sind.

Eine dritte Beschränkung ergibt sich aus der mangelnden Sensitivität des Modellgeltungstests für Verletzungen der sequentiellen Prozeßtheorie, die multinomialen Ereignisbaummodellen häufig zugrunde liegt. Sequentielle Prozeßtheorien lassen sich zwar im Rahmen von Verarbeitungsbaummodellen häufig gut darstellen (vgl. Abschnitt 11.1.2). Letztlich geprüft wird aber nur, ob die Kategoriewahrscheinlichkeiten den Restriktionen genügen, die das System der Modellgleichungen ihnen auferlegt. Dies kann auch dann der Fall sein, wenn die wahre zeitliche Abfolge der Prozeßkomponenten eine ganz andere ist als im Ereignisbaum angenommen. Kurz: Der Regelfall ist, daß es zu einem gegebenen Verarbeitungsbaummodell ein anderes gibt, das a) datenäquivalent ist und b) eine andere Prozeßtheorie zum Ausdruck bringt. Wer also wirklich an der strengen Überprüfung von Prozeßtheorien interessiert ist, wird im Regelfall über ein einfaches Verarbeitungsbaummodell zu einer Ein-Punkt-Datenerhebung hinausgehen müssen, sei es durch Markoff-Ketten-Modellierung von Zeitreihendaten (z.B. Bäuml, 1996; Batchelder et al., 1997; Erdfelder, 1993; vgl. Abschnitt 6.5) oder durch explizite Einbeziehung der Zeitvariable in den Modellierungsansatz (z.B. Falmagne et al., 1990; Luce, 1986; Townsend & Ashby, 1983; vgl. auch Abschnitt 6.8). Mit dem letztgenannten Ansatz verläßt man im Regelfall den Rahmen multinomialer Modellbildung.

Mit der dritten Beschränkung hängt die Schwierigkeit eng zusammen, mittels multinomialer Modelle zwischen seriellen und parallelen Informationsverarbeitungstheorien zu diskriminieren. Es gibt Versuche, parallele versus serielle Prozesse im Rahmen multinomialer Verarbeitungsbaummodelle gegeneinander zu testen (Batchelder & Riefer, 1999; Brown, 1998; vgl. auch Abschnitt 7.5.2). Ihr Grundgedanke ist die stochastische Unabhängigkeit versus Abhängigkeit von Prozessen, die auf unterschiedlichen Ebenen eines Verarbeitungsbaums angesiedelt sind. Ist die bedingte Wahrscheinlichkeit eines Prozesses auf der zweiten Ebene unabhängig vom Ausgang des Prozesses auf der ersten Ebene, so sind die betref-

fenden Prozesse unabhängig. Man kann dann ihre Reihenfolge im Baummodell vertauschen und erhält ein äquivalentes Modell. Batchelder und Riefer (1999) und Brown (1998) bewerten dies als Hinweis auf die Parallelität der involvierten Prozesse, während stochastische Abhängigkeit, die Nichtaustauschbarkeit der Ebenen eines Ereignisbaummodells zur Folge hat, als Hinweis auf die sequentielle Natur der Prozesse gewertet wird. Das Problem bei dieser Argumentation ist, daß stochastische Abhängigkeit versus Unabhängigkeit von Prozessen auf der einen und Serialität versus Parallelität von Prozessen auf der anderen Seite kausal nicht miteinander verknüpft sind. Man kann sich problemlos zwei zeitlich nacheinander ablaufende Prozesse A und B vorstellen, wobei die Erfolgswahrscheinlichkeit des B-Prozesses unabhängig vom Ausgang des A-Prozesses ist. Umgekehrt sind auch zwei zeitlich parallele Prozesse A und A´ denkbar, deren Ausgänge korreliert sind, weil sie beispielsweise beide durch dieselben Hintergrundvariablen beeinflußt werden. Der Test auf stochastische Unabhängigkeit versus Abhängigkeit von Prozessen ist also nur sehr locker mit der Frage nach der Serialität versus Parallelität von Prozessen verknüpft. Dieser Unsicherheit sind sich Batchelder und Riefer (1999) auch durchaus bewußt, denn sie urteilen sehr vorsichtig, stochastisch unabhängige Prozesse seien „*perhaps* a case of parallel processing" (a.a.O., S. 80, Hervorhebung vom Verfasser der vorliegenden Arbeit).

Ein technisches Problem, das sich in diesem Zusammenhang ebenfalls stellt, ist die mangelnde Trennschärfe von Tests der stochastischen Unabhängigkeit versus Abhängigkeit. Typischerweise paßt mindestens ein paralleles und mindestens ein serielles Modell nahezu gleich gut auf die Daten, ein Ergebnis, das auch Brown (1998) beobachtet hat. In einer solchen Situation hilft nur ein umfassendes Validierungsprogramm zu den Parametern des allgemeineren (parameterreicheren) Modells. Auf keinen Fall sollte ein gut auf die Daten passendes Modell nur deshalb verworfen werden, weil es mehr Parameter als ein anderes, auf die Daten ebenfalls passendes Modell umfaßt. Parameterarme Modelle können durchaus „zufällig" gut auf die Daten eines Experiments passen, obwohl eigentlich ein anderes, allgemeineres Modell angemessen ist (vgl. Abschnitt 7.5.2).

Unabhängig von diesem technischen Problem bleibt der Vorbehalt bestehen, daß Serialität versus Parallelität von kognitiven Prozessen nicht einfach als Abhängigkeit versus Unabhängigkeit von Prozessen „operationalisierbar" ist. Strengere und annahmeärmere Tests der Serialität versus Parallelität von Prozessen sind möglich, wenn man zu simultanen Analysen von Antworthäufigkeiten und Antwortzeiten übergeht (Townsend & Ashby, 1983; vgl. Abschnitt 6.8.1 und Abschnitt 6.8.2).

Eine fünfte Beschränkung ergibt sich aus der nur bedingten Eignung multinomialer Modelle für den direkten Theorienvergleich. Diese Beschränkung wiegt um so schwerer, als multinomiale Modelle gerade für diesen Zweck oft relativ un-

kritisch eingesetzt wurden (vgl. Kapitel 7 und den Einleitungsteil zu Kapitel 10). Problemlos anwendbar sind Theorienvergleiche mittels multinomialer Modelle aber nur dann, wenn sich die konkurrierenden Theorien gleichermaßen gut anhand eines bestimmten experimentellen Paradigmas prüfen lassen und sie darüber hinaus gleichermaßen gut im Rahmen multinomialer Modelle formalisierbar sind. Beide Bedingungen können bei sehr ähnlichen Theorien, die sich nur hinsichtlich der Annahmen zu einigen wenigen Prozessen unterscheiden und die sich auf ein ganz bestimmtes experimentelles Paradigma beziehen, durchaus einmal erfüllt sein. Häufig läuft der Theorienvergleich dann auf einen Vergleich hierarchisch ineinander geschachtelter Modelle hinaus, wobei das strengere Modell durch bestimmte Parameterfixierungen oder Gleichsetzungsrestriktionen aus dem allgemeineren Modell hervorgeht. Sofern die Hypothesentests auf der Basis von Teststärkeüberlegungen so konstruiert werden, daß eine rationale Entscheidung zwischen beiden Modellen bei geringen α- und β-Fehlerrisiken möglich wird, ist gegen solche Theorienvergleiche nichts einzuwenden (vgl. Kapitel 4.3).

Laufen Theorienvergleiche dagegen nicht auf hierarchisch ineinander geschachtelte Modelle hinaus oder sind die Theorien von höherem Allgemeinheitsgrad, so daß sie anhand eines bestimmten Paradigmas nicht erschöpfend geprüft werden können — das betrifft die meisten „interessanten" Theorien der kognitiven Psychologie —, so sollten die Ergebnisse einzelner Untersuchungen immer mit großer Vorsicht bewertet werden. Ein Vergleich in einem Paradigma A kann ganz andere Ergebnisse erbringen als ein Vergleich in einem Paradigma B, wie in Kapitel 10.1 und Kapitel 10.2 anhand der Überprüfung konkurrierender Theorien zum Falschinformationseffekt klar gezeigt wurde. Der Schluß hieraus kann nur lauten, daß man sich mit voreiligen Schlüssen sehr zurückhalten muß. Erst umfangreiche Untersuchungen auf der Basis paradigmenübergreifender Forschungsprogramme können ein solides Fundament für die vergleichende Bewertung konkurrierender Theorien liefern. Diese Schlußfolgerung ist natürlich für modellbasierte Theorienvergleiche nicht spezifisch, aber sie wird durch modellbasiertes Vorgehen auch keineswegs gegenstandslos, wie es gelegentlich suggeriert wird (vgl. Kapitel 7).

Eine sechste Beschränkung ergibt sich aus der Qualität des Experiments oder der Experimente, auf die ein multinomiales Modell angewendet wird. Die Schlußfolgerungen aus Anwendungen multinomialer Modelle können nie besser sein als die Experimente, aus denen sie erwachsen sind. Fehler im experimentellen Design oder in der Versuchsprozedur werden durch die multinomiale Modellierung nicht rückgängig gemacht, sondern schlagen im Regelfall auf die Parameterwerte und gegebenenfalls auch auf die Anpassungsstatistiken durch. Sind etwa Versuchsmaterial und Versuchsbedingungen mangels Ausbalancierung miteinander konfundiert, so muß damit gerechnet werden, daß die Parameterschätzungen

nicht mehr sinnvoll interpretierbar sind, selbst wenn die Modellanpassungsstatistik keine Modellverletzung signalisiert. Gleiches gilt, wenn versäumt wurde, Versuchspersonen den Versuchsbedingungen randomisiert zuzuweisen usw. Die Regeln guten Experimentierens (zusammenfassend: Bredenkamp, 1969; 1980; 1996; in press; Erdfelder, 1996) sind also nicht nur im Bereich der traditionellen experimentellen Psychologie von essentieller Bedeutung, sondern auch und gerade bei modellbasierter Analyse experimenteller Daten.

11.4 Die Zukunft des multinomialen Modellierens

Betrachtet man den Trend der Publikationen zu multinomialen Modellen in der kognitiven Psychologie (vgl. Batchelder & Riefer, 1999, und Kapitel 7 der vorliegenden Arbeit), so ist zu erwarten, daß die Bedeutung dieser Modellklasse in Zukunft eher wachsen als sinken wird. Sofern die Grenzen des Ansatzes, die im letzten Absatz besprochen wurden, gebührend beachtet werden, ist diese Entwicklung durchaus zu begrüßen.

Zu hoffen ist, daß Schwächen bisheriger Anwendungen in Zukunft mehr und mehr vermieden werden. Hierzu wird die inzwischen verfügbare Computersoftware, die sehr leistungsfähig und benutzerfreundlich ist, ihren Beitrag leisten wird (Hu, 1999; Rothkegel, 1999; vgl. Abschnitt 4.7). Rationale statistische Entscheidungen über Modellgeltungs- und Parameterhypothesen unter simultaner Kontrolle des α- und des β-Fehlers sollten der Regelfall werden. Die praktische Seite dieser Kontrolle ist inzwischen so einfach geworden, daß es keine Entschuldigung mehr gibt, wenn man auf sie verzichtet (vgl. dazu Kapitel 4.7). Gleiches gilt für systematische Valididierungsstudien zu multinomialen Modellen, die ganz selbstverständlich und routinemäßig bei neu entwickelten Modellen durchgeführt werden sollten. Erst wenn ein Modell diese Hürden genommen hat, kann das Modell für weitere Anwendungen ohne Einschränkung empfohlen werden.

Zu erwarten ist, daß Fehler in der Handhabung multinomialer Modellierungen und in der Interpretation ihrer Ergebnisse in Zukunft seltener werden. Inzwischen leicht vermeidbare Fehler sind etwa die Interpretation von Parameterschätzwerten aus inadäquaten Modellen oder die inadäquate Umsetzungen von Ungleichheitsrestriktionen zu den Modellparametern (vgl. Abschnitt 7.9). Wie sich derartige Restriktionen auf einfache Art und Weise praktisch realisieren lassen, wurde in Abschnitt 4.3.2.1 ausgiebig besprochen. Gleiches gilt für die adäquate Prüfung der Hypothese einer Null-Wechselwirkung (vgl. ebenfalls Abschnitt 4.3.2.1). Auch hier gilt: Die Techniken sind inzwischen so leicht anwendbar, daß es keine Entschuldigung mehr gibt, wenn man in Zukunft trotz klarer Indikation auf sie verzichtet.

Zu wünschen ist, daß die Grenzen multinomialer Modellierungsansätze in Zukunft immer weiter ausgedehnt werden können. Ein ganz wichtiger Bereich ist hierbei die paradigmenübergreifende multinomiale Modellierung, wie sie in Kapitel 8 anhand unterschiedlicher Quellendiskriminationstests ansatzweise illustriert wurde. In dem Maße, in dem es gelingt, mehrere empirische Paradigmen im Rahmen eines Modells zu repräsentieren, kommt man dem Ziel umfassender Theorieprüfungen mittels multinomialer Modelle näher.

Eine zweite Entwicklung, die in diesem Zusammenhang förderlich sein kann, ist die Ausweitung der empirischen Datenbasis von Vp-Urteilen und Vp-Reaktionen — der bisherigen Domäne multinomialer Modelle — auf Antwortzeiten, Blickbewegungsdaten und physiologische Indikatoren. Solange diese Daten kategorial repräsentiert werden können, spricht absolut nichts dagegen, multinomiale Modelle hierauf anzuwenden. Wenn sie nicht adäquat kategorial repräsentierbar sind, kann man multinomiale Modelle entsprechend erweitern (vgl. beispielsweise Abschnitt 6.8.2). Daß physiologische Daten mit bestimmten kognitiven Prozessen hoch korreliert sein können, wurde bereits eindrucksvoll nachgewiesen (z.B. Martinez & Kesner, 1998). Es liegt daher auf der Hand, sie in multinomiale Modelle kognitiver Prozesse miteinzubeziehen, wann immer eine Theorie über Verhaltensprognosen hinaus auch physiologische Prognosen ermöglicht.

Eine dritte Entwicklung, die in Zukunft eine größere Rolle spielen sollte, ist die mathematisch-logische Verknüpfung von psychologischer Theorie auf der einen und multinomialem Modell auf der anderen Seite. Batchelder und Riefer (1999) waren in ihrem Überblicksreferat noch sehr skeptisch: „Of course, it is not reasonable to expect that a tight, mathematical connection can always be drawn between an MPT model and a more complete theory for any given phenomenon" (a.a.O., S. 76). Sie gingen davon aus, daß allein aufgrund der typischerweise verbalen Formulierung der zu prüfenden psychologischen Theorie keine streng mathematisch-logische Brücke zu dem multinomialen Modell konstruiert werden kann, das letztlich auf Daten angewendet wird. Auch wenn diese Sichtweise für viele Modellanwendungen in der Vergangenheit zutreffend gewesen sein mag, so muß das doch in Zukunft nicht der Regelfall bleiben. Dies gilt insbesondere für weitgehend formalisierte Gedächtnistheorien, beispielsweise für globale Gedächtnismodelle wie SAM, TODAM, MINERVA2 usw. (zusammenfassend: Clark & Gronlund, 1996). Brandt (2000) hat kürzlich exemplarisch gezeigt, daß Hintzmans MINERVA2 (Hintzman, 1986; 1988) — wenn es konsequent durchdacht und auf empirische Daten bezogen wird — letztlich in ein multinomiales Modell zu Ja-Nein-Rekognitionstest einmündet.[106] Der Weg von den MINERVA2-Annahmen

[106]Allerdings handelt es sich hierbei nicht um ein MVB-Modell, sondern um ein parametrisiertes multinomiales Modell, das eher Ähnlichkeiten zur Signalentdeckungstheorie aufweist. Der wesentliche Unterschied zur Signalentdeckungstheorie ist, daß das Modell in termini der MI-

zum multinomialen Modell ist dabei ein rein mathematisch-deduktiver, der mit sehr wenige technischen Zusatzannahmen auskommt. Man kann auch sagen, daß Brandts (2000) Modell eine Weiterentwicklung der MINERVA2-Theorie zu einem Datenanalysemodell ist. Herkömmliche Überprüfungsansätze zu MINERVA2, die eher informellen Charakter haben, stellen sich auf diesem Hintergrund als inadäquat heraus. Die modellbasierten Analysen führen teilweise zu anderen Bewertungen empirischer Daten und treiben auf diese Weise die Forschung zu globalen Gedächtnismodellen voran.

Das letzte Beispiel zeigt, daß multinomiale Modelle nicht auf die beiden Funktionen, Werkzeuge der Theorieprüfung oder der Ableitung von Meßverfahren zu sein, beschränkt bleiben müssen. Diese Funktionen sind durchaus wichtig, und sie bleiben auch in Zukunft erhalten. Darüber hinaus können multinomiale Modelle aber die Theorieentwicklung selbst vorantreiben und so neue Perspektiven für die Analyse kognitionspsychologischer Daten eröffnen. Hierin liegt ein großes, bislang noch kaum erschlossenes Potential für zukünftige Forschungsanstrengungen.

11.5 Zusammenfassung

In diesem abschließenden Kapitel wurde die Rolle multinomialer Modelle in der kognitionspsychologischen Forschung aus verschiedenen Perspektiven beleuchtet. Es wurde zunächst argumentiert, daß multinomiale Modelle dem Ziel der Befreiung vom „Joch der precision-importance trade-off function" näher kommen als andere Ansätze, beispielsweise die Repräsentationstheorie des Messens. Leitlinien der Konstruktion geeigneter Beobachtungssysteme und fragestellungsadäquater Modelle wurden vorgestellt, wobei allerdings betont wurde, daß der Prozeß der Modellkonstruktion ein kreativer Akt ist, der prinzipiell nicht normierbar ist. Sodann wurde der Nutzen bisheriger Anwendungen von multinomialen Modellen hervorgehoben, der einerseits darin besteht, Meßmodelle für kognitive Prozesse zu liefern, und andererseits darin, die Formalisierung und Prüfung kognitionspsychologischer Theorien voranzutreiben. Die Vorteile gegenüber traditionellen, auf Ad-hoc-Maßen basierenden Methoden der kognitiven Psychologie wurden betont.

Die letzten beiden Abschnitte des Kapitels beschäftigten sich mit den gegenwärtigen Grenzen multinomialer Modellierungsansätze und Perspektiven für zukünftige Entwicklungen. Es ist zu erwarten, daß multinomiale Modelle in der Zukunft eine eher noch wichtigere Rolle in der kognitionspsychologischen Forschung spielen werden als das in der Gegenwart der Fall ist. Dies gilt nicht nur für

---

NERVA2-Modellparameter parametrisiert ist, deren Schätzwerte somit direkt im Rahmen der Theorie interpretierbar sind.

den Kontext der Theorieprüfung und Meßmodellanwendung, sondern zunehmend auch für den Kontext der Theorieentwicklung.

Literaturverzeichnis

Abramowitz, M. & Stegun, I.A. (Eds.). (1964). *Handbook of mathematical functions.* (Applied Mathematics Series 55). Washington, DC: National Bureau of Standards.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Aickin, M. (1984). Some fallacies in the computation of paternity probabilities. *American Journal of Human Genetics, 36,* 904-915.

Aitkin, M., Anderson, D. & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, 144A,* 419-461.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csáki (Eds.), *Second international symposium on information theory* (S. 267-281). Budapest: Akadémiai Kiadó.

Andersen, E.B. (1990). *The statistical analysis of categorical data.* Heidelberg: Springer.

Andersen, E.B. (1996). Log-lineare Modelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 303-314). Weinheim: Psychologie Verlags Union.

Anderson, J.A. (1995). *An introduction to neural networks.* Cambridge, MA: MIT Press.

Anderson, J.R. (1982a). Acquisition of cognitive skill. *Psychological Review, 89,* 369-406.

Anderson, J.R. (1992). Automaticity and the ACT* theory. *The American Journal of Psychology, 105,* 165-180.

Anderson, J.R. (1993). *Rules of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, N.H. (1982b). *Foundations of information integration theory.* New York: Academic Press.

Andres, J. (1996). Das Allgemeine Lineare Modell. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden.* Weinheim: Psychologie Verlags Union.

Ashby, F.G. (Ed.). (1992). *Multidimensional models of perception and cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F.G., Prinzmetal, W., Ivry, R. & Maddox, W.T. (1996). A formal theory of feature binding in object perception. *Psychological Review, 103,* 165-192.

Ashton, G.C. (1980). Mismatches in genetic markers in a large family study. *American Journal of Human Genetics, 32,* 601-613.

Atkinson, R.C., Bower, G.H. & Crothers, E.J. (1965). *An introduction to mathematical learning theory.* New York: Wiley.

Ayers, M.S. & Reder, L.M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review, 5,* 1-21.

Azencott, R. & Dacunha-Castelle, D. (1986). *Series of irregular observations.* New York: Springer.

Baddeley, A.D. (1997). *Human memory. Theory and practice.* (Rev. ed.). Hove, GB: Psychology Press.

Bailey, N.T.J. (1951). Testing the solubility of maximum likelihood equations in the routine application of scoring methods. *Biometrics, 7,* 268-274.

Bakan, D. (1954). A generalization of Sidman´s results on group and individual functions, and a criterion. *Psychological Bulletin, 51,* 63-64.

Baker, R.R. & Bellis, M.A. (1995). *Sperm competition: Copulation, masturbation, and infidelity*. London: Chapman & Hall.

Bamber, D. & van Santen, J.P.H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29,* 443-473.

Bamber, D. & van Santen, J.P.H. (2000). How to assess a model´s testability and identifiability. *Journal of Mathematical Psychology, 44,* 20-40.

Baschek, I.L., Bredenkamp, J., Oehrle, B. & Wippich, W. (1977). Bestimmung der Bildhaftigkeit (I), Konkretheit (C) und der Bedeutungshaltigkeit (m**1) von 800 Substantiven. *Zeitschrift für Experimentelle und Angewandte Psychologie, 24,* 353-396.

Batchelder, W.H. (1991). Getting wise about minimum distance measures. (Review of T.R.C. Read and N.A.C. Cressie: Goodness-of-fit statistics for discrete multivariate data. New York: Springer, 1988). *Journal of Mathematical Psychology, 35,* 267-273.

Batchelder, W.H. (1994). *Mathematical models and measurement theory*. Handout, Irvine, CA: University of California at Irvine,

Batchelder, W.H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10,* 331-344.

Batchelder, W.H. (1999). Multinomial processing tree models of cognition. In W. Hacker & M. Rinck (Eds.), *Zukunft gestalten. Bericht über den 41. Kongreß der Deutschen Gesellschaft für Psychologie in Dresden 1998* (S. 65-73). Lengerich: Pabst.

Batchelder, W.H. & Boyd, J.P. (1985). Review of "Models for behavior: Stochastic processes in psychology" by Thomas D. Wickens. *Journal of Mathematical Psychology, 29,* 122-127.

Batchelder, W.H., Chosak-Reiter, J., Shankle, W.R. & Dick, M.B. (1997). A multinomial modeling analysis of memory deficits in Alzheimer´s and vascular dementia. *Journal of Gerontology: Psychological Sciences and Social Sciences, 52B,* P206-P215.

Batchelder, W.H. & Crowther, C.S. (1997). Multinomial processing tree models of factorial categorization. *Journal of Mathematical Psychology, 41,* 45-55.

Batchelder, W.H., Hu, X. & Riefer, D.M. (1994). Analysis of a model for source monitoring. In G.H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (S. 51-65). New York: Springer.

Batchelder, W.H., Kumbasar, E. & Boyd, J.P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology, 22,* 29-58.

Batchelder, W.H. & Riefer, D.M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review, 87,* 375-397.

Batchelder, W.H. & Riefer, D.M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology, 39,* 129-149.

Batchelder, W.H. & Riefer, D.M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97,* 548-564.

Batchelder, W.H. & Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57-86.

Batchelder, W.H., Riefer, D.M. & Hu, X. (1994). Measuring memory factors in source monitoring: Reply to Kinchla. *Psychological Review, 101,* 172-176.

Batchelder, W.H., Riefer, D.M. & Walton, E.F. (1990). *A multinomial model analysis of suggestibility effects in eyewitness memory*. Vortrag gehalten auf der Meeting of the European Mathematical Psychology Group (EMPG), Bristol, GB, 1990.

Batchelder, W.H. & Romney, A.K. (1986). The statistical analysis of a general condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Ed.), *Information pooling and group decision making* (S. 103-112). Greenwich, CO: SAI Press.

Batchelder, W.H. & Romney, A.K. (1988). Test theory without an answer key. *Psychometrika, 53,* 71-92.

Batchelder, W.H. & Romney, A.K. (1989). New results in test theory without an answer key. In E.E. Roskam (Ed.), *Mathematical psychology in progress* (S. 229-248). Berlin: Springer.

Bäuml, K.-H. (1991a). Experimental analysis of storage and retrieval processes involved in retroactive inhibition: The effect of presentation mode. *Acta Psychologica, 77,* 103-119.

Bäuml, K.-H. (1991b). Retroaktive Hemmung: Der Einfluß des interpolierten Kategorienmaterials auf die Verfügbarkeit von Informationen. *Zeitschrift für experimentelle und angewandte Psychologie, 38,* 169-187.

Bäuml, K.-H. (1991c). Retroaktive Hemmung: Kein Einfluß der Kategorisierungsart - konzeptuell versus phonetisch-visuell - auf die Verfügbarkeit von Information. *Zeitschrift für Psychologie, 199,* 177-190.

Bäuml, K.-H. (1992). Diskriminationslernen bei rotierten Gesichtern. *Zeitschrift für experimentelle und angewandte Psychologie, 34,* 1-17.

Bäuml, K.-H. (1996). A Markov model for measuring storage loss and retrieval failure in retroactive inhibition. *Acta Psychologica, 92,* 231-250.

Bayen, U.J. (1990). *Zur Lokalisation von Altersdifferenzen im episodischen Gedächtnis Erwachsener: Eine Querschnittuntersuchung auf der Basis eines mathematischen Modells* (Berichte aus dem Psychologischen Institut der Universität Bonn, Band 12, Heft 2). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Bayen, U.J., Erdfelder, E. & Bearden, J.N. (1998). *Aging and hindsight: Bias or memory problem?* Vortrag gehalten auf der 7th Biannual Cognitive Aging Conference, Atlanta, GA, USA, April 1998.

Bayen, U.J. & Murnane, K. (1996). Aging and the use of perceptual and temporal information in source memory tasks. *Psychology and Aging, 11,* 293-303.

Bayen, U.J., Murnane, K. & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 197-215.

Bayen, U.J., Nakamura, G.V., Dupuis, S. & Yang, C.-L. (2000). The use of schematic knowledge about sources in source monitoring. *Memory & Cognition, 28,* 480-500.

Bearden, J.N., Zukerman, J., Enneking, A., Erdfelder, E. & Bayen, U.J. (1999). *Aging and the hindsight bias: Effects of anchor availability.* Vortrag gehalten auf der North Carolina Cognition Group´s Annual Meeting., North Carolina State University, Raleigh, NC., February 1999.

Begg, I.M., Anas, A. & Farinacci, S. (1992). Dissociation of processes in belief: source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121,* 446-458.

Bekerian, D.A. & Bowers, J.M. (1983). Eyewitness testimony: Were we misled? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 139-145.

Belli, R.F. (1989). Influences of misleading postevent information: Misinformation interference and acceptance. *Journal of Experimental Psychology: General, 118,* 72-85.

Belli, R.F. (1993). Failures of interpolated tests in inducing memory impairment with final modified tests: Evidence unfavourable to the blocking hypothesis. *American Journal of Psychology, 106,* 407-427.

Belli, R.F., Lindsay, S., Gales, M.S. & McCarthy, T.T. (1994). Memory impairment and source misattribution in postevent misinformation experiments with short retention intervalls. *Memory & Cognition, 22,* 40-54.

Belli, R.F., Windschitl, P.D., McCarthy, T.T. & Winfrey, S.E. (1992). Detecting memory impairment with a modified test procedure: Manipulating retention interval with centrally presented event items. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 356-367.

Bender, R.H., Wallsten, T.S. & Ornstein, P.A. (1996). Age differences in encoding and retrieving details of a pediatric examination. *Psychonomic Bulletin & Review, 3,* 188-198.

Beran, R. (1986). Simulated power functions. *Annals of Statistics, 14,* 151-173.

Bernbach, H.A. (1967). Decision processes in memory. *Psychological Review, 74,* 462-480.

Bernbach, H.A. (1971). Recognition memory and signal detection analyses. *Psychological Review, 78,* 274.

Bernstein, F. (1925). Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Zeitschrift für Abstammungs- und Vererbungslehre, 37,* 237-270.

Berry, K.J. & Mielke, P.W. (1995). Exact cumulative probabilities for the multinomial distribution. *Educational and Psychological Measurement, 55,* 769-772.

Birch, M.W. (1964). A new proof of the Pearson-Fisher theorem. *Annals of Mathematical Statistics, 35,* 817-824.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (S. 395-479). Reading, MA: Addison-Wesley.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Blackwell, H.R. (1953). *Psychophysical thresholds. Experimental studies of methods of measurement* (University of Michigan Engineering Research Institute Bulletin Nr. 36). Ann Arbor, MI: Engineering Research Institute.

Blackwell, H.R. (1963). Neural theories of simple visual discrimination. *Journal of the Optical Society of America, 53,* 129-160.

Blumen, I.M., Kogan, M. & McCarthy, P.J. (1955). *The industrial mobility of labor as a probability process.* Ithaca: Cornell University Press.

Bosch, K. (1986). *Elementare Einführung in die Wahrscheinlichkeitsrechnung.* Braunschweig: Vieweg.

Bower, G.H. & Trabasso, T. (1964). Concept identification. In R.C. Atkinson (Ed.), *Studies in mathematical psychology* (S. 32-94). Stanford: Stanford University Press.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290-302.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Brainerd, C.J. (1985). Model-based approaches to storage and retrieval development. In C.J. Brainerd & M. Pressley (Ed.), *Basic processes in memory development: Progress in cognitive development research* (S. 143-208). New York: Springer.

Brainerd, C.J. & Reyna, V.F. (1998a). Fuzzy-trace theory and children´s false memories. *Journal of Experimental Child Psychology, 71,* 81-129.

Brainerd, C.J. & Reyna, V.F. (1998b). When things that were never experienced are easier to „remember" than things that were. *Psychological Science, 9,* 484-489.

Brainerd, C.J., Reyna, V.F., Howe, M.L. & Kingma, J. (1990). The development of forgetting and reminiscence. *Monographs of the Society for Research in Child Development, 55,* Serial No. 222.

Brainerd, C.J., Reyna, V.F. & Mojardin, A.H. (1999). Conjoint recognition. *Psychological Review, 106,* 160-179.

Brandt, M. (2000). *Multinomiale Modellierung globaler Matchingprozesse im menschlichen Gedächtnis. (Unveröffentlichtes Manuskript).* Bonn: Rheinische Friedrich-Wilhelms-Universität Bonn.

Bredenkamp, J. (1969). Experiment und Feldexperiment. In C.F. Graumann (Ed.), *Handbuch der Psychologie, Band 7: Sozialpsychologie, 1. Halbband: Theorien und Methoden* (S. 332-374). Göttingen: Hogrefe.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt am Main: Akademische Verlagsgesellschaft.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente.* Darmstadt: Steinkopff.

Bredenkamp, J. (1984). Weitere kritische Anmerkungen zur Anwendung der Informationsintegrationstheorie in der Entwicklungspsychologie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychoplogie, 16,* 134-145.

Bredenkamp, J. (1986). *Dürfen wir psychologische Hypothesen statistisch prüfen?* (Berichte aus dem Psychologischen Institut der Universität Bonn, Band 16, Heft 2). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Bredenkamp, J. (1990). Kognitionspsychologische Untersuchungen eines Rechenkünstlers. In H. Feger (Ed.), *Wissenschaft und Verantwortung. Festschrift für Karl Josef Klauer* (S. 47-70). Göttingen: Hogrefe.

Bredenkamp, J. (1996). Grundlagen experimenteller Methoden. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Ed.), *Handbuch Quantitative Methoden* (S. 37-46). Weinheim: Beltz.

Bredenkamp, J. (1998a). Die Erfassung unbewußter Verarbeitungsprozesse. In K.C. Klauer & H. Westmeyer (Eds.), *Psychologische Methoden und soziale Prozesse* (S. 30-46). Lengerich: Pabst.

Bredenkamp, J. (1998b). *Lernen, Erinnern, Vergessen*. München: C.H. Beck´sche Verlagsbuchhandlung.

Bredenkamp, J. (in press). Laboratory experiment: methodology. In N.J. Smelser & P.B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences*. Amsterdam, NL: Pergamon.

Bredenkamp, J. & Erdfelder, E. (1985). Multivariate Varianzanalyse nach dem V-Kriterium. *Psychologische Beiträge, 27,* 127-154.

Bredenkamp, J. & Erdfelder, E. (1996). Methoden der Gedächtnispsychologie. In D. Albert & K.-H. Stapf (Eds.), *Gedächtnis (= Enzyklopädie der Psychologie, Themenbereich C, Serie II, Band 4)* (S. 1-94). Göttingen: Hogrefe.

Bredenkamp, J. & Wippich, W. (1977a). *Lern- und Gedächtnispsychologie. Band I*. Stuttgart: Kohlhammer.

Bredenkamp, J. & Wippich, W. (1977b). *Lern- und Gedächtnispsychologie. Band II*. Stuttgart: Kohlhammer.

Brent, R.P. (1973). *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice Hall.

Bridgeman, P.W. (1927). *The logic of modern physics*. New York: Macmillan.

Broadbent, D.E. (1967). Word-frequency effect and response bias. *Psychological Review, 74,* 1-15.

Bröder, A. (1999). *„Take the Best — Ignore the Rest". Untersuchungen zur psychologischen Realität einer begrenzt rationalen Entscheidungsheuristik*. Dissertation, Psychologisches Institut der Universität Heidelberg, Heidelberg.

Bröder, A. & Bredenkamp, J. (1996). SLIP-Technik, Prozeßdissoziationsmodell und multinomiale Modellierung: Neue Werkzeuge zum experimentellen Nachweis "Freudscher Versprecher"? *Zeitschrift für Experimentelle Psychologie, 43,* 175-202.

Bröder, A. & Erdfelder, E. (1997). Hindsight Bias: beeinträchtigte Erinnerung, verzerrte Rekonstruktion oder beides? In H. Mandl (Ed.), *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996* (S. 31-36). Göttingen, Germany: Hogrefe.

Broffitt, J.D. & Randles, R.H. (1977). A power approximation for the chi-square goodness-of-fit test: simple hypothesis case. *Journal of the American Statistical Association, 72,* 604-607.

Brown, V. (1998). Comparing parallel and sequential multinomial models of letter identification. In C.E. Dowling, F.S. Roberts & P. Theuns (Ed.), *Recent progress in mathematical psychology* (S. 253-284). Mahwah, NJ: Lawrence Erlbaum Associates.

Buchner, A. & Erdfelder, E. (1996). On assumptions of, relations between, and evaluations of some process dissociation measurement models. *Consciousness and Cognition, 5,* 581-594.

Buchner, A., Erdfelder, E., Steffens, M. & Martensen, H. (1997). The nature of memory processes underlying recognition judgments in the process dissociation procedure. *Memory & Cognition, 25,* 508-517.

Buchner, A., Erdfelder, E. & Vaterrodt-Plünnecke, B. (1995). Toward unbiased measurement of conscious and unconscious memory processes within the process dissociation framework. *Journal of Experimental Psychology: General, 124,* 137-160.

Buchner, A., Faul, F. & Erdfelder, E. (1992). *G•Power: A priori, post-hoc, and compromise power analyses for the Macintosh* [Computer-Programm]. Bonn, Germany: Bonn University.

Buchner, A., Steffens, M., Erdfelder, E. & Rothkegel, R. (1997). A multinomial model to assess fluency and recollection in a sequence learning task. *The Quarterly Journal of Experimental Psychology, 50A,* 631-663.

Buchner, A., Steffens, M.C. & Rothkegel, R. (1998). On the role of fragmentary knowledge in a sequence learning task. *Quarterly Journal of Experimental Psychology, 51A,* 251-281.

Buchner, A. & Wippich, W. (in press). On the reliability of implicit and explicit memory measures. *Cognitive Psychology.*

Bühler, K. (1908). Antwort auf die von W. Wundt erhobenen Einwände gegen die Methode der Selbstbeobachtung an experimentell erzeugten Erlebnissen. *Archiv für die gesamte Psychologie, 12,* 93-122.

Bunge, M. & Ardila, R. (1987). *Philosophy of psychology.* New York: Springer.

Buss, D.M., Larsen, R.J. & Westen, D. (1996). Sex differences in jealousy: Not gone, not forgotten, and not explained by alternative hypotheses. *Psychological Science, 7,* 373-375.

Buss, D.M., Larsen, R.J., Westen, D. & Semmelroth, J. (1992). Sex differences in jealousy: Evolution, physiology, and psychology. *Psychological Science, 3,* 251-255.

Buunk, B.P., Angleitner, A., Oubaid, V. & Buss, D.M. (1996). Sex differences in jealousy in evolutionary and cultural perspective: Tests from the Netherlands, Germany, and the United States. *Psychological Science, 7,* 359-363.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carnap, R. (1936). Testability and meaning, Part 1. *Philosophy of Science, 3,* 420-471.

Carnap, R. (1937). Testability and meaning, Part 2. *Philosophy of Science, 4,* 2-40.

Carter, D.S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. *Educational and Psychological Measurement, 39,* 261-266.

Ceppellini, R., Siniscalco, M. & Smith, C.A.B. (1955). The estimation of gene frequencies in a random-mating population. *Annals of Human Genetics, 20*, 97-115.

Chandler, C.C. (1991). How memory for an event is influenced by related events: Interference in modified recognition tests. *Journal of Experimental Psychology: Learning,Memory,and Cognition, 17*, 115-125.

Chechile, R.A. (1987). Trace susceptibility theory. *Journal of Experimental Psychology: General, 116*, 203-222.

Chechile, R.A. (1998). A new method for estimating model parameters for multinomial data. *Journal of Mathematical Psychology, 42*, 432-471.

Chechile, R.A. & Meyer, D.L. (1976). A Bayesian procedure for separately estimating storage and retrieval components of forgetting. *Journal of Mathematical Psychology, 13*, 269-295.

Chechile, R.A. & Soraci, S.A.J. (1999). Evidence for a multiple-process account of the generation effect. *Memory, 7*, 483-508.

Christiaansen, R.E. & Ochalek, K. (1983). Editing misleading information from memory: Evidence for the coexistence of original and postevent information. *Memory & Cognition, 11*, 467-475.

Clark, S.E. & Gronlund, S.D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37-60.

Clark, S.J. & Desharnais, R.A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160-168.

Cliff, N. (1992). Abstract measurement theory and the revolution that has never happened. *Psychological Science, 3*, 186-190.

Clogg, C.C. (1977). *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users* (Working paper Nr. 1977-09). University Park, PA: Population Issues Research Center, Pennsylvania State University.

Cochran, W.G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics, 23*, 315-345.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conway, M. & Dewhurst, S. (1995). The self and recollective experience. *Applied Cognitive Psychology, 9*, 1-19.

Conway, M.A., Gardiner, J.M., Perfect, T.J., Anderson, S.J. & Cohen, G.M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General, 110*, 393-413.

Cosmides, L. & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best. Essays on evolution and optimality* (S. 277-306). Cambridge, Mass.: MIT Press.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J.H. Barkow, L. Cosmides & J. Tooby (Ed.), *The adapted mind. Evolutionary psychology and the generation of culture* (S. 163-228). Oxford: Oxford University Press.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Craik, F.I.M. (1994). Memory changes in normal aging. *Current Directions in Psychological Science, 3,* 155-158.

Cressie, N. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, B46,* 440-464.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43,* 171-192.

Crowther, C.S., Batchelder, W.H. & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review, 102,* 396-408.

Curran, T. & Hintzman, D.L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 531-547.

Curran, T. & Hintzman, D.L. (1997). Consequences and causes of correlations in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Daly, M. & Wilson, M. (1987). Evolutionary psychology and family violence. In C. Crawford, M. Smith & D. Krebs (Eds.), *Sociobiology and psychology: Ideas, issues, and applications* (S. 293-309). Hillsdale: Erlbaum.

Daly, M. & Wilson, M. (1988). *Homicide.* New York: Aldine.

Daly, M. & Wilson, M. (1996). Violence against stepchildren. *Current Directions in Psychological Science, 5,* 77-81.

Daly, M., Wilson, M. & Weghorst, S.J. (1982). Male sexual jealousy. *Ethology and Sociobiology, 3,* 11-27.

Davies, R.B. & Crouchly, R. (1986). The Mover-Stayer model: Requiescat in peace. *Sociological Methods & Research, 14,* 356-380.

Dawes, R.M. & Moore, M. (1980). Die Guttman-Skalierung orthodoxer und randomisierter Reaktionen. In F. Petermann (Ed.), *Einstellungsmessung – Einstellungsforschung* (S. 117-133). Göttingen: Hogrefe.

De Leeuw, J. (1994). Statistics and the sciences. In I. Borg & P.P. Mohler (Eds.), *Trends and perspectives in empirical social research* (S. 139-148). Berlin: de Gruyter.

Dehn, D. & Engelkamp, J. (1997). Process dissociation procedure: Double dissociations following divided attention and speeded responding. *The Quarterly Journal of Experimental Psychology, 50A,* 318-336.

Dehn, D. & Erdfelder, E. (1998). What kind of bias is hindsight bias? *Psychological Research, 61,* 135-146.

Dehn, D.M. (1998). *Automatische und kontrollierte Prozesse beim Erinnern von Handlungen: Methodische und theoretische Probleme.* Lengerich: Pabst Science Publisher.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via EM-algorithm. *Journal of the Royal Statistical Society, B, 39,* 1-22.

Deppe, W. (1977). *Formale Modelle in der Psychologie.* Stuttgart: Kohlhammer.

DeSteno, D.A. & Salovey, P. (1996). Evolutionary origins of sex differences in jealousy. *Psychological Science, 7,* 367-372.

Dingler, H. (1931). *Der Zusammenbruch der Wissenschaft und der Primat der Philosophie* (2. ed.). München: Reinhardt.

Dodson, C.S., Holland, P.W. & Shimamura, A.P. (1998). On the recollection of specific- and partial-source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1121-1136.

Dodson, C.S. & Johnson, M.K. (1996). Some problems with the process dissociation approach to memory. *Journal of Experimental Psychology: General, 125,* 181-194.

Dodson, C.S., Prinzmetal, W. & Shimamura, P. (1998). Using Excel to estimate parameters from observed data: An example from source memory data. *Behavior Research Methods, Instruments, & Computers, 30,* 517-526.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523-533.

Donaldson, W., MacKenzie, T.M. & Underhill, C.F. (1996). A comparison of recollective memory and source monitoring. *Psychonomic Bulletin & Review, 3,* 486-490.

Dorfman, D.D. & Alf, E.J. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervalls - rating method data. *Journal of Mathematical Psychology, 6,* 487-496.

Dorfman, D.D. & Berbaum, K.S. (1986). RSCORE-J: Pooled rating-method data: A computer program for analyzing pooled ROC curves. *Behavior Research Methods, Instruments, & Computers, 18,* 452-462.

Drost, F.C., Kallenberg, W.C.M., Moore, D.S. & Oosterhoff, J. (1989a). Asymptotic error bounds for power approximations to multinomial tests of fit. In L.J. Gleser, M.D. Perlman, S.J. Press & A.R. Sampson (Eds.), *Contributions to probability and statistics. Essays in honor of Ingram Olkin* (S. 429-446). New York: Springer.

Drost, F.C., Kallenberg, W.C.M., Moore, D.S. & Oosterhoff, J. (1989b). Power approximations to multinomial tests of fit. *Journal of the American Statistical Association, 84,* 130-141.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51,* 380-417.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B. & Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected information. *Biometrika, 65,* 457-487.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Egan, J.P. (1958). *Recognition memory and the operating characteristic* (Technical Note AFCRC-TN Nr. 58-51). Indiana University Hearing and Communication Laboratory.

Ehrlichman, H. & Barrett, J. (1983). Right hemisphere specialization for mental imagery: A review of the evidence. *Brain and Cognition, 2,* 55-76.

Einstein, G.O. & Hunt, R.R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 588-598.

Elandt-Johnson, R.C. (1971). *Probability models and statistical methods in genetics.* New York: John Wiley & Sons.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Engelkamp, J., Biegelmann, U. & McDaniel, M.A. (1998). Relational and item-specific information: Trade-off and redundancy. *Memory, 6,* 307-333.

Engen, T. (1972). Psychophysics. In J.W. Kling & L.A. Riggs (Eds.), *Experimental psychology* (S. 47-86). New York: Holt, Rinehart and Winston.

Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des β-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle. *Zeitschrift für Sozialpsychologie, 15,* 18-32.

Erdfelder, E. (1985). Die Überprüfung von Alternativhypothesen. Bemerkungen zum Kommentar von Pieter Koele. *Zeitschrift für Sozialpsychologie, 16,* 59-62.

Erdfelder, E. (1989). *Maximum likelihood analysis of binomial mixtures: A manual for users of BINOMIX* (Berichte aus dem Psychologischen Institut der Universität Bonn, Band 15, Heft 2). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Erdfelder, E. (1990a). Deterministic developmental hypotheses, probabilistic rules of manifestation, and the analysis of finite mixture distributions. In A. von Eye (Ed.), *Statistical methods in longitudinal research. Vol. 2: Time series and categorical longitudinal data* (S. 471-509). New York: Academic Press.

Erdfelder, E. (1990b). Probabilistische Ereignisbäume als restringierte univariate Latent-Class-Modelle. In D. Frey (Ed.), *Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990. Band 1* (S. 590-591). Göttingen: Hogrefe.

Erdfelder, E. (1991). Prädiktionsanalyse und die Analyse latenter Klassen: Welches Verfahren für welche Fragestellung? In A. von Eye (Ed.), *Prädiktionsanalyse* (S. 157-191). Weinheim: Psychologie Verlags Union.

Erdfelder, E. (1992a). *Multinomiale Modelle zum Hindsight-Bias-Paradigma*. Vortrag gehalten auf der 34. Tagung experimentell arbeitender Psychologen, Osnabrück, Germany, 12.-16. April, 1992.

Erdfelder, E. (1992b). *Parameter estimation and goodness-of-fit tests for Batchelder and Riefer's Storage-Retrieval Model: A manual for users of the SRM program* (Berichte aus dem Psychologischen Institut der Universität Bonn, Band 18, Heft 2). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Erdfelder, E. (1993a). BINOMIX: A BASIC program for maximum likelihood analyses of finite and beta-binomial mixture distributions. *Behavior Research Methods, Instruments, & Computers, 25*, 416-418.

Erdfelder, E. (1993b). Markov modeling of imagery and verbal processes in human memory. In R. Steyer, K.F. Wender & K.F. Widaman (Ed.), *Psychometric methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (S. 135-140). Stuttgart: Fischer.

Erdfelder, E. (1994). Erzeugung und Verwendung empirischer Daten. In T. Herrmann & W.H.Tack (Eds.), *Methodologische Grundlagen der Psychologie (= Enzyklopädie der Psychologie, Serie I, Band 1)* (S. 47-97). Göttingen: Hogrefe.

Erdfelder, E. (1996). Experiment. In G. Strube, B. Becker, C. Freksa, U. Hahn, K. Opwis & G. Palm (Ed.), *Wörterbuch der Kognitionswissenschaft* (S. 164-169). Stuttgart: Klett-Cotta.

Erdfelder, E. & Bayen, U.J. (1991). Episodisches Gedächtnis im Alter: Methodologische und empirische Argumente für einen Zugang über mathematische Modelle. In D. Frey (Ed.), *Bericht über den 37. Kongreß der*

Deutschen Gesellschaft für Psychologie in Kiel 1990, Band 2 (S. 172-180). Göttingen, Germany: Hogrefe.

Erdfelder, E., Brandt, M. & Bröder, A. (1995). *Multinomiale Verarbeitungsbaum-modellierung*. Vortrag gehalten auf der Methodenworkshop im Rahmen des DFG-Schwerpunktprogramms „Informationsverarbeitung im sozialen Kontext", Otzenhausen, 1. -3. 6. 95.

Erdfelder, E. & Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann & W.H.Tack (Eds.), *Methodologische Grundlagen der Psychologie (Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 1)* (S. 604-648). Göttingen: Hogrefe.

Erdfelder, E. & Bredenkamp, J. (1998). Recognition of script-typical versus script-atypical information: Effects of cognitive elaboration. *Memory & Cognition, 26,* 922-938.

Erdfelder, E., Bröder, A. & Brandt, M. (1998). Zur Bedeutung des Antwortformats für die Häufigkeit von Konjunktionsfehlern (Abstract). In H. Lachnit, A. Jacobs & F. Rösler (Ed.), *Experimentelle Psychologie. Abstracts der 40. Tagung experimentell arbeitender Psychologen, Marburg, 6.-9. 4. 1998* (S. 72-73). Lengerich: Pabst.

Erdfelder, E. & Buchner, A. (1995). *Process dissociation measurement models: Good versus better* (Berichte aus dem Psychologischen Institut der Universität Bonn, Band 21, Heft 3). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Erdfelder, E. & Buchner, A. (1998a). Decomposing the hindsight bias: A multinomial processing tree model for separating recollection and reconstruction in hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 387-414.

Erdfelder, E. & Buchner, A. (1998b). Process dissociation measurement models: Threshold theory or detection theory? *Journal of Experimental Psychology: General, 127*, 83-96.

Erdfelder, E., Faul, F. & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1-11.

Erdfelder, E. & Funke, J. (1988). *Entwicklung eines Polynomial-Tests für die Ausreißer- Alternative und Anwendung auf ein kognitionspsychologisches Beispiel* (Berichte aus dem Psychologischen Institut der Universität Bonn, Band 14, Heft 4). Bonn: Rheinische Friedrich-Wilhelms-Universität.

Erdfelder, E., Murnane, K. & Bayen, U.J. (1995). Die Messung kognitiver Prozesse im Paradigma der Quellendiskrimination. In K. Pawlik (Ed.), *Bericht über den 39. Kongreß der Deutschen Gesellschaft für Psychologie in Hamburg 1994* (S. 541-546). Göttingen: Hogrefe.

Ericsson, K.A. & Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis. Verbal reports as data (revised edition)*. Cambridge: MIT Press.

Ertel, S. (1965). Standardisierung eines Eindrucksdifferentials. *Zeitschrift für experimentelle und angewandte Psychologie, 12,* 22-58.

Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitteilungen der Anthropologischen Gesellschaft in Wien, 68,* 9-53.

Essler, W.K. (1982). *Wissenschaftstheorie I. Definition und Reduktion*. Freiburg: Verlag Karl Alber.

Essler, W.K. & Trapp, R. (1977). Some ways of operationally introducing dispositional predicates with regard to scientific and ordinary practice. In R. Tuomela (Ed.), *Dispositions* (S. 109-134). Dordrecht, Niederlande: Reidel.

Estes, W.K. (1956). The problem of inference from curves based on groups data. *Psychological Bulletin, 53,* 134-140.

Estes, W.K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology, 12,* 263-282.

Evans, J.S.B.T. (1977). Toward a statistical theory of reasoning. *Quarterly Journal of Experimental Psychology, 29,* 621-635.

Falmagne, J.-C. (1989). A latent trait theory via a stochastic learning theory for a knowledge space. *Psychometrika, 54,* 283-303.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P. & Johannesen, L. (1990). Introduction to knowledge spaces: How to built, test, and search them. *Psychological Review, 97,* 201-224.

Falmagne, J.C. (1985). *Elements of psychophysical theory*. Oxford: Clarendon Press.

Falmagne, J.C., Iverson, G.J. & Marcovici, S. (1979). Binaural loudness summation: Probabilistic theory and data. *Psychological Review, 86,* 25-43.

Fechner, G.T. (1860). *Elemente der Psychophysik.* Leipzig: Breitkopf und Härtel.

Feger, H. & Graumann, C.F. (1983). Beobachtung und Beschreibung von Erleben und Verhalten. In H. Feger & J. Bredenkamp (Eds.), *Datenerhebung (Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 2)* (S. 76-134). Göttingen: Hogrefe.

Ferguson, S.A., Hashtroudi, S. & Johnson, M.K. (1992). Age differences in using source relevant cues. *Psychology and Aging, 7,* 443-452.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123-129.

Fienberg, S.E. (1977). *The analysis of cross-classified categorical data.* Cambridge, MA: MIT Press.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

Fischhoff, B. (1982). For those condemned to study the past: Heuristics and biases in hindsight. In D. Kahneman, P. Slovic & A. Tversky (Ed.), *Judgment under*

uncertainty: Heuristics and biases (S. 335-351). New York: Cambridge University Press.

Flexser, A.J. & Tulving, E. (1978). Retrieval independence in recognition and recall. *Psychological Review, 85,* 153-171.

Formann, A.K. (1984). *Die Latent-Class-Analyse.* Weinheim: Beltz.

Formann, A.K. (1992). Latent class models with order restrictions. *Methodika, 6,* 131-149.

Funke, J. (1996). Methoden der kognitiven Psychologie. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 515-528). Weinheim: Psychologie Verlags Union.

Gadenne, V. (1976). *Die Gültigkeit psychologischer Untersuchungen*. Stuttgart: Kohlhammer.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: Mohr.

Gallistel, C.R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Gallo, D.A., Roberts, M.J. & Seamon, J.G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review, 4,* 271-276.

Garcia-Pérez, M.A. (1987). A finite state theory of performance in multiple-choice tests. In E.E. Roskam & R. Suck (Eds.), *Progress in Mathematical Psychology. Vol. I* (S. 455-464). Amsterdam: Elsevier.

Garcia-Pérez, M.A. (1989a). Item-sampling, guessing, partial information and decision-making in achievement testing. In E.E. Roskam (Ed.), *Mathematical psychology in progress* (S. 249-265). Berlin, Germany: Springer.

Garcia-Pérez, M.A. (1989b). La corrección del azar en pruebas objectivas: un enfoque basado en una nueva teoria de stados finitos. *Investigaciones Psicológicas, 6, 33-62.*

Garcia-Pérez, M.A. (1990). A comparison of two models of performance in objective tests: Finite states versus continuous distributions. *British Journal of Mathematical and Statistical Psychology, 43,* 73-91.

Garcia-Pérez, M.A. (1993). In defence of ´none of the above´. *British Journal of Mathematical and Statistical Psychology, 46,* 213-229.

Garcia-Pérez, M.A. (1994). Parameter estimation and goodness-of-fit testing in multinomial models. *British Journal of Mathematical and Statistical Psychology, 47,* 247-282.

Garcia-Pérez, M.A. (1999). MPROB: Computation of multinomial probabilities. *Behavior Research Methods, Instruments, & Computers, 31,* 701-705.

Garcia-Pérez, M.A. (in press). Exact finite-sample significance and confidence regions for goodness-of-fit statistics in one-way multinomials. *British Journal of Mathematical and Statistical Psychology, 53.*

Garcia-Pérez, M.A. & Frary, R.B. (1989). Psychometric properties of finite-state scores versus number-correct and formula scores: A simulation study. *Applied Psychological Measurement, 13,* 403-417.

Garcia-Pérez, M.A. & Frary, R.B. (1991a). Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology, 44,* 45-73.

Garcia-Pérez, M.A. & Frary, R.B. (1991b). Testing finite state models of performance in objective tests using items with ´none of the above´ as an option. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical Psychology: Current developments* (S. 273-291). New York: Springer.

Gardiner, J.M. & Gregg, V.H. (1997). Recognition memory with little or no remembering: Implications for a detection model. *Psychonomic Bulletin & Review, 4,* 474-479.

Gardiner, J.M. & Java, R.I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18,* 23-30.

Gardiner, J.M. & Java, R.I. (1993). Recognising and remembering. In A.F. Collins, S.E. Gathercole, M.A. Conway & P.E. Morris (Eds.), *Theories of memory* (S. 163-188). Hove, GB: Lawrence Erlbaum Associates.

Gardiner, J.M., Java, R.I. & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology, 50,* 114-122.

Gardiner, J.M. & Radomski, E. (1999). Awareness of recognition memory for Polish and English folk songs in Polish and English folk. *Memory, 7,* 461-470.

Gardiner, J.M., Ramponi, C. & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness and Cognition, 7,* 1-26.

Gardiner, J.M., Richardson-Klavehn, A. & Ramponi, C. (1997). On reporting recollective experiences and "direct access to memory systems". *Psychological Science, 8,* 391-394.

Gardiner, J.M., Richardson-Klavehn, A. & Ramponi, C. (1998). Limitations of the signal detection model of the remember-know paradigm: A reply to Hirshman. *Consciousness and Cognition, 7,* 285-288.

Garner, W.R. (1974). *The processing of information and structure.* Potomac, MD: Lawrence Erlbaum Associates.

Gegenfurtner, K. (1992). PRAXIS: Brent´s algorithm for function minimization. *Behavior Research Methods, Instruments, & Computers, 24,* 560-564.

Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie.* München: Ernst Reinhardt Verlag.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology. Vol. 2* (S. 83-115). New York: John Wiley & Sons.

Gigerenzer, G. (1998). Surrugates for theories. *Theory & Psychology, 8,* 195-204.

Gigerenzer, G. & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43,* 127-171.

Gigerenzer, G. & Richter, H.R. (1990). Context effects and their interaction with development. *Cognitive Development, 5,* 235-264.

Gigerenzer, G. & Strube, G. (1983). Are there limits to binaural additivity of loudness? *Journal of Experimental Psychology: Human Perception and Performance, 9,* 126-136.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1999). *Das Reich des Zufalls. Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen*. Heidelberg: Spektrum Akademischer Verlag.

Glanzer, M., Hilford, A., Kim, K. & Adams, J.K. (1999). Further tests of dual-process theory: A reply to Yonelinas (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 522-523.

Glanzer, M., Kim, K., Hilford, A. & Adams, J.K. (1999). Slope of the receiver-operating chracteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500-513.

Gleißner, U. & Erdfelder, E. (1994). *Heuristische und algorithmische Strategien beim Wason Selection Task*. Vortrag gehalten auf der 36. Tagung experimentell arbeitender Psychologen, München, 28.-31.3.1994.

Golden, R.M. (1995). Making correct statistical inferences using a wrong probability model. *Journal of Mathematical Psychology, 39,* 3-20.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215-231.

Goodman, L.A. (1979). On the estimation of parameters in latent structure analysis. *Psychometrika, 44,* 1979.

Goodman, L.A. & Kruskal, W. (1974a). Empirical evaluation of formal theory. *Journal of Mathematical Sociology, 3,* 187-196.

Goodman, L.A. & Kruskal, W. (1974b). More about empirical evaluation of formal theory. *Journal of Mathematical Sociology, 3,* 211-213.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16,* 1-14.

Graesser, A.C. & Nakamura, G.V. (1982). The impact of a schema on comprehension and memory. In G.H. Bower (Ed.), *The psychology of learning and motivation, Volume 16* (S. 59-109). New York: Academic Press.

Graf, P. & Komatsu, S.-I. (1994). Process-dissociation procedure: Handle with caution! *European Journal of Cognitive Psychology, 6,* 113-129.

Green, D.M. & Birdsall, T.G. (1978). Detection and recognition. *Psychological Review, 85,* 192-206.

Green, D.M. & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.

Green, E., Flynn, M.S. & Loftus, E.F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior, 21,* 207-219.

Greeno, J.G. (1968). Identifiability and statistical properties of two-stage learning with no successes in the initial stage. *Psychometrika, 33,* 173-216.

Greeno, J.G. & Bjork, R.A. (1973). Mathematical learning theory and the new „mental forestry". *Annual Review of Psychology, 24,* 81-116.

Greeno, J.G., James, C.T., DaPolito, F. & Polson, P.G. (1978). *Associate learning: A cognitive analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Greenwald, A.C. & Draine, S.C. (1997). Do subliminal stimuli enter the mind unnoticed? Tests with a new method. In J.D. Cohen & J.W. Schooler (Eds.), *Scientific approaches to consciousness* (S. 83-108). Mahwah, NJ: Erlbaum.

Greve, W. (1996). Erkenne Dich selbst? Argumente zur Bedeutung der "Perspektive der ersten Person". *Sprache & Kognition, 15,* 104-119.

Grier, J.B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75,* 424-429.

Grizzle, J.E., Starmer, C.F. & Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics, 25,* 489-504.

Groeben, N. & Scheele, B. (1977). *Argumente für eine Psychologie des reflexiven Subjekts.* Darmstadt: Steinkopff.

Groeben, N. & Westmeyer, H. (1981). *Kriterien psychologischer Forschung* (2. ed.). München: Juventa.

Guenther, W.C. (1977). Power and sample size for approximate chi-square tests. *American Statistician, 31,* 83-85.

Guttman, L. (1981). What is not what in theory construction. In I. Borg (Ed.), *Multidimensional data representations: When and why* (S. 47-64). Ann Arbor, MI: Mathesis Press.

Haberman, S.J. (1978). *Analysis of qualitative data (Volume 1).* New York: Academic Press.

Haberman, S.J. (1979). *Analysis of qualitative data (Volume 2).* New York: Academic Press.

Hager, W. (1987). Grundlagen einer Versuchsplanung zu Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Ed.), *Allgemeine Experimentelle Psychologie* (S. 43-264). Stuttgart: Gustav Fischer Verlag.

Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen.* Göttingen: Hogrefe.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Eds.), *Hypothesenprüfung (Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Band 5)* (S. 24-238). Göttingen: Hogrefe.

Hall, P. & Wilson, S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics, 47,* 757-762.

Hamilton, D.L. & Gifford, R.K. (1976). Illusory correlations in interpersonal perception: A cognitive basis for stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392-407.

Harris, C.R. & Christenfeld, N. (1996). Gender, jealousy, and reason. *Psychological Science, 7,* 364-366.

Harvey, P.D. (1985). Reality monitoring in mania and schizophrenia. *Journal of Nervous and Mental Disease, 173,* 67-72.

Hasher, L., Stoltzfus, E.R., Zacks, R.T. & Rypma, B. (1991). Age and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 163-169.

Hautus, M.J. (1997). Calculating estimates of sensitivity from group data: Pooled vs. averaged estimators. *Behavior Research Methods, Instruments, & Computers, 29,* 556-562.

Hawkins, S.A. & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107,* 311-327.

Hays, W.L. (1973). *Statistics for the social sciences* (2. ed.). New York: Holt, Rinehart and Winston.

Hell, W. (1993). Gedächtnistäuschungen. In W. Hell, K. Fiedler & G. Gigerenzer (Ed.), *Kognitive Täuschungen* (S. 13-38). Heidelberg: Spektrum Akademischer Verlag.

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M. & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition, 16,* 533-538.

Henkel, L.A. & Franklin, N. (1998). Reality monitoring of physically similar and conceptually related objects. *Memory & Cognition, 26,* 659-673.

Henkel, L.A., Johnson, M.K. & DeLeonardis, D.M. (1998). Aging and source monitoring: Cognitive processes and neuropsychological correlates. *Journal of Experimental Psychology: General, 127,* 251-268.

Herrmann, T. (1979). *Psychologie als Problem. Herausforderungen der psychologischen Wissenschaft.* Stuttgart: Klett-Cotta.

Herrmann, T. (1983). Nützliche Fiktionen. Anmerkungen zur Funktion kognitionspsychologischer Theoriebildung. *Sprache & Kognition, 2,* 88-99.

Herrmann, T. & Stapf, K.H. (1971). Über theoretische Konstruktionen in der Psychologie. *Psychologische Beiträge, 13,* 336-354.

Hertwig, R. (1994). *Why Dr. Gould´s homonculus doesn´t think like Dr. Gould: The "conjunction fallacy" reconsidered.* Unveröffentlichte Doktorarbeit, Universität Konstanz, Konstanz.

Hertwig, R. & Chase, V.M. (in press). Many reasons or just one: How response mode affects reasoning in the conjunction problem. *Thinking & Reasoning.*

Hertwig, R., Gigerenzer, G. & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review, 104,* 194-202.

Hicks, J.L. & Marsch, R.L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review, 6,* 117-122.

Higham, P.A. (1998). Believing details known to have been suggested. *British Journal of Psychology, 89,* 265-283.

Hildebrand, D.K., Laing, J.D. & Rosenthal, H. (1977). *Prediction analysis of cross classifications.* New York: John Wiley & Sons.

Hillstrom, A.P. & Logan, G.D. (1997). Process dissociation, cognitive architecture, and response time: Comments on Lindsay and Jacoby (1994). *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1561-1578.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple trace memory model. *Psychological Review, 93,* 411-428.

Hintzman, D.L. (1988). Judgments of frequency and recognition memory in a multiple trace model. *Psychological Review, 95,* 528-551.

Hintzman, D.L. (1991). Why are formal models useful in psychology? In W.E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (S. 39-56). Hillsdale, NJ: Erlbaum.

Hintzman, D.L. (1993). On variability, Simpson's paradox, and the relation between recognition and recall: Reply to Tulving and Flexser. *Psychological Review, 100,* 143-148.

Hintzman, D.L. & Curran, T. (1997). More than one way to violate independence: Reply to Jacoby and Shrout (1997). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 511-513.

Hirshman, E. (1998a). On the logic of testing the independence assumption in the process-dissociation procedure. *Memory & Cognition, 26,* 857-859.

Hirshman, E. (1998b). On the utility of the signal detection model of the remember-know paradigm. *Consciousness and Cognition, 7,* 103-107.

Hirshman, E. & Bjork, R.A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning,Memory,and Cognition, 14,* 484-494.

Hirshman, E. & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science, 9,* 61-65.

Hirshman, E. & Lanning, K. (1999). Is there a special association between self judgements and conscious recollection? *Applied Cognitive Psychology, 13,* 29-42.

Hirshman, E. & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory & Cognition, 25,* 345-351.

Hodge, M.H. & Otani, H. (1996). Beyond category sorting and pleasantness rating: Inducing relational and item-specific processing. *Memory & Cognition, 24,* 110-115.

Hodkin, B. (1987). Performance model analysis in class inclusion: An illustration with two language conditions. *Developmental Psychology, 23,* 683-689.

Horton, K.D. & Vaughan, D.C. (1999). Analyzing estimates of automatic and conscious retrieval in between- and within-subjects designs. *Behavior Research Methods, Instruments, & Computers, 31,* 347-352.

Hosmane, B. (1987). An empirical investigation of chi-square tests for the hypothesis of no three-factor interaction in I x J x K contingency tables. *Journal of Statistical Computation and Simulation, 28,* 167-178.

Howe, M.L. (1991). Misleading children´s story recall: Forgetting and reminiscence of facts. *Developmental Psychology, 27,* 746-762.

Howe, M.L. (1995). Interference effects in young children´s long-term retention. *Developmental Psychology, 31,* 579-596.

Howe, M.L. & Brainerd, C.J. (1989). Development of children´s long-term retention. *Developmental Review, 9,* 301-340.

Howe, M.L., Courage, M.L. & Bryant-Brown, L. (1993). Reinstating preschoolers´ memories. *Developmental Psychology, 29,* 854-869.

Howe, M.L. & O´Sullivan, J.T. (1997). What children´s memories tell us about recalling our childhoods: A review of storage and retrieval processes in the development of long-term retention. *Developmental Review, 17,* 148-204.

Howe, M.L. & Rabinowitz, M.F. (1996). Reasoning from memory: A lifespan inquiry into the necessity of remembering when reasoning about class inclusion. *Journal of Experimental Child Psychology, 61,* 1-42.

Hu, X. (1990). *Computer program for source monitoring (SOURCE, Version 1.0)* [Computer-Programm]. Irvine, CA: University of California at Irvine, School of Social Sciences.

Hu, X. (1995). *Statistical inference program for multinomial binary tree models (MBT, Version 1.1)* [Computer-Programm]. Irvine, CA: University of California at Irvine, School of Social Sciences.

Hu, X. (1998). *General processing tree models (GPT, version 1.0)* [Computer-Programm]. Memphis, TN: University of Memphis.

Hu, X. (1999a). *General processing tree (GPT) models with non-binary parameters.* Manuscript submitted for publication.

Hu, X. (1999b). Multinomial processing tree models: An implementation. *Behavior Research Methods, Instruments, & Computers, 31,* 689-695.

Hu, X. & Batchelder, W.H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59,* 21-47.

Hu, X., Marks, W. & Isenberg, A. (1996). *Retrieval inhibition in directed forgetting: A source monitoring approach*. Vortrag gehalten auf der 29th Annual Mathematical Psychology Meeting, Chapel Hill, NC, August 1996.

Hu, X. & Phillips, G.A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments, & Computers, 31,* 220-234.

Hulme, C., Roodenrys, S. & Schweikert, R. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1217-1232.

Humphreys, M.S. & Bain, J.D. (1983). Recognition memory: A cue and information analysis. *Memory & Cognition, 11,* 583-600.

Humphreys, M.S. & Bowyer, P.A. (1980). Sequential testing effects and the relationship between recognition and recognition failure. *Memory & Cognition, 8,* 271-277.

Humphreys, M.S. & Greeno, J.G. (1970). Interpretation of the two-stage analysis of paired-associate memorizing. *Journal of Mathematical Psychology, 7,* 275-292.

Hunt, R.R. & Einstein, G.O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20,* 497-514.

Hurwicz, L. (1950). Generalization of the concept of identification. In T.C. Koopmans (Ed.), *Statistical inference in dynamic economic models* (S. 245-257). New York: Wiley.

Hutchinson, T.P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. *British Journal of Mathematical and Statistical Psychology, 35,* 71-89.

Inoue, C. & Bellezza, F.S. (1998). The detection model of recognition using know and remember judgments. *Memory & Cognition, 26,* 299-308.

Iseler, A. (1999). *Explikative und deduktive Komponenten der "Herleitung" statistischer Vorhersagen aus psychologischen Hypothesen*. Vortrag gehalten auf der 4. Tagung der Fachgruppe Methoden in der Deutschen Gesellschaft für Psychologie e.V., Leipzig, 23. 9. - 25. 9. 1999.

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513-541.

Jacoby, L.L. (1998). Invariance in automatic influences of memory: Toward a user´s guide for the process dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1-36.

Jacoby, L.L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 3-22.

Jacoby, L.L., Begg, I.M. & Toth, J.P. (1997). In defense of functional independence: Violations of assumptions underlying the process dissociation procedure? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 484-495.

Jacoby, L.L. & Shrout, P.E. (1997). Toward a psychometric analysis of violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 505-510.

Jacoby, L.L., Yonelinas, A.P. & Jennings, J.M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J.D. Cohen & J.W. Schooler (Eds.), *Scientific approaches to consciousness* (S. 13-47). Mahwah, NJ: Lawrence Erlbaum Associates.

Jöhnk, M.D. (1964). Erzeugung von beta-verteilten und gamma-verteilten Zufallszahlen. *Metrika, 8,* 5-15.

Johnson, M.K., Foley, M.A. & Leach, K. (1988). The consequences for memory of imagining in another person´s voice. *Memory & Cognition, 16,* 337-342.

Johnson, M.K., Kounios, J. & Reeder, J.A. (1994). Time-course studies of reality monitoring and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1409-1419.

Johnson, M.K., Nolde, S.F. & De Leonardis, D.M. (1996). Emotional focus and source monitoring. *Journal of Memory and Language, 35,* 135-156.

Johnson, N.L. & Kotz, S. (1970a). *Distributions in statistics. Continuous univariate distributions - 1.* New York: Wiley.

Johnson, N.L. & Kotz, S. (1970b). *Distributions in statistics. Continuous univariate distributions - 2.* New York: Wiley.

Johnson, N.L. & Kotz, S. (1972). *Distributions in statistics: Continuous multivariate distributions.* New York: Wiley.

Johnson-Laird, P.N. & Byrne, R.M.J. (1991). *Deduction.* Hove: Erlbaum.

Jones, M.R. (1971). From probability learning to sequential processing: A critical review. *Psychological Bulletin, 76,* 153-185.

Joordens, S. & Merikle, P.M. (1993). Independence or redundancy? Two models of conscious and unconscious influences. *Journal of Experimental Psychology: General, 122,* 462-467.

Kaernbach, C. (1991). Poisson signal-detection theory: Link between threshold models and the Gaussian assumption. *Perception & Psychophysics, 50,* 498-506.

Kail, R., Hale, C.A., Leonard, L.B. & Nippold, M. (1984). Lexikal storage and retrieval in language-impaired children. *Applied Psycholinguistics, 5,* 37-49.

Kallenberg, W.C.M., Oosterhoff, J. & Schriever, B.F. (1985). The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association, 80,* 959-968.

Kant, I. (1798/1968). Anthropologie in pragmatischer Hinsicht. In W. Weischedel (Ed.), *Immanuel Kant. Schriften zur Anthropologie, Geschichtsphilosophie, Politik und Pädagogik 2* (S. 397-690). Frankfurt: Suhrkamp (Original erschienen 1798).

Kaufman, L. (1979). *Perception. The world transformed.* New York: Oxford University Press.

Keefe, R.S.E., Courtney, M.C., Bayen, U.J. & Harvey, P.D. (1999). Source monitoring deficits in patients with schizophrenia: A multinomial modeling analysis. *Psychological Medicine, 29,* 903-914.

Kendler, T.S. & Kendler, H.H. (1959). Reversal and nonreversal shifts in kindergarten children. *Journal of Experimental Psychology, 58,* 56-60.

Kihlstrom, J.F., Barnhardt, T.M. & Tataryn, D.J. (1992). Implicit perception. In R.F. Bornstein & T.S. Pittman (Ed.), *Perception without awareness* (S. 17-54). New York: Guilford Press.

Kikuno, H. (1995). A study of the suggestibility effect on eyewitness testimony - discussion on the integration and the coexist hypothesis. *Japanese Journal of Psychology, 66,* 116-120.

Kinchla, R.A. (1994). Comments on Batchelder and Riefer's multinomial model for source monitoring. *Psychological Review, 101,* 166-171.

Kintsch, W. (1970). *Learning, memory, and conceptual processes.* New York: John Wiley & Sons.

Kintsch, W. & Morris, C.J. (1965). Application of a Markov model to free recall and recognition. *Journal of Experimental Psychology, 69,* 200-206.

Klauer, K.C. (1988). Tests von Latent-Trait-Annahmen: Ein Schluß von der Population auf die Person. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9,* 97-104.

Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213-228.

Klauer, K.C. (1996a). Parameterschätzung. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 99-107). Weinheim: Psychologie Verlags Union.

Klauer, K.C. (1996b). Urteilerübereinstimmung bei dichotomen Kategoriensystemen. *Diagnostica, 42,* 101-118.

Klauer, K.C. & Batchelder, W.H. (1996). Structural analysis of subjective categorical data. *Psychometrika, 61,* 199-240.

Klauer, K.C. & Meiser, T. (in press). A source-monitoring analysis of illusory correlations. *Personality and Social Psychology Bulletin.*

Klauer, K.C., Musch, J. & Naumer, B. (1999). *On belief bias in syllogistic reasoning.* Manuscript submitted for publication.

Klauer, K.C. & Oberauer, K. (1995). Testing the mental model theory of propositional reasoning. *The Quarterly Journal of Experimental Psychology, 48A*, 671-687.

Klauer, K.C. & Wegener, I. (1998). Unraveling social categorization in the "Who Said What?" paradigm. *Journal of Personality and Social Psychology, 75*, 1155-1178.

Klauer, K.C., Wegener, I. & Ehrenberg, K. (1999). *Person-based and category-based processing: The effects of relative group size.* Manuscript submitted for publication.

Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Klein, K.-M. (1989). *Gedächtnispsychologische Untersuchungen an epileptischen Patienten*. Unveröffentlichte Diplomarbeit, Universität Bonn, Bonn.

Kliegl, R. & Lindenberger, U. (1993). Modeling intrusions and correct recall in episodic memory: Adult age differences in encoding of list context. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 617-637.

Kohnert, A. (1996). *Grenzen des Rückschaufehlers. Die Verzerrung von Erinnerungen an früheres Wissen durch neue Informationen*. Bonn: Holos.

Komatsu, S.-I., Graf, P. & Uttl, B. (1995). Process-dissociation procedure: Core assumptions fail, sometimes. *European Journal of Cognitive Psychology, 7*, 19-40.

Koopmans, T.C. & Reiersøl, O. (1950). The identification of structural characteristics. *Annals of Mathematical Statistics, 21*, 165-181.

Krantz, D.H. (1969). Threshold theories of signal detection. *Psychological Review, 76*, 308-324.

Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement. Volume 1: Additive and polynomial representations.* New York: Academic Press.

Krauth, J. (1982). Formulation and experimental verification of models in propositional reasoning. *Quarterly Journal of Experimental Psychology, 34A*, 285-298.

Krauth, J. & Lienert, G.A. (1995). *Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin (Reprint).* Weinheim: Psychologie Verlags Union.

Krengel, U. (1991). *Einführung in die Wahrscheinlichkeitstheorie und Statistik* (3. ed.). Braunschweig: Vieweg.

Krüger, T. (1995). *Untersuchungen zur Beziehung bewußter und unbewußter Gedächtnisprozesse.* Diplomarbeit, Universität Bonn, Bonn.

Krüger, T. (1999). *Die Erfassung bewußter und unbewußter Gedächtnisprozesse. Die "Prozeß-Dissoziations-Prozedur" - Probleme und Perspektiven einer neuen Methode.* Lengerich: Pabst.

Krüger, T. & Vaterrodt-Plünnecke, B. (1997). Die stochastische Beziehung bewußter und automatischer Gedächtnisprozesse: Eine Erweiterung der Prozeß-Dissoziationsprozedur. *Zeitschrift für Experimentelle Psychologie, 44,* 220-245.

Kuhn, T.S. (1962). *The structure of scientific revolutions.* Chicago: The University of Chicago Press.

Kumbasar, E. (in press). Methods for analyzing three-way cognitive network data. *Journal of Quantitative Anthropology.*

Kumbasar, E., Romney, E.K. & Batchelder, W.H. (1994). Systematic biases in social perception. *American Journal of Sociology, 100,* 477-505.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (S. 91-195). London: Cambridge University Press.

Langeheine, R. (1986). Log-lineare Modelle. In H.J. Hummel, W. Jagodzinski & R. Langeheine (Eds.), *Techniken der empirischen Sozialforschung: Band 8. Kausalanalyse* (S. 122-195). München: R. Oldenburg Verlag.

Langeheine, R. (1988). New developments in latent class theory. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (S. 77-108). New York: Plenum.

Langeheine, R., Pannekoek, J. & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24,* 492-516.

Langeheine, R. & Rost, J. (1996). Latent-Class-Analyse. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 315-332). Weinheim: Psychologie Verlags Union.

Langeheine, R. & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research, 18,* 416-441.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association, 73,* 253-263.

Lasdon, L.S., Waren, A., Jain, A. & Ratner, M. (1978). Design and testing of a generalized reduced gradient code for nonlinear programming. *A C M Transactions on Mathematical Software, 4,* 34-50.

Lazarsfeld, P.F. & Henry, N.W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lee, C.C. (1987). Chi-squared tests for and against an order restriction on multinomial parameters. *Journal of the American Statistical Association, 82,* 611-618.

Levin, B. (1983). On calculations involving the maximum cell frequency. *Communications in Statistics - Theory and Methods, 12,* 1299-1327.

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology, 71,* 331-338.

Levy, K.J. (1980). The randomized response technique and large sample tests concerning the parameters of a multinomial distribution. *Educational and Psychological Measurement, 40,* 701-707.

Lewin, K. (1951/1982). Formalisierung und Fortschritt in der Psychologie. In C.F. Graumann (Ed.), *Feldtheorie. Kurt-Lewin-Werkausgabe, Band 4* (S. 41-72). Bern: Huber (Original erschienen 1951).

Li, C.C. & Chakravarti, A. (1985). Basic fallacies in the formulation of the paternity index. *American Journal of Human Genetics, 37,* 809-818.

Lienert, G.A. (1973). *Verteilungsfreie Methoden in der Biostatistik. Band 1.* Meisenheim: Hain.

Lienert, G.A. (Ed.). (1988). *Angewandte Konfigurationsfrequenzanalyse.* Frankfurt: Athenäum.

Light, L.L., LaVoie, D., Valencia-Laver, D., Albertson Owens, S.A. & Mead, G. (1992). Direct and indirect measures of memory for modality in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1284-1297.

Lindsay, D.S. (1990). Misleading suggestions can impair eyewitnesses' ability to remember event details. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1077-1083.

Lindsay, D.S. & Johnson, M.K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17,* 349-358.

Lindsay, D.S., Johnson, M.K. & Kwon, P. (1991). Developmental changes in memory source monitoring. *Journal of Experimental Child Psychology, 52,* 297-318.

Link, S.W. (1982). Correcting response measures for guessing and partial information. *Psychological Bulletin, 92,* 469-486.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Loftus, E.F. (1979). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Loftus, E.F., Miller, D.G. & Burns, H.J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19-31.

Loftus, E.F., Schooler, J.W. & Wagenaar, W.A. (1985). The fate of memory: Comment on McCloskey and Zaragoza. *Journal of Experimental Psychology: General, 114,* 375-380.

Luce, R.D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R.D. (1963a). Detection and recognition. In R.D. Luce, R.R. Bush & E. Galanter (Eds.), *Handbook of Mathematical Psychology. Vol. 1* (S. 245-308). New York: John Wiley & Sons.

Luce, R.D. (1963b). A threshold theory for simple detection experiments. *Psychological Review, 70,* 61-79.

Luce, R.D. (1986). *Response times. Their role in infering elementary mental organization*. New York: Oxford University Press.

Luce, R.D., Krantz, D.H., Suppes, P. & Tversky, A. (1990). *Foundations of measurement. Volume III.* New York: Academic Press.

Lukas, J. (1996). *Psychophysik der Raumwahrnehmung*. Weinheim: Psychologie Verlags Union.

Macmillan, N.A. & Creelman, C.D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin, 107,* 401-413.

Macmillan, N.A. & Creelman, C.D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

Macmillan, N.A. & Kaplan, H.L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false alarm rates. *Psychological Bulletin, 98,* 185-199.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252-271.

Manning, C.G. & Loftus, E.F. (1996). Eyewitness testimony and memory distortion. *Japanese Psychological Research, 38,* 5-13.

Marche, T.A. & Howe, M.L. (1995). Preschoolers report misinformation despite accurate memory. *Developmental Psychology, 31,* 554-567.

Maris, E. (1992). *Psychometric models for psychological processes and structures*. Unveröffentlichte Dissertation, Katholieke Universiteit, Leuven, Belgien.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

Marsh, R.L. & Hicks, J.L. (1998). Test formats change source-monitoring decision processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1137-1151.

Martinez, J. & Kesner, R. (Ed.). (1998). *Neurobiology of learning and memory*. San Diego, CA: Academic Press.

Massaro, D.W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McBride, D.M. & Dosher, B.A. (1999). Forgetting rates are comparable in conscious and automatic memory: A process-dissociation study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 583-607.

McCloskey, M. & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against the memory impairment hypothesis. *Journal of Experimental Psychology: General, 114,* 1-16.

McDaniel, M.A., Waddill, P.J. & Einstein, G.O. (1988). A contextual account of the generation effect: A three factor theory. *Journal of Memory and Language, 27,* 521-536.

McGurk, H. & MacDonald, T. (1976). Hearing lips and seeing vioces. *Nature, 264,* 746-748.

McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika, 21,* 331-347.

McLachlan, G.J. & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: Wiley.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103-115.

Meiser, T. (2000). Gedächtnisvorteil ausgeführter Handlungen: Anmerkungen zur multimodalen Gedächtnistheorie. *Zeitschrift für Experimentelle Psychologie, 47,* 1-12.

Meiser, T. & Bröder, A. (1999). *Memory for multidimensional source information.* Manuscript submitted for publication.

Meiser, T. & Hewstone, M. (1999). *Crossed categorization effects on the formation of illusory correlations.* Manuscript submitted for publication.

Meng, R.C. & Chapman, D.G. (1966). The power of chi square tests for contingency tables. *Journal of the American Statistical Association, 61,* 965-975.

Meyer, M. (1994). *Serielle versus parallele Verarbeitung visueller Information. Eine experimentelle Untersuchung am Beispiel eines Konzentrationstests.* Bonn: Holos.

Mickey, M.R., Gjertson, D.W. & Terasaki, P.I. (1986). Empirical validation of the Essen-Möller probability of paternity. *American Journal of Human Genetics, 39,* 123-132.

Miller, E. & Lewis, P. (1977). Recognition memory in elderly patients with depression and dementia: A signal detection analysis. *Journal of Abnormal Psychology, 86,* 84-86.

Mitra, S.K. (1958). On the limiting power function of the frequency chi-square test. *Annals of Mathematical Statistics, 29,* 1221-1233.

Molenaar, I.W. & Lewis, C. (1996). Bayes-Statistik. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 145-156). Weinheim: Psychologie Verlags Union.

Mooijaart, A. & van der Heijden, P.G.M. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika, 57,* 261-269.

Moore, D.S. (1986). Tests of chi-squared type. In R.B. D´Agostino & M.A. Stephens (Eds.), *Goodness-of-fit techniques* (S. 63-95). New York: Marcel Dekker, Inc.

Mulder, G., Sanders, A.F. & van Galen, G.P. (1995). Discrete and continuous information processing (special issue). *Acta Psychologica, 90.*

Müller, B. (1996a). Nutzung von Wissen über Funktionskonzepte: Regel-, struktur-, oder beispielbasiert. *Zeitschrift für Experimentelle Psychologie, 43,* 144-174.

Müller, B. (1996b). Zulässigkeitsurteile und Bediensequenzen: Der Einfluß zeitlicher Sukzessivität auf die Kompositionsbildung. *Zeitschrift für Psychologie, 204,* 281-303.

Müller, B. & Funke, J. (1995). Diskrete dynamische Systeme: Kompositionen höherer Ordnung bei Bediensequenzen. *Zeitschrift für Psychologie, 203,* 153-172.

Mulligan, N.W. (1996). The effects of perceptual interference at encoding on implicit memory, explicit memory, and memory for source. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1067-1087.

Mulligan, N.W. & Hirshman, E. (1997). Measuring the bases of recognition memory: An investigation of the process dissociation framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 280-304.

Multhaup, K.S., De Leonardis, D.M. & Johnson, M.K. (1999). Source memory and eyewitness suggestibility in older adults. *Journal of General Psychology, 126,* 74-84.

Murdock, B.B. (1974). *Human memory: Theory and data.* Potomac, MD: Lawrence Erlbaum Associates.

Murnane, K. & Bayen, U.J. (1998). Measuring memory for source: Some theoretical assumptions and technical limitations. *Memory & Cognition, 26,* 674-677.

Nagin, D.S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods, 4,* 139-157.

Nakatani, L.H. (1970). Comments on Broadbent´s response bias model for stimulus recognition. *Psychological Review, 77,* 574-576.

Nash, J.C. (1979). *Compact numerical methods for computers: Linear algebra and function minimization.* New York: Halsted Press.

Neyman, J. (1949). Contribution to the theory of the χ^2 test. *Proceedings of the first Berkeley symposium on mathematical statistics and probability,* 239-273.

Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16,* 1-32.

Niederée, R. (1994). There is more to measurement than just measurement: Measurement theory, symmetry, and substantive theorizing. *Journal of Mathematical Psychology, 38,* 527-594.

Niederée, R. & Mausfeld, R. (1996). Skalenniveau, Invarianz und „Bedeutsamkeit". In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 385-398). Weinheim: Psychologie Verlags Union.

Nilsson, L.-G. & Gardiner, J.M. (1993). Identifying exceptions in a database of recognition failure studies from 1973 to 1992. *Memory & Cognition, 21,* 397-410.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Nishimura, H., Hashikawa, K., Doi, K., Iwaki, T., Watanabe, Y., Kusuoka, H., Nishimura, T. & Kubo, T. (1999). Sign language ´heard´ in the auditory cortex. *Nature, 397,* 116.

O'Grady, K.E. (1981). Probabilities and critical values for z, chi square, r, t, and F. *Behavior Research Methods & Instrumentation, 13,* 55-56.

Oakes, P.J. (1987). The salience of social categories. In C.J. Turner, M.A. Hogg, P.J. Oakes, S.D. Reicher & M.S. Wetherell (Eds.), *Rediscovering the social group: A self categorization theory* (S. 117-141). Oxford: Blackwell.

Oldenbürger, H.-A. (1996). Exploratorische, graphische und robuste Datenanalyse. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 71-86). Weinheim: Psychologie Verlags Union.

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science, 6,* 155-156.

Olson, C.L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin, 83,* 579-586.

Ortega, J.M. & Rheinboldt, W.C. (1970). *Iterative solution of nonlinear equations in several variables.* New York: Academic Press.

Ott, R., Curio, I. & Scholz, O.B. (2000). Implicit memory to auditorily presented threatening stimuli - A process-dissociation approach. *Perceptual and Motor Skills, 90,* 131-146.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart, and Winston.

Paivio, A. (1986). *Mental representations. A dual coding approach.* Oxford, GB: Oxford University Press.

Paivio, A., Yuille, J., C. & Madigan, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph, 76.*

Payne, D.G., Elie, C.J., Blackwell, J.M. & Neuschatz, J.S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language, 35,* 261-285.

Payne, D.G., Toglia, M.P. & Anastasi, J.S. (1994). Recognition performance level and the magnitude of the misinformation effect in eyewitness memory. *Psychonomic Bulletin & Review, 1,* 376-382.

Perruchet, P. & Amorim, M.-A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 785-800.

Pfister, H.-R. & Fischer, K. (1992). *Hindsight-Bias bei Wissensfragen: Urteilsstrategie und Encodierungsspezifität.* (Technischer Bericht Nr. Nr. JU 53/92.). Berlin: TU Berlin, Institut für Psychologie.

Phillips, S. & Levine, M. (1975). Probing for hypotheses with adults and children: Blank trials and introtacts. *Journal of Experimental Psychology: General, 104,* 327-354.

Pohl, R.F. (1992). Der Rückschau-Fehler: Systematische Verfälschung der Erinnerung bei Experten und Novizen. *Kognitionswissenschaft, 3,* 38-44.

Pohl, R.F. (1995). Der Rückschaufehler - Eine systematische Verfälschung der Erinnerung. *Report Psychologie, 21,* 596-609.

Pohl, R.F. (1998). The effects of feedback source and plausibility on hindsight bias. *European Journal of Cognitive Psychology, 20,* 191-212.

Pohl, R.F. & Gawlik, B. (1995). Hindsight bias and the misinformation effect: Separating blended recollections from other recollection types. *Memory, 3,* 21-55.

Pohl, R.F. & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes, 67,* 49-58.

Polson, P.G. & Huizinga, D. (1974). Statistical methods for absorbing Markov-chain models for learning: Estimation and identification. *Psychometrika, 39,* 3-22.

Pongratz, L.J. (1998). Die Kontroverse zwischen Wilhelm Wundt (1832-1920) und Karl Bühler (1879-1963). Analyse einer Wende der Psychologie. In W. Baumgartner, F.-P. Burkhard & F. Wiedman (Eds.), *Zur Entwicklung und Bedeutung der Würzburger Schulen (Brentano-Studien, Band 7)* (S. 255-266). Dettelbach: Röll.

Popper, K.R. (1934/1982). *Logik der Forschung* (7. ed.). Tübingen: J.C.B. Mohr (Original erschienen 1934).

Postman, L. & Underwood, B.J. (1973). Critical issues in interference theory. *Memory & Cognition, 1,* 19-40.

Powell, M.J.D. (1962). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal, 7,* 155-162.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical recipes in FORTRAN. The art of scientific computing* (2. ed.). Cambridge: Cambridge University Press.

Prinzmetal, W., Henderson, D. & Ivry, R. (1995). Loosening the constraints on illusory conjunctions: Assessing the roles of exposure duration and attention.

Journal of Experimental Psychology: Human Perception and Performance, 21, 1362-1375.

Puff, C.R. (Ed.). (1982). *Handbook of research methods in human memory and cognition.* New York: Academic Press.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, B11,* 68-84.

Rabinowitz, F.M., Howe, M.L. & Lawrence, J.A. (1989). Class inclusion and working memory. *Journal of Experimental Child Psychology, 48,* 379-409.

Raftery, A.E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, B48,* 249-250.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21,* 89-102.

Rajaram, S. & Roediger, H.L.I. (1997). Remembering and knowing as states of consciousness during retrieval. In J.D. Cohen & J.W. Schooler (Eds.), *Scientific approaches to consciousness* (S. 213-240). Mahwah, NJ: Erlbaum.

Rao, C.R. (1962). Efficient estimates and optimum inference in large samples (with discussion). *Journal of the Royal Statistical Society, B 24,* 46-72.

Rao, C.R. (1963). Criteria of estimation in large samples. *Sankhya, A44,* 189-206.

Rasch, G. (1960 / 1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Read, T.R.C. (1984). Small-sample comparisons for the power divergence goodness-of-fit statistics. *Journal of the American Statistical Association, 79,* 929-935.

Read, T.R.C. & Cressie, N.A.C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review, 69,* 329-343.

Rey, W.J.J. (1978). *Robust statistical methods.* New York: Springer.

Richardson, J.T.E. (1990). Imagery and memory in brain-damaged patients. In P.J. Hampson, D.F. Marks & J.T.E. Richardson (Eds.), *Imagery. Current developments* (S. 351-372). London: Routledge.

Richardson-Klavehn, A. & Bjork, R.A. (1988). Measures of memory. *Annual Review of Psychology, 39,* 475-543.

Richardson-Klavehn, A., Gardiner, J.M. & Java, R.I. (1995). Memory: Task dissociations, process dissociations, and dissociations of consciousness. In G. Underwood (Ed.), *Implicit cognition* (S. 85-158). Oxford: Oxford University Press.

Riefer, D.M. (1982). The advantages of mathematical modeling over traditional methods in the analysis of category clustering. *Journal of Mathematical Psychology, 26,* 97-123.

Riefer, D.M. & Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318-339.

Riefer, D.M. & Batchelder, W.H. (1991a). Age differences in storage and retrieval: A multinomial modeling analysis. *Bulletin of the Psychonomic Society, 29,* 415-418.

Riefer, D.M. & Batchelder, W.H. (1991b). Statistical inference for multinomial processing tree models. In J.-P. Doignon & J.-C. Falmagne (Eds.), *Mathematical Psychology: Current developments* (S. 313-335). New York: Springer.

Riefer, D.M. & Batchelder, W.H. (1995). A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition, 23,* 611-630.

Riefer, D.M., Hu, X. & Batchelder, W.H. (1994). Response strategies in source monitoring. *Journal of Experimental Psychology Learning, Memory, and Cognition, 20,* 680-693.

Riefer, D.M. & LaMay, M.L. (1998). Memory for common and bizarre stimuli: A storage-retrieval analysis. *Psychonomic Bulletin & Review, 5,* 312-317.

Riefer, D.M. & Rouder, J.N. (1992). A multinomial modeling analysis of the mnemonic benefits of bizarre imagery. *Memory & Cognition, 20,* 601-611.

Rieskamp, J. & Hoffrage, U. (1999). When do people use simple heuristics and how can we tell? In G. Gigerenzer & P. Todd (Eds.), *Simple heuristics that make us smart* (S. 141-167). New York: Oxford University Press.

Rietz, C., Rudinger, G. & Andres, J. (1996). Lineare Strukturgleichungsmodelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 253-268). Weinheim: Psychologie Verlags Union.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90,* 38-71.

Roediger, H.L. & McDermott, K.B. (1994). The problem of differing false-alarm rates for the process dissociation procedure: Comment on Verfaellie and Treadwell (1993). *Neuropsychology, 8,* 284-288.

Roediger, H.L.I., Jacoby, D. & McDermott, K.B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language, 35,* 300-318.

Romano, J.P. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association, 83,* 698-708.

Romney, A.K., Batchelder, W.H. & Weller, S.C. (1987). Recent applications of cultural consensus theory. *American Behavioral Scientist, 31,* 163-177.

Romney, A.K., Weller, S.C. & Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88,* 313-338.

Rose, R.G., Rose, P.R., King, N. & Perez, A. (1975). Bilingual memory for related and unrelated sentences. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 599-606.

Roskam, E.E. (1979). Eine Fallstudie über Forschungsmethodik. *Zeitschrift für Sozialpsychologie, 10,* 114-133.

Roskam, E.E. (1985). Current issues in item response theory. In E.E. Roskam (Ed.), *Measurement and personality assessment* (S. 3-19). Amsterdam: Elsevier.

Roskam, E.E. (1996a). Beobachtung und Daten. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 3-22). Weinheim: Psychologie Verlags Union.

Roskam, E.E. (1996b). Latent-Trait-Modelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 431-458). Weinheim: Beltz.

Ross, B.H. & Bower, G.H. (1981). Comparison of models of associative recall. *Memory & Cognition, 9,* 1-16.

Ross, S.M. (1997). *Introduction to probability models.* San Diego, CA: Academic Press.

Rost, J. (1985). A latent-class model for rating data. *Psychometrika, 50,* 37-49.

Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie.* Bern: Huber.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau, 50,* 140-156.

Rost, J. & Erdfelder, E. (1996). Mischverteilungsmodelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 333-348). Weinheim: Psychologie Verlags Union.

Roth, G. (1997). *Das Gehirn und seine Wirklichkeit. Kognitive Neurobiologie und ihre philosophischen Konsequenzen.* Frankfurt, Germany: Suhrkamp.

Rothkegel, R. (1998). *AppleTree (Version 1.2.1)* [Computer-Programm]. Trier: Fachbereich 1-Psychologie der Universität Trier.

Rothkegel, R. (1999). AppleTree: A multinomial processing tree modeling program for Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 31,* 696-700.

Rouder, J.N. & Batchelder, W.H. (1998). Multinomial models for measuring storage and retrieval processes in paired associate learning. In C.E. Dowling, F.S. Roberts & P. Theuns (Eds.), *Recent progress in mathematical psychology.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rubin, D.B. (1991). EM and beyond. *Psychometrika, 56,* 241-254.

Rudas, T. (1986). A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. *Journal of Statistical Computation and Simulation, 24*, 107-120.

Rudinger, G., Andres, J. & Rietz, C. (1994). Structural equation models for studying intellectual development. In D. Magnusson, L.R. Bergman, G. Rudinger & B. Törestad (Ed.), *Problems and methods in longitudinal research: Stability and change* (S. 274-307). Cambridge: Cambridge University Press.

Rudinger, G., Chaselon, F., Zimmermann, E. & Henning, H.J. (1985). *Qualitative Daten. Neue Wege sozialwissenschaftlicher Methodik.* München: Urban und Schwarzenberg.

Ruoß, M. (1997). Schmerzpatienten zeigen einen höheren Hindsight Bias. *Zeitschrift für Experimentelle Psychologie, 44*, 561-588.

Russo, J.E., Johnson, E.J. & Stephens, D.L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759-761.

Russo, R. & Andrade, J. (1995). The directed forgetting effect in word fragment completion: An application of the process dissociation procedure. *The Quarterly Journal of Experimental Psychology, 48A*, 405-423.

Russo, R., Cullis, A.M. & Parkin, A.J. (1998). Consequences of violating the assumption of independence in the process dissociation procedure: A word fragment completion study. *Memory & Cognition, 26*, 617-632.

Saegert, J., Hamayan, E. & Ahmar, H. (1975). Memory for language of input in polyglots. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 607-613.

Saur, B. (1998). *Hypnose und Gedächtnis. Experimentelle Überprüfung eines suggestiven Verfahrens zur Reduzierung des Hindsight Bias.* Unveröffentlichte Diplomarbeit, Psychologisches Institut, Universität Bonn, Bonn.

Schilken, E. (1998). *Speicherung und Abruf von verbalem Material bei Patienten mit Temporallappenepilepsie.* Unveröffentliche Diplomarbeit, Psychologisches Institut, Universität Bonn, Bonn.

Schmitter-Edgecombe, M. (1999). Effects of divided attention and time course on automatic and controlled components of memory in older adults. *Psychology and Aging, 14*, 331-345.

Scholz, O.B. (1997). Das Unbewußte als Informationsverarbeitungsprozeß und seine Bedeutung für die moderne klinische Psychologie. *Zeitschrift für Psychologie, 205*, 327-356.

Scholz, O.B., Ott, R. & Müller-Sinik, K. (1997). Beziehung zwischen Parametern der impliziten und expliziten Informationsverarbeitung und psychosomatischen Selbstberichtsmaßen. *Verhaltenstherapie, 7*, 217-225.

Schönemann, P.H. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. In I. Borg & P.P. Mohler (Eds.), *Trends and perspectives in empirical social research* (S. 149-160). Berlin: de Gruyter.

Schooler, J.W., Foster, R.A. & Loftus, E.F. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition, 16,* 243-251.

Schreiber, T.A. & Sergant, S.D. (1998). The role of commitment in producing misinformation effects in eyewitness memory. *Psychonomic Bulletin & Review, 5,* 443-448.

Schumacher-Bittner, S. (1993). *Fehlerhafte Augenzeugen-Aussagen aufgrund widersprüchlicher Informationen. Einflüsse auf den Verfälschungseffekt.* Bonn: Holos.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Schwarz, W. (1991). Rational analysis will not throw off the yoke of the precision-importance trade-off function. *Behavioral and Brain Sciences, 14,* 501-502.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory and Cognition, 21,* 168-175.

Shao, J. & Tu, D. (1995). *The jackknife and bootstrap.* New York: Springer.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika, 72,* 133-144.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49,* 263-269.

Siemer, M. (1997). What exactly is random about random effects? *Methods of Psychological Research - Online, 2,* 139-151.

Sievers, W. (1998). *Bootstrap-Inferenzverfahren.* Aachen: Shaker.

Simonoff, J.S. (1986). Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *Journal of the American Statistical Association, 81,* 1005-1011.

Sixtl, F. (1993). *Der Mythos des Mittelwertes. Neue Methodenlehre der Statistik.* München: Oldenbourg.

Slakter, M.J. (1968). Accuracy of an approximation to the power of the chi-square goodness-of-fit test with small but equal expected frequencies. *Journal of the American Statistical Association, 63,* 912-918.

Slotnick, S.D., Klein, S.A., Dodson, C.S. & Shimamura, A.P. (1999). *An analysis of multinomial and signal detection models of source memory.* Manuscript submitted for publication.

Smith, P.L. (1998). Attention and luminance detection: A quantitative analysis. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 105-133.

Snodgrass, J.G. & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34-50.

Sporer, S.L. (1993). Der Falschinformationseffekt: Verschmelzung, Koexistenz oder Abruffehler? In L. Montada (Ed.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992, Band 2.* (S. 345-356). Göttingen: Hogrefe.

Sporer, S.L. & aus dem Kahmen, A. (1993). Zur suggestiven Wirkung irreführender nachträglicher Informationen. *Hypnose und Kognition, 10,* 1-19.

Stahlberg, D. & Eller, F. (1993). Hindsight-Effekte: Eine urteilstheoretische Erklärung. In L. Montada (Ed.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992, Band 2* (S. 735-741). Göttingen, Germany: Hogrefe.

Stahlberg, D. & Maass, A. (1998). Hindsight Bias: Impaired memory or biased reconstruction? In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology, Vol. 8* (S. 105-132). John Wiley & Sons.

Steffens, M.C., Buchner, A., Martensen, H. & Erdfelder, E. (in press). Further evidence on the similarity of memory processes in the process dissociation procedure and in source monitoring. *Memory & Cognition*.

Stegmüller, W. (1973a). *Logische Analyse der Struktur ausgereifter physikalischer Theorien: ‚Non-statement view' von Theorien (Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band II, Studienausgabe, Teil D).* Berlin: Springer.

Stegmüller, W. (1973b). *Statistisches Schließen - Statistische Begründungen - Statistische Analyse (Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Band II, 1. Halbband).* Berlin: Springer.

Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. In D. Scarborough & S. Sternberg (Eds.), *Methods, models, and conceptual issues. An invitation to cognitive science, Volume 4* (S. 703-863). Cambridge, MA: MIT Press.

Stevens, J.P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin, 88,* 728-737.

Stevens, S.S. (1935). The operational definition of psychological concepts. *Psychological Review, 42,* 517-526.

Steyer, R. & Eid, M. (1993). *Messen und Testen*. Berlin: Springer.

Stolz, J.A. & Merikle, P.A. (1995). *Conscious and unconscious influences of memory: Temporal dynamics*. Vortrag gehalten auf der 36th annual meeting of the Psychonomic Society, Los Angeles, November 1995.

Strack, F. & Bless, H. (1994). Memory for non-occurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33,* 203-217.

Strack, F. & Förster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science, 6,* 352-358.

Stumpf, H. (1996). Klassische Testheorie. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 411-430). Weinheim: Psychologie Verlags Union.

Suppes, P. (1983). Warum Formalisierung in der Wissenschaft erwünscht ist. In W. Balzer & M. Heidelberger (Eds.), *Zur Logik empirischer Theorien* (S. 24-39). Berlin: de Gruyter.

Suppes, P., Krantz, D.H., Luce, R.D. & Tversky, A. (1989). *Foundations of measurement. Volume II.* New York: Academic Press.

Suppes, P. & Zinnes, J. (1963). Basic measurement theory. In R.D. Luce, R.R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology, Volume 1* (S. 1-76). New York: John Wiley & Sons.

Sutton, H.E. & Wagner, R.P. (1985). *Genetics, a human concern*. New York: McMillan.

Swets, J.A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100-117.

Swets, J.A. (1998). Separating discrimination and decision in detection, recognition, and matters of life and death. In D. Scarborough & S. Sternberg (Eds.), *Methods, models, and conceptual issues. An invitation to cognitive science, Volume 4* (2 ed.) (S. 635-702). Cambrige, MA: MIT Press.

Tack, W.H. (1971). Das Entscheidbarkeitsproblem bei der Verwendung formaler Modelle. *Psychologische Beiträge, 13,* 355-365.

Tack, W.H. (1976). *Stochastische Lernmodelle*. Stuttgart: Kohlhammer.

Taneichi, N. & Sekiya, Y. (1995). Power approximations of the test of homogeneity for multinomial populations. *Journal of the Japan Statistical Society, 25,* 97-109.

Tanner, J.M. (1970). Physical growth. In P.H. Mussen (Ed.), *Carmichael´s manual of child psychology, Volume 1* (3. ed.) (S. 77-155). New York: Wiley.

Tanner, W.P.J. & Swets, J.A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401-409.

Thomas, H. (1995). Modeling class inclusion strategies. *Developmental Psychology, 31,* 170-179.

Thornhill, N.W. (1990). The evolutionary significance of incest rules. *Ethology and Sociobiology, 11,* 113-129.

Thornhill, N.W. (1991). An evolutionary analysis of rules regulating human inbreeding and marriage. *Behavioral and Brain Sciences, 14,* 247-293.

Thornhill, N.W. & Thornhill, R. (1987). Evolutionary theory and rules of mating and marriage pertaining to relatives. In C. Crawford, M. Smith & D. Krebs

(Eds.), *Sociobiology and psychology; Ideas, issues, and applications* (S. 373-400). Hillsdale: Erlbaum.

Tooby, J. & Cosmides, L. (1992). The psychological foundations of culture. In J.H. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind. Evolutionary psychology and the generation of culture* (S. 19-136). New York: Oxford University Press.

Townsend, J.T. & Ashby, F.G. (1983). *Stochastic modeling of elementary psychological processes.* Cambridge: Cambridge University Press.

Treisman, A. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*, 107-141.

Tukey, J.W. (1969). Analyzing data: Sanctification or detection work? *American Psychologist*, *24*, 83-91.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1-12.

Tulving, E. & Madigan, S.A. (1970). Memory and verbal learning. *Annual Review of Psychology*, *21*, 437-484.

Tulving, E. & Psotka, J. (1971). Retroactive inhibition in free recall: Inaccessibility of information available in the memory store. *Journal of Experimental Psychology*, *87*, 1-8.

Tulving, E. & Wiseman, S. (1975). Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society*, *6*, 79-82.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315.

Tversky, A. & Koehler, D.J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547-567.

van de Pol, F. & Langeheine, R. (1990). Mixed Markov latent class models. In C.C. Clogg (Ed.), *Sociological methodology 1990* (S. 213-247). Oxford: Blackwell.

van de Pol, F., Langeheine, R. & de Jong, W. (1991). *PANMARK user manual: PANel anlysis using MARKov chains* [Computer-Programm]. Voorburg: Netherlands Central Bureau of Statistics.

Vaterrodt-Plünnecke, B. (1994). Multinomiale Modellierung impliziter Gedächtnisprozesse: Ein alternativer Ansatz. *Zeitschrift für experimentelle und angewandte Psychologie*, *41*, 295-314.

Vaterrodt-Plünnecke, B., Krüger, T., Gerdes, H. & Bredenkamp, J. (1996). Prozeß-Dissoziations-Prozedur: Prüfbare Meßmodelle zur Erfassung von kontrollierten, automatischen und Antworttendenz-Prozessen. *Zeitschrift für Experimentelle Psychologie*, *43*, 483-519.

Verfaellie, M. & Treadwell, J.R. (1993). Status of recognition memory in amnesia. *Neuropsychology*, *7*, 5-13.

Verplanck, W.S. (1962). Unaware of where´s awareness. Some verbal operants-notates, monents, and notants. In C.W. Eriksen (Ed.), *Behavior and awareness-A symposium of research and interpretation* (S. 130-158). Durham, NC: Duke University Press.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data — results of a Monte Carlo study. *Methods of Psychological Research, 2*, 29-48.

von Eye, A. (1990). *Introduction to configural frequency analysis. The search for types and antitypes in cross-classifications.* Cambridge, MA: Cambridge University Press.

Vorberg, D. (1977). *On the equivalence of serial and parallel processing systems.* Vortrag gehalten auf der 10th Annual Mathematical Psychology Meeting, University of California at Los Angeles, CA, 1977.

Vorberg, D. & Schmidt, R. (1975). A testable property of some discrete-state models for recognition memory. *Psychological Review, 82*, 316-324.

Wagenaar, W.A. & Boer, J.P. (1987). Misleading postevent information: Testing parameterized models of integration in memory. *Acta Psychologica, 66*, 291-306.

Wainwright, M.J. & Reingold, E.M. (1996). Response-bias correction in the process dissociation procedure: Approaches, assumptions, and Evaluation. *Consciousness and Cognition, 5*, 232-254.

Wald, A. (1950). Note on identification of economic relations. In T.C. Koopmans (Ed.), *Statistical inference in dynamic economoc models* (S. 238-244). New York: Wiley.

Wallsten, T.S., Bender, R.H. & Li, Y. (1999). Dissociating judgment from response processes in statement verification: The effects of experience on each component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 96-115.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63-69.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Watson, J.B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158-177.

Waugh, N.C. & Norman, D.A. (1965). Primary memory. *Psychological Review, 72*, 89-104.

Wegener, I. (1999). *Soziale Kategorisierung im situativen Kontext: Kognitive Flexibilität bei der Personenwahrnehumung.* Dissertation, Rheinische Friedrich-Wilhelms-Universität, Bonn.

Weir, B.S. (1990). *Genetic data analysis.* Sunderland, MA: Sinauer.

Welham, C.V.J. (1990). Incest. An evolutionary model. *Ethology and Sociobiology, 11,* 97-111.

Westermann, R. (1987a). *Strukturalistische Theorienkonzeption und empirische Forschung in der Psychologie. Eine Fallstudie.* Berlin: Springer.

Westermann, R. (1987b). Wissenschaftstheoretische Grundlagen der experimentellen Psychologie. In G. Lüer (Ed.), *Allgemeine Experimentelle Psychologie* (S. 5-42). Stuttgart: Gustav Fischer Verlag.

Westermann, R. & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11,* 117-146.

Westmeyer, H. (1972). *Logik der Diagnostik. Grundlagen einer normativen Diagnostik.* Stuttgart: Kohlhammer.

Westmeyer, H. (Ed.). (1989). *Psychological theories from a structuralist point of view.* New York: Springer.

Westmeyer, H. (1992). *The structuralist program in psychology: Foundations and applications.* Toronto: Hogrefe & Huber Publishers.

Westmeyer, H. (1998). Psychologische Methoden und soziale Konventionen. In K.C. Klauer & H. Westmeyer (Eds.), *Psychologische Methoden und soziale Prozesse* (S. 250-266). Lengerich: Pabst.

White, H. (1982). Maximum likelihood estimation for misspecified models. *Econometrica, 50,* 1-25.

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Wickens, T.D. (1982). *Models for behavior. Stochastic processes in psychology.* San Francisco: Freeman.

Wickens, T.D. (1993). Analysis of contingency tables with between-subjects variability. *Psychological Bulletin, 113,* 191-204.

Wickens, T.D. (1998a). Categorical data analysis. *Annual Review of Psychology, 48,* 537-558.

Wickens, T.D. (1998b). Drawing conclusions from data: Statistical methods for coping with uncertainty. In D. Scarborough & S. Sternberg (Eds.), *Methods, models, and conceptual issues. An invitation to cognitive science, Volume 4* (2. ed.) (S. 584-634). Cambrige, MA: MIT Press.

Wiggins, L.M. (1973). *Panel analysis.* Amsterdam: Elsevier.

Wilkinson, A.C., DeMarinis, M. & Riley, S.J. (1983). Developmental and individual differences in rapid remembering. *Child Development, 54,* 898-911.

Wilkinson, A.C. & Koestler, R. (1983). Repeated recall: A new model and tests of its generality from childhood to old age. *Journal of Experimental Psychology: General, 112,* 423-451.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics, 9,* 60-62.

Willmes, K. (1998). Neyman-Pearson-Theorie statistischen Testens. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Eds.), *Handbuch Quantitative Methoden* (S. 109-122). Weinheim: Beltz.

Wilson, A. (1975). The inference of covert hypotheses by verbal reports in concept-learning research. *Quarterly Journal of Experimental Psychology, 27,* 313-322.

Wilson, M.I. & Daly, M. (1996). Male sexual proprietariness and violence against wives. *Current Directions in Psychological Science, 5,* 2-7.

Windmann, S. & Krüger, T. (1998). Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and Cognition, 7,* 603-633.

Windschitl, P.D. (1996). Memory for faces: Evidence for retrieval-based impairment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1101-1122.

Wippich, W. & Bredenkamp, J. (1979). *Bildhaftigkeit und Lernen.* Darmstadt: Steinkopff.

Wolfram, S. (1992). *Mathematica. Ein System für Mathematik auf dem Computer.* Bonn: Addison-Wesley Publishing Company.

Woodworth, R.S. (1938). *Experimental psychology.* New York: Holt.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11,* 95-103.

Wundt, W. (1907). Über Ausfrageexperimente und über die Methoden zur Psychologie des Denkens. *Psychologische Studien, 3,* 301-360.

Wundt, W. (1908). Kritische Nachlese zur Ausfragemethode. *Archiv für die gesamte Psychologie, 11,* 445-459.

Yantis, S., Meyer, D.E. & Smith, J.E.K. (1991). Analysis of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin, 110,* 350-374.

Yarnold, J.K. (1972). Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. *Annals of Mathematical Statistics, 43,* 1566-1580.

Yellott, J.I. (1971). Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology, 8,* 159-199.

Yonelinas, A.P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341-1354.

Yonelinas, A.P. (1999). Recognition memory ROCs and the dual-process signal-detection model: Comment on Glanzer, Kim, Hilford, and Adams (1999). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 514-521.

Yonelinas, A.P. & Jacoby, L.L. (1996a). Response bias and the process dissociation procedure. *Journal of Experimental Psychology: General, 125,* 422-434.

Yonelinas, A.P. & Jacoby, L.L. (1996b). „Response bias and the process-dissociation procedure": Correction to Yonelinas and Jacoby (1996). *Journal of Experimental Psychology: General, 126*, 18.

Yonelinas, A.P., Kroll, N.E.A., Dobbins, J., Lazzara, M. & Knight, R.T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember-know, process dissociation, and receiver operating characteristics data. *Neuropsychology, 12*, 323-339.

Yonelinas, A.P., Regehr, G. & Jacoby, L.L. (1995). Incorporating response bias in a dual-process theory of memory. *Journal of Memory and Language, 34*, 821-835.

Zaragoza, M.S. & Koshmider, J.W., III. (1989). Misled subjects may know more than their performance implies. *Journal of Experimental Psychology: Learning,Memory,and Cognition, 15*, 246-255.

Zaragoza, M.S. & Lane, S.M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 934-945.

Zaragoza, M.S. & Mitchell, K.J. (1996). Repeated exposure to suggestion and the creation of false memories. *Psychological Science, 7*, 294-300.

Zaragoza, M.S. & Muench, J. (1989). *Source confusion following exposure to misleading postevent information*. Vortrag gehalten auf der 30. Annual Meeting of the Psychonomic Society, Atlanta, Georgia, November 1989.

Zickar, M.J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science, 7*, 104-115.

Anhang

Dieser Anhang führt Modellbeschreibungsdateien (EQN-Dateien) für die in der Arbeit behandelten MVB-Modelle auf. Zu einigen dieser EQN-Dateien werden auch passende Dateneingabedateien (MDT-Dateien) mit den in der vorliegenden Arbeit analysierten Datensätzen angegeben.

Mit Hilfe der EQN-Dateien können die behandelten Modelle unter Verwendung des Windows-Programms „GPT" (General Processing Tree, Hu, 1998; Hu & Phillips, 1999), des DOS-Programms „MBT" (Multinomial Binary Tree, Hu, 1995; 1999b) oder des Macintosh-Programms „AppleTree" von Rothkegel (1998; 1999) sehr einfach auf beliebige Datensätze angewandt werden, die in MDT-Dateien zu spezifizieren sind. Die jeweils aktuellsten Versionen von GPT und MBT sowie Beispiele für syntaktisch einwandfreie EQN- und MDT-Dateien können unter der Adresse http://xhuoffice.psyc.memphis.edu/gpt/index.htm z.Zt. kostenfrei über das Internet bezogen werden. Eine WorldWideWeb- (WWW-) Hilfe zum GPT-Programm, das wegen seines großen Leistungsumfangs nicht ganz leicht zu bedienen ist, wird unter derselben Adresse angeboten. Eine Kurzbeschreibung von MBT findet man bei Hu (1999b), eine knappe Erläuterung von GPT bei Hu und Phillips (1999). Die jeweils neueste Version von AppleTree kann einschließlich EQN- und MDT-Beispieldateien z.Zt. ebenfalls kostenlos über das Internet bezogen werden (http://www.psychologie.uni-trier.de:8000/projects/AppleTree.html). Eine Kurzbeschreibung bietet Rothkegel (1999).

Die zu analysierenden MVB-Modelle müssen den Programmen zunächst in geeigneter Form eingegeben werden. Dies geschieht mittels spezieller ASCII-Textdateien, den sogenannten EQN-Dateien.[107] EQN-Dateien werden hier so beschrieben, daß sie von allen drei o.g. Programmen lesbar sind. Programmspezifische Besonderheiten, die teilweise eine Liberalisierung der Syntax von EQN-Dateien ermöglichen, können den oben zitierten Quellen entnommen werden. EQN-Dateien kann ein beliebiger, maximal achtstelliger Name gegeben werden, dem die Extension „.EQN" folgen muß. Die erste Zeile einer EQN-Datei gibt jeweils die Anzahl der folgenden Zeilen in der Datei an, die zugleich der Gesamtzahl der Äste des beschriebenen MVB-Modells entspricht. Danach folgen Beschreibungen aller Äste des Modells derart, daß — jeweils durch mindestens eine Leerstelle getrennt — (a) zunächst die Nummer des Baumes, zu dem der betreffende Ast gehört, (b) die Nummer der empirischen Kategorie, in die der Ast einmündet, und (c) die Wahrscheinlichkeit des Astes als Produkt von Parametern θ_s oder ihren Komplementen $(1-\theta_s)$ angegeben wird.[108]

---

[107] In GPT ist alternativ auch eine graphische Eingabe des Modells möglich.

[108] Letzteres gilt für AppleTree und MBT. GPT kann auch nichtbinäre MVB-Modelle direkt analysieren, d.h. ohne den Umweg über eine binäre Reparametrisierung (vgl. Abschnitt 3.5).

Die Bezeichnung der Parameter kann frei gewählt werden, darf aber nicht mehr als vier Zeichen umfassen. Zwischen den Parametern oder ihren Komplementen ist jeweils das Produktzeichen „*" *ohne Leerstellen* anzugeben. Komplemente zu Parametern müssen in der Form (1−θ), d.h. in Klammern und ohne Leerstellen angegeben werden.

Die Numerierung der Bäume und Kategorien muß durchlaufend erfolgen, d.h. wenn eine beliebige natürliche Zahl *n* als Nummer irgendwo auftaucht (*n* > 1), dann muß auch *n*−1 auftauchen. Kategorien aus verschiedenen Bäumen (bzw. Versuchsbedingungen) müssen immer verschiedene Nummern bekommen, auch wenn die Kategorien inhaltlich das Gleiche bedeuten.

Der „Größe" der verarbeitbaren Modelle (Maximale Anzahl Bäume, Kategorien, Parameter, Äste, Äste pro Kategorie) variiert zwischen den angegebenen Programmen und ist bei GPT und AppleTree auch maschinenabhängig. Bei MBT wird die maximale Größe des analysierbaren Modells jeweils nach dem Starten des Programms angegeben.

Die MDT-Dateien können ebenfalls einen maximal achtstelligen Namen bekommen und müssen mit der Extension „.MDT" abgeschlossen werden. Sie müssen eine kurze, frei wählbare Beschreibung des Datensatzes in der ersten Zeile enthalten. Danach folgen die zu analysierenden Kategoriehäufigkeiten, und zwar derart, daß in jeder Zeile zunächst die Nummer der empirischen Kategorie, danach die in der Stichprobe beobachtete Häufigkeit der betreffenden Kategorie angegeben wird, und zwar durch mindestens eine Leerstelle getrennt. Da die Kategorien aus verschiedenen Bäumen immer unterschiedliche Nummern tragen, erübrigt sich eine Angabe der Baumnummer. Von zentraler Bedeutung ist jedoch, daß die Numerierung der empirischen Kategorien in der MDT-Datei und der verwendeten EQN-Datei identisch ist.

1) *Datei „EA.EQN" zum Enkodierung-Abruf-Modell von Batchelder and Riefer (1980; 1986).*

Das in Kapitel 3.5.1 beschriebene Enkodierung-Abruf-Modell umfaßt zwei Ereignisbäume (für Itempaare und Einzelitems) mit acht Ästen für vier plus zwei Ereigniskategorien, die hier in der gleichen Weise wie in Abschnitt 3.5.1 durchnumeriert werden. Das Modell ist in der hier vorgestellten Version saturiert, wird aber durch die Restriktion $u = a$ testbar.

Die EQN-Datei ist folgendermaßen aufgebaut:

```
8
1       1       c*r
1       4       c*(1-r)
1       2       (1-c)*u*u
1       3       (1-c)*u*(1-u)
1       3       (1-c)*(1-u)*u
1       4       (1-c)*(1-u)*(1-u)
2       5       a
2       6       (1-a)
```

2) *Dateien „1HT.EQN" und „2HT.EQN"zum Ein- und Zwei-Hochschwellen-Modell für Rekognitions- und Signalentdeckungstests*

Die in Abschnitt 3.5.2 vorgestellten Ein- und Zwei-Hochschwellen-Modelle umfassen jeweils zwei Ereignisbäume (für alte und neue Items) für insgesamt vier Ereigniskategorien (Treffer, Verpasser sowie falsche Alarme und korrekte Zurückweisungen). Die Modelle sind saturiert.

Die EQN-Datei zum Ein-Hochschwellen-Modell lautet:

```
5
1       1       r1ht
1       1       (1-r1ht)*b1ht
1       2       (1-r1ht)*(1-b1ht)
2       3       b1ht
2       4       (1-b1ht)
```

Die analoge Datei zum Zwei-Hochschwellen-Modell ist:

```
6
 1       1       r2ht
 1       1       (1-r2ht)*b2ht
 1       2       (1-r2ht)*(1-b2ht)
 2       4       r2ht
 2       3       (1-r2ht)*b2ht
 2       4       (1-r2ht)*(1-b2ht)
```

3) *Datei „LCAEA.EQN" zum Zwei-Klassen-Enkodierung-Abruf-Modell aus Abschnitt 5.4.2.1*

Das in Abschnitt 5.4.2.1 vorgestellte Enkodierung-Abruf-Modell (EA-Modell) mit zwei latenten Klassen ist ein Modell mit insgesamt 16 Ereignisbäumen, zwei für jede von acht Versuchsgruppen. Die Parameter des Modells sind so restringiert wie es für den Versuchsplan aus Tabelle 5.1 sinnvoll ist.

Die EQN-Datei zum Modell lautet:

```
128
 1        1       p*c10*r10
 1        4       p*c10*(1-r10)
 1        2       p*(1-c10)*u10*u10
 1        3       p*(1-c10)*u10*(1-u10)
 1        3       p*(1-c10)*(1-u10)*u10
 1        4       p*(1-c10)*(1-u10)*(1-u10)
 1        1       (1-p)*c20*r20
 1        4       (1-p)*c20*(1-r20)
 1        2       (1-p)*(1-c20)*u20*u20
 1        3       (1-p)*(1-c20)*u20*(1-u20)
 1        3       (1-p)*(1-c20)*(1-u20)*u20
 1        4       (1-p)*(1-c20)*(1-u20)*(1-u20)
 2        5       p*u10
 2        6       p*(1-u10)
 2        5       (1-p)*u20
 2        6       (1-p)*(1-u20)
 3        7       p*c10*r10
 3       10       p*c10*(1-r10)
 3        8       p*(1-c10)*u11*u11
 3        9       p*(1-c10)*u11*(1-u11)
 3        9       p*(1-c10)*(1-u11)*u11
 3       10       p*(1-c10)*(1-u11)*(1-u11)
 3        7       (1-p)*c20*r20
 3       10       (1-p)*c20*(1-r20)
 3        8       (1-p)*(1-c20)*u21*u21
 3        9       (1-p)*(1-c20)*u21*(1-u21)
 3        9       (1-p)*(1-c20)*(1-u21)*u21
 3       10       (1-p)*(1-c20)*(1-u21)*(1-u21)
 4       11       p*u11
 4       12       p*(1-u11)
 4       11       (1-p)*u21
 4       12       (1-p)*(1-u21)
 5       13       p*c10*r11
 5       16       p*c10*(1-r11)
 5       14       p*(1-c10)*u10*u10
 5       15       p*(1-c10)*u10*(1-u10)
 5       15       p*(1-c10)*(1-u10)*u10
 5       16       p*(1-c10)*(1-u10)*(1-u10)
 5       13       (1-p)*c20*r21
 5       16       (1-p)*c20*(1-r21)
 5       14       (1-p)*(1-c20)*u20*u20
 5       15       (1-p)*(1-c20)*u20*(1-u20)
 5       15       (1-p)*(1-c20)*(1-u20)*u20
 5       16       (1-p)*(1-c20)*(1-u20)*(1-u20)
 6       17       p*u10
 6       18       p*(1-u10)
 6       17       (1-p)*u20
 6       18       (1-p)*(1-u20)
 7       19       p*c10*r11
 7       22       p*c10*(1-r11)
 7       20       p*(1-c10)*u11*u11
 7       21       p*(1-c10)*u11*(1-u11)
 7       21       p*(1-c10)*(1-u11)*u11
 7       22       p*(1-c10)*(1-u11)*(1-u11)
 7       19       (1-p)*c20*r21
 7       22       (1-p)*c20*(1-r21)
 7       20       (1-p)*(1-c20)*u21*u21
 7       21       (1-p)*(1-c20)*u21*(1-u21)
 7       21       (1-p)*(1-c20)*(1-u21)*u21
 7       22       (1-p)*(1-c20)*(1-u21)*(1-u21)
 8       23       p*u11
 8       24       p*(1-u11)
 8       23       (1-p)*u21
 8       24       (1-p)*(1-u21)
 9       25       p*c11*r10
 9       28       p*c11*(1-r10)
 9       26       p*(1-c11)*u10*u10
 9       27       p*(1-c11)*u10*(1-u10)
```

```
 9       27       p*(1-c11)*(1-u10)*u10
 9       28       p*(1-c11)*(1-u10)*(1-u10)
 9       25       (1-p)*c21*r20
 9       28       (1-p)*c21*(1-r20)
 9       26       (1-p)*(1-c21)*u20*u20
 9       27       (1-p)*(1-c21)*u20*(1-u20)
 9       27       (1-p)*(1-c21)*(1-u20)*u20
 9       28       (1-p)*(1-c21)*(1-u20)*(1-u20)
10       29       p*u10
10       30       p*(1-u10)
10       29       (1-p)*u20
10       30       (1-p)*(1-u20)
11       31       p*c11*r10
11       34       p*c11*(1-r10)
11       32       p*(1-c11)*u11*u11
11       33       p*(1-c11)*u11*(1-u11)
11       33       p*(1-c11)*(1-u11)*u11
11       34       p*(1-c11)*(1-u11)*(1-u11)
11       31       (1-p)*c21*r20
11       34       (1-p)*c21*(1-r20)
11       32       (1-p)*(1-c21)*u21*u21
11       33       (1-p)*(1-c21)*u21*(1-u21)
11       33       (1-p)*(1-c21)*(1-u21)*u21
11       34       (1-p)*(1-c21)*(1-u21)*(1-u21)
12       35       p*u11
12       36       p*(1-u11)
12       35       (1-p)*u21
12       36       (1-p)*(1-u21)
13       37       p*c11*r11
13       40       p*c11*(1-r11)
13       38       p*(1-c11)*u10*u10
13       39       p*(1-c11)*u10*(1-u10)
13       39       p*(1-c11)*(1-u10)*u10
13       40       p*(1-c11)*(1-u10)*(1-u10)
13       37       (1-p)*c21*r21
13       40       (1-p)*c21*(1-r21)
13       38       (1-p)*(1-c21)*u20*u20
13       39       (1-p)*(1-c21)*u20*(1-u20)
13       39       (1-p)*(1-c21)*(1-u20)*u20
13       40       (1-p)*(1-c21)*(1-u20)*(1-u20)
14       41       p*u10
14       42       p*(1-u10)
14       41       (1-p)*u20
14       42       (1-p)*(1-u20)
15       43       p*c11*r11
15       46       p*c11*(1-r11)
15       44       p*(1-c11)*u11*u11
15       45       p*(1-c11)*u11*(1-u11)
15       45       p*(1-c11)*(1-u11)*u11
15       46       p*(1-c11)*(1-u11)*(1-u11)
15       43       (1-p)*c21*r21
15       46       (1-p)*c21*(1-r21)
15       44       (1-p)*(1-c21)*u21*u21
15       45       (1-p)*(1-c21)*u21*(1-u21)
15       45       (1-p)*(1-c21)*(1-u21)*u21
15       46       (1-p)*(1-c21)*(1-u21)*(1-u21)
16       47       p*u11
16       48       p*(1-u11)
16       47       (1-p)*u21
16       48       (1-p)*(1-u21)
```

Ein hierzu nahezu ideal passender Datensatz ist:

```
Ideale Daten zu LCAEA.EQN
1   85
2   44
3   158
4   513
5   208
6   592
```

```
7   85
8   172
9   259
10  285
11  448
12  352
13  147
14  44
15  158
16  451
17  208
18  592
19  147
20  172
21  259
22  228
23  448
24  352
25  147
26  19
27  84
28  550
29  208
30  592
31  147
32  87
33  159
34  407
35  448
36  352
37  282
38  19
39  84
40  416
41  208
42  592
43  282
44  87
45  159
46  273
47  448
48  352
===
```

4) *Datei „ABOMODEL.EQN" zum ABO-Blutgruppen-Vererbungsmodell aus Abschnitt 6.2 sowie die Blutgruppendatei „ABODATA.MDT".*

Das Blutgruppen-Vererbungsmodell aus Abschnitt 6.2 welches die Möglichkeit berücksichtigt, daß der Präsumtivvater mit Wahrscheinlichkeit 1–*p* nicht der biologische Vater ist, kann mit der folgenden EQN-Datei analysiert werden:

```
324
1       22      a*a*a*a*p
1       22      a*a*a*a*(1-p)*a
1       30      a*a*a*a*(1-p)*(1-a)*x
1       22      a*a*a*a*(1-p)*(1-a)*(1-x)
1       24      a*a*a*(1-a)*x*p
1       24      a*a*a*(1-a)*x*(1-p)*a
1       32      a*a*a*(1-a)*x*(1-p)*(1-a)*x
1       24      a*a*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       22      a*a*a*(1-a)*(1-x)*p
1       22      a*a*a*(1-a)*(1-x)*(1-p)*a
1       30      a*a*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       22      a*a*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       32      a*a*(1-a)*x*a*p
1       24      a*a*(1-a)*x*a*(1-p)*a
```

```
1       32      a*a*(1-a)*x*a*(1-p)*(1-a)*x
1       24      a*a*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       31      a*a*(1-a)*x*(1-a)*x*p
1       23      a*a*(1-a)*x*(1-a)*x*(1-p)*a
1       31      a*a*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       23      a*a*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       31      a*a*(1-a)*x*(1-a)*(1-x)*p
1       23      a*a*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       31      a*a*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       23      a*a*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       22      a*a*(1-a)*(1-x)*a*p
1       22      a*a*(1-a)*(1-x)*a*(1-p)*a
1       30      a*a*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       22      a*a*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       23      a*a*(1-a)*(1-x)*(1-a)*x*p
1       23      a*a*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       31      a*a*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       23      a*a*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       21      a*a*(1-a)*(1-x)*(1-a)*(1-x)*p
1       21      a*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       29      a*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       21      a*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       54      a*(1-a)*x*a*a*p
1       54      a*(1-a)*x*a*a*(1-p)*a
1       62      a*(1-a)*x*a*a*(1-p)*(1-a)*x
1       54      a*(1-a)*x*a*a*(1-p)*(1-a)*(1-x)
1       56      a*(1-a)*x*a*(1-a)*x*p
1       56      a*(1-a)*x*a*(1-a)*x*(1-p)*a
1       64      a*(1-a)*x*a*(1-a)*x*(1-p)*(1-a)*x
1       56      a*(1-a)*x*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       54      a*(1-a)*x*a*(1-a)*(1-x)*p
1       54      a*(1-a)*x*a*(1-a)*(1-x)*(1-p)*a
1       62      a*(1-a)*x*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       54      a*(1-a)*x*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       64      a*(1-a)*x*(1-a)*x*a*p
1       56      a*(1-a)*x*(1-a)*x*a*(1-p)*a
1       64      a*(1-a)*x*(1-a)*x*a*(1-p)*(1-a)*x
1       56      a*(1-a)*x*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       63      a*(1-a)*x*(1-a)*x*(1-a)*x*p
1       55      a*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*a
1       63      a*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       55      a*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       63      a*(1-a)*x*(1-a)*x*(1-a)*(1-x)*p
1       55      a*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       63      a*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       55       a*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       54      a*(1-a)*x*(1-a)*(1-x)*a*p
1       54      a*(1-a)*x*(1-a)*(1-x)*a*(1-p)*a
1       62      a*(1-a)*x*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       54      a*(1-a)*x*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       55      a*(1-a)*x*(1-a)*(1-x)*(1-a)*x*p
1       55      a*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       63      a*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       55      a*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       53      a*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*p
1       53      a*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       61      a*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       53      a*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       22      a*(1-a)*(1-x)*a*a*p
1       22      a*(1-a)*(1-x)*a*a*(1-p)*a
1       30      a*(1-a)*(1-x)*a*a*(1-p)*(1-a)*x
1       22      a*(1-a)*(1-x)*a*a*(1-p)*(1-a)*(1-x)
1       24      a*(1-a)*(1-x)*a*(1-a)*x*p
1       24      a*(1-a)*(1-x)*a*(1-a)*x*(1-p)*a
1       32      a*(1-a)*(1-x)*a*(1-a)*x*(1-p)*(1-a)*x
1       24      a*(1-a)*(1-x)*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       22      a*(1-a)*(1-x)*a*(1-a)*(1-x)*p
1       22      a*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*a
1       30      a*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       22      a*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       32      a*(1-a)*(1-x)*(1-a)*x*a*p
1       24      a*(1-a)*(1-x)*(1-a)*x*a*(1-p)*a
```

```
1       32      a*(1-a)*(1-x)*(1-a)*x*a*(1-p)*(1-a)*x
1       24      a*(1-a)*(1-x)*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       31      a*(1-a)*(1-x)*(1-a)*x*(1-a)*x*p
1       23      a*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*a
1       31      a*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       23      a*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       31      a*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*p
1       23      a*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       31      a*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       23      a*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       22      a*(1-a)*(1-x)*(1-a)*(1-x)*a*p
1       22      a*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*a
1       30      a*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       22      a*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       23      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*p
1       23      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       31      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       23      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       21      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*p
1       21      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       29      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       21      a*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       62      (1-a)*x*a*a*a*p
1       62      (1-a)*x*a*a*a*(1-p)*a
1       58      (1-a)*x*a*a*a*(1-p)*(1-a)*x
1       58      (1-a)*x*a*a*a*(1-p)*(1-a)*(1-x)
1       64      (1-a)*x*a*a*(1-a)*x*p
1       64      (1-a)*x*a*a*(1-a)*x*(1-p)*a
1       60      (1-a)*x*a*a*(1-a)*x*(1-p)*(1-a)*x
1       60      (1-a)*x*a*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       62      (1-a)*x*a*a*(1-a)*(1-x)*p
1       62      (1-a)*x*a*a*(1-a)*(1-x)*(1-p)*a
1       58      (1-a)*x*a*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       58      (1-a)*x*a*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       60      (1-a)*x*a*(1-a)*x*a*p
1       64      (1-a)*x*a*(1-a)*x*a*(1-p)*a
1       60      (1-a)*x*a*(1-a)*x*a*(1-p)*(1-a)*x
1       60      (1-a)*x*a*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       59      (1-a)*x*a*(1-a)*x*(1-a)*x*p
1       63      (1-a)*x*a*(1-a)*x*(1-a)*x*(1-p)*a
1       59      (1-a)*x*a*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       59      (1-a)*x*a*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       59      (1-a)*x*a*(1-a)*x*(1-a)*(1-x)*p
1       63      (1-a)*x*a*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       59      (1-a)*x*a*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       59      (1-a)*x*a*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       58      (1-a)*x*a*(1-a)*(1-x)*a*p
1       62      (1-a)*x*a*(1-a)*(1-x)*a*(1-p)*a
1       58      (1-a)*x*a*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       58      (1-a)*x*a*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       59      (1-a)*x*a*(1-a)*(1-x)*(1-a)*x*p
1       63      (1-a)*x*a*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       59      (1-a)*x*a*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       59      (1-a)*x*a*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       57      (1-a)*x*a*(1-a)*(1-x)*(1-a)*(1-x)*p
1       61      (1-a)*x*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       57      (1-a)*x*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       57      (1-a)*x*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       46      (1-a)*x*(1-a)*x*a*a*p
1       46      (1-a)*x*(1-a)*x*a*a*(1-p)*a
1       42      (1-a)*x*(1-a)*x*a*a*(1-p)*(1-a)*x
1       42      (1-a)*x*(1-a)*x*a*a*(1-p)*(1-a)*(1-x)
1       48      (1-a)*x*(1-a)*x*a*(1-a)*x*p
1       48      (1-a)*x*(1-a)*x*a*(1-a)*x*(1-p)*a
1       44      (1-a)*x*(1-a)*x*a*(1-a)*x*(1-p)*(1-a)*x
1       44      (1-a)*x*(1-a)*x*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       46      (1-a)*x*(1-a)*x*a*(1-a)*(1-x)*p
1       46      (1-a)*x*(1-a)*x*a*(1-a)*(1-x)*(1-p)*a
1       42      (1-a)*x*(1-a)*x*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       42      (1-a)*x*(1-a)*x*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       44      (1-a)*x*(1-a)*x*(1-a)*x*a*p
1       48      (1-a)*x*(1-a)*x*(1-a)*x*a*(1-p)*a
```

```
1       44      (1-a)*x*(1-a)*x*(1-a)*x*a*(1-p)*(1-a)*x
1       44      (1-a)*x*(1-a)*x*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       43      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*x*p
1       47      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*a
1       43      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       43      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       43      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*(1-x)*p
1       47      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       43      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       43      (1-a)*x*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       42      (1-a)*x*(1-a)*x*(1-a)*(1-x)*a*p
1       46      (1-a)*x*(1-a)*x*(1-a)*(1-x)*a*(1-p)*a
1       42      (1-a)*x*(1-a)*x*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       42      (1-a)*x*(1-a)*x*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       43      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*x*p
1       47      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       43      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       43      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       41      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*p
1       45      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       41      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       41      (1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       46      (1-a)*x*(1-a)*(1-x)*a*a*p
1       46      (1-a)*x*(1-a)*(1-x)*a*a*(1-p)*a
1       42      (1-a)*x*(1-a)*(1-x)*a*a*(1-p)*(1-a)*x
1       42      (1-a)*x*(1-a)*(1-x)*a*a*(1-p)*(1-a)*(1-x)
1       48      (1-a)*x*(1-a)*(1-x)*a*(1-a)*x*p
1       48      (1-a)*x*(1-a)*(1-x)*a*(1-a)*x*(1-p)*a
1       44      (1-a)*x*(1-a)*(1-x)*a*(1-a)*x*(1-p)*(1-a)*x
1       44      (1-a)*x*(1-a)*(1-x)*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       46      (1-a)*x*(1-a)*(1-x)*a*(1-a)*(1-x)*p
1       46      (1-a)*x*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*a
1       42      (1-a)*x*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       42      (1-a)*x*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       44      (1-a)*x*(1-a)*(1-x)*(1-a)*x*a*p
1       48      (1-a)*x*(1-a)*(1-x)*(1-a)*x*a*(1-p)*a
1       44      (1-a)*x*(1-a)*(1-x)*(1-a)*x*a*(1-p)*(1-a)*x
1       44      (1-a)*x*(1-a)*(1-x)*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*x*p
1       47      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*a
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*p
1       47      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       42      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*a*p
1       46      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*a
1       42      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       42      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*p
1       47      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       43      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       41      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*p
1       45      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       41      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       41      (1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       22      (1-a)*(1-x)*a*a*a*p
1       22      (1-a)*(1-x)*a*a*a*(1-p)*a
1       26      (1-a)*(1-x)*a*a*a*(1-p)*(1-a)*x
1       18      (1-a)*(1-x)*a*a*a*(1-p)*(1-a)*(1-x)
1       24      (1-a)*(1-x)*a*a*(1-a)*x*p
1       24      (1-a)*(1-x)*a*a*(1-a)*x*(1-p)*a
1       28      (1-a)*(1-x)*a*a*(1-a)*x*(1-p)*(1-a)*x
1       20      (1-a)*(1-x)*a*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       22      (1-a)*(1-x)*a*a*(1-a)*(1-x)*p
1       22      (1-a)*(1-x)*a*a*(1-a)*(1-x)*(1-p)*a
1       26      (1-a)*(1-x)*a*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       18      (1-a)*(1-x)*a*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       28      (1-a)*(1-x)*a*(1-a)*x*a*p
1       24      (1-a)*(1-x)*a*(1-a)*x*a*(1-p)*a
```

```
1       28      (1-a)*(1-x)*a*(1-a)*x*a*(1-p)*(1-a)*x
1       20      (1-a)*(1-x)*a*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       27      (1-a)*(1-x)*a*(1-a)*x*(1-a)*x*p
1       23      (1-a)*(1-x)*a*(1-a)*x*(1-a)*x*(1-p)*a
1       27      (1-a)*(1-x)*a*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       19      (1-a)*(1-x)*a*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       27      (1-a)*(1-x)*a*(1-a)*x*(1-a)*(1-x)*p
1       23      (1-a)*(1-x)*a*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       27      (1-a)*(1-x)*a*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       19      (1-a)*(1-x)*a*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       18      (1-a)*(1-x)*a*(1-a)*(1-x)*a*p
1       22      (1-a)*(1-x)*a*(1-a)*(1-x)*a*(1-p)*a
1       26      (1-a)*(1-x)*a*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       18      (1-a)*(1-x)*a*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       19      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*x*p
1       23      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       27      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       19      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       17      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*(1-x)*p
1       21      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       25      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       17      (1-a)*(1-x)*a*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       38      (1-a)*(1-x)*(1-a)*x*a*a*p
1       38      (1-a)*(1-x)*(1-a)*x*a*a*(1-p)*a
1       42      (1-a)*(1-x)*(1-a)*x*a*a*(1-p)*(1-a)*x
1       34      (1-a)*(1-x)*(1-a)*x*a*a*(1-p)*(1-a)*(1-x)
1       40      (1-a)*(1-x)*(1-a)*x*a*(1-a)*x*p
1       40      (1-a)*(1-x)*(1-a)*x*a*(1-a)*x*(1-p)*a
1       44      (1-a)*(1-x)*(1-a)*x*a*(1-a)*x*(1-p)*(1-a)*x
1       36      (1-a)*(1-x)*(1-a)*x*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       38      (1-a)*(1-x)*(1-a)*x*a*(1-a)*(1-x)*p
1       38      (1-a)*(1-x)*(1-a)*x*a*(1-a)*(1-x)*(1-p)*a
1       42      (1-a)*(1-x)*(1-a)*x*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       34      (1-a)*(1-x)*(1-a)*x*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       44      (1-a)*(1-x)*(1-a)*x*(1-a)*x*a*p
1       40      (1-a)*(1-x)*(1-a)*x*(1-a)*x*a*(1-p)*a
1       44      (1-a)*(1-x)*(1-a)*x*(1-a)*x*a*(1-p)*(1-a)*x
1       36      (1-a)*(1-x)*(1-a)*x*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       43      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*x*p
1       39      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*a
1       43      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       35      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       43      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*(1-x)*p
1       39      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       43      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       35      (1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       34      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*a*p
1       38      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*a*(1-p)*a
1       42      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       34      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       35      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*x*p
1       39      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       43      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       35      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       33      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*p
1       37      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       41      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       33      (1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       6       (1-a)*(1-x)*(1-a)*(1-x)*a*a*p
1       6       (1-a)*(1-x)*(1-a)*(1-x)*a*a*(1-p)*a
1       10      (1-a)*(1-x)*(1-a)*(1-x)*a*a*(1-p)*(1-a)*x
1       2       (1-a)*(1-x)*(1-a)*(1-x)*a*a*(1-p)*(1-a)*(1-x)
1       8       (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*x*p
1       8       (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*x*(1-p)*a
1       12      (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*x*(1-p)*(1-a)*x
1       4       (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*x*(1-p)*(1-a)*(1-x)
1       6       (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*(1-x)*p
1       6       (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*a
1       10      (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*(1-a)*x
1       2       (1-a)*(1-x)*(1-a)*(1-x)*a*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       12      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*a*p
1       8       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*a*(1-p)*a
```

```
1       12      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*a*(1-p)*(1-a)*x
1       4       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*a*(1-p)*(1-a)*(1-x)
1       11      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*x*p
1       7       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*a
1       11      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*(1-a)*x
1       3       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*x*(1-p)*(1-a)*(1-x)
1       11      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*p
1       7       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*a
1       11      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*x
1       3       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
1       2       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*a*p
1       6       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*a
1       10      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*(1-a)*x
1       2       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*a*(1-p)*(1-a)*(1-x)
1       3       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*p
1       7       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*a
1       11      (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*x
1       3       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*x*(1-p)*(1-a)*(1-x)
1       1       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*p
1       5       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*a
1       9       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*x
1       1       (1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-a)*(1-x)*(1-p)*(1-a)*(1-x)
```

Neben dem schon erwähnten Parameter *p* für die Wahrscheinlichkeit, daß der Präsumtivvater auch der biologische Vater ist, tauchen in dieser Datei die Wahrscheinlichkeiten *a* für das A-Blutgruppenallel und die bedingte Wahrscheinlichkeit *x* für ein B-Blutgruppenallel, gegeben daß kein A-Allel vorliegt, auf. Die Wahrscheinlichkeit eines B-Blutgruppenallels ergibt sich einfach aus dem Produkt $(1-a)\cdot x$, die eines 0-Blutgruppenallels aus dem Produkt $(1-a)\cdot(1-x)$.

Da es insgesamt 4 Blutgruppenphänotypen gibt, sind $4\cdot4\cdot4 = 64$ Blutgruppenkombinationen von Mutter, Präsumtivvater und Kind a priori in Erwägung zu ziehen. Die Nummern der Beobachtungskategorien in der EQN-Datei ergeben sich durch Durchnumerierung aller 64 Kombinationen, wobei die Blutgruppenphänotypen in der Reihenfolge 0, A, B und AB betrachtet werden. Der Blutgruppenindex des Kindes läuft am schnellsten, der des Präsumtivvaters am zweitschnellsten und der der Mutter am langsamsten. Insgesamt acht der 64 Beobachtungskategorien kommen in der EQN-Datei nicht vor, da sie ausgeschlossen sind, wenn die nominelle Mutter die biologische Mutter ist. Dies betrifft alle Fälle mit Blutgruppe 0 bei der Mutter und Blutgruppe AB beim Kind (Kategorien 13 bis 16) sowie alle Fälle mit Blutgruppe AB bei der Mutter und Blutgruppe 0 beim Kind (Kategorien 49 bis 52).[109]

Die in Abschnitt 6.2 berichteten Analysen basieren auf den Daten von Mickey et al. (1986, S. 126/127, vgl. auch Tabelle 6.2 in der vorliegenden Arbeit). Sie ergeben sich aus der Summierung der einander korrespondierenden Häufigkeiten aus Tabelle 3 (für die nicht ausgeschlossenen Väter) und Tabelle 4 (für die ausge-

[109]Dies ist eine kleine Abweichung von der eingangs beschriebenen Syntax, die eine durchgehende Numerierung verlangt. Diese ist nötig, wenn die EQN-Datei mit MBT analysiert werden soll; eine Analyse mit AppleTree ist jedoch auch bei nicht durchgehender Numerierung problemlos möglich. Um den Bezug zu den 64 Zellen der Tabelle 6.2 nicht zu verlieren, wurde hier keine durchgehende Numerierung gewählt.

schlossenen Väter). Die daraus resultierende MDT-Datei für alle 64-8 = 56 möglichen Blutgruppenkombinationen (ohne die Kategorien 13 bis 16 und 49 bis 52) sieht wie folgt aus:

```
ABO-Paternity-Data von Mickey et al. 1986, Tabelle 3 plus Tabelle 4
01    239
02    127
03    38
04    3
05    11
06    124
07    3
08    16
09    4
10    3
11    22
12    7
17    93
18    56
19    8
20    0
21    173
22    203
23    20
24    16
25    2
26    2
27    18
28    5
29    4
30    0
31    15
32    2
33    33
34    11
35    2
36    1
37    2
38    6
39    1
40    1
41    25
42    11
43    12
44    1
45    1
46    14
47    0
48    0
53    12
54    15
55    2
56    0
57    8
58    8
59    1
60    0
61    2
62    6
63    1
64    3
====
```

5) *Datei „LCA8KL.EQN" zur Analyse des restringierten 8-Klassen-LCA-Modells für die Untersuchung von Levine (1996)*

Das in Abschnitt 2.5 beschriebene und in Abschnitt 6.3 analysierte Modell besteht aus einem Baum mit 128 Ästen, 16 Kategorien und 15 Parametern. Die 16 Kategorien entsprechenden den 16 Zellen der 2^4-Kontingenztafel möglicher Antwortmuster und sind in der lexikographischen Reihenfolge (r, r, r, r), (r, r, r, l), (r, r, l, r), (r, r, l, l), ..., (l, l, l, l) von 1 bis 16 durchnumeriert worden, wobei l für „links" und r für „rechts" steht. Die sieben Parameter $h1$ bis $h7$ sind Hilfsparameter, aus denen sich die acht Klassengrößenparameter $p(h_i)$ durch Produktbildung wie folgt ergeben:

$p(h_1) = h_1 \cdot h_2 \cdot h_3$
$p(h_2) = h_1 \cdot h_2 \cdot (1-h_3)$
. . .
$p(h_8) = (1-h_1) \cdot (1-h_7) \cdot (1-h_6)$.

Die restlichen acht *a*- und *b*-Parameter entsprechen den in Tabelle 2.2 angegebenen Antwortfehlerwahrscheinlichkeiten.

Da sich aus einem bestimmten Modell durch Permutation der Klassenindizes äquivalente Modelle ergeben, ist das Modell zunächst einmal nicht identifizierbar. Identifizierbarkeit läßt sich jedoch leicht herstellen, indem man z.B. eine Ordnungsrestriktion auf die Klassengrößenparameter legt, wie dies in Abschnitt 4.3.2.1 beschrieben wurden. Das würde im gegebenen Falle jedoch zu einem MVB-Modell mit sehr vielen Ästen führen, weshalb hierauf verzichtet wurde. Alternative Möglichkeiten, zu identifizierbaren Submodellen zu gelangen, ergeben sich aus Gleichheitsrestriktionen zu den *a*- bzw. *b*-Antwortfehlerwahrscheinlichkeiten.

Es folgt der Inhalt der EQN-Datei:

```
128
1 1 h1*h2*h3*a1*(1-b2)*a3*(1-b4)
1 2 h1*h2*h3*a1*(1-b2)*a3*b4
1 3 h1*h2*h3*a1*(1-b2)*(1-a3)*(1-b4)
1 4 h1*h2*h3*a1*(1-b2)*(1-a3)*b4
1 5 h1*h2*h3*a1*b2*a3*(1-b4)
1 6 h1*h2*h3*a1*b2*a3*b4
1 7 h1*h2*h3*a1*b2*(1-a3)*(1-b4)
1 8 h1*h2*h3*a1*b2*(1-a3)*b4
1 9 h1*h2*h3*(1-a1)*(1-b2)*a3*(1-b4)
1 10 h1*h2*h3*(1-a1)*(1-b2)*a3*b4
1 11 h1*h2*h3*(1-a1)*(1-b2)*(1-a3)*(1-b4)
1 12 h1*h2*h3*(1-a1)*(1-b2)*(1-a3)*b4
1 13 h1*h2*h3*(1-a1)*b2*a3*(1-b4)
1 14 h1*h2*h3*(1-a1)*b2*a3*b4
1 15 h1*h2*h3*(1-a1)*b2*(1-a3)*(1-b4)
1 16 h1*h2*h3*(1-a1)*b2*(1-a3)*b4
1 1 h1*h2*(1-h3)*(1-b1)*a2*(1-b3)*a4
1 2 h1*h2*(1-h3)*(1-b1)*a2*(1-b3)*(1-a4)
1 3 h1*h2*(1-h3)*(1-b1)*a2*b3*a4
1 4 h1*h2*(1-h3)*(1-b1)*a2*b3*(1-a4)
1 5 h1*h2*(1-h3)*(1-b1)*(1-a2)*(1-b3)*a4
1 6 h1*h2*(1-h3)*(1-b1)*(1-a2)*(1-b3)*(1-a4)
1 7 h1*h2*(1-h3)*(1-b1)*(1-a2)*b3*a4
1 8 h1*h2*(1-h3)*(1-b1)*(1-a2)*b3*(1-a4)
```

```
1 9 h1*h2*(1-h3)*b1*a2*(1-b3)*a4
1 10 h1*h2*(1-h3)*b1*a2*(1-b3)*(1-a4)
1 11 h1*h2*(1-h3)*b1*a2*b3*a4
1 12 h1*h2*(1-h3)*b1*a2*b3*(1-a4)
1 13 h1*h2*(1-h3)*b1*(1-a2)*(1-b3)*a4
1 14 h1*h2*(1-h3)*b1*(1-a2)*(1-b3)*(1-a4)
1 15 h1*h2*(1-h3)*b1*(1-a2)*b3*a4
1 16 h1*h2*(1-h3)*b1*(1-a2)*b3*(1-a4)
1 1 h1*(1-h2)*h4*a1*a2*a3*a4
1 2 h1*(1-h2)*h4*a1*a2*a3*(1-a4)
1 3 h1*(1-h2)*h4*a1*a2*(1-a3)*a4
1 4 h1*(1-h2)*h4*a1*a2*(1-a3)*(1-a4)
1 5 h1*(1-h2)*h4*a1*(1-a2)*a3*a4
1 6 h1*(1-h2)*h4*a1*(1-a2)*a3*(1-a4)
1 7 h1*(1-h2)*h4*a1*(1-a2)*(1-a3)*a4
1 8 h1*(1-h2)*h4*a1*(1-a2)*(1-a3)*(1-a4)
1 9 h1*(1-h2)*h4*(1-a1)*a2*a3*a4
1 10 h1*(1-h2)*h4*(1-a1)*a2*a3*(1-a4)
1 11 h1*(1-h2)*h4*(1-a1)*a2*(1-a3)*a4
1 12 h1*(1-h2)*h4*(1-a1)*a2*(1-a3)*(1-a4)
1 13 h1*(1-h2)*h4*(1-a1)*(1-a2)*a3*a4
1 14 h1*(1-h2)*h4*(1-a1)*(1-a2)*a3*(1-a4)
1 15 h1*(1-h2)*h4*(1-a1)*(1-a2)*(1-a3)*a4
1 16 h1*(1-h2)*h4*(1-a1)*(1-a2)*(1-a3)*(1-a4)
1 1 h1*(1-h2)*(1-h4)*(1-b1)*(1-b2)*(1-b3)*(1-b4)
1 2 h1*(1-h2)*(1-h4)*(1-b1)*(1-b2)*(1-b3)*b4
1 3 h1*(1-h2)*(1-h4)*(1-b1)*(1-b2)*b3*(1-b4)
1 4 h1*(1-h2)*(1-h4)*(1-b1)*(1-b2)*b3*b4
1 5 h1*(1-h2)*(1-h4)*(1-b1)*b2*(1-b3)*(1-b4)
1 6 h1*(1-h2)*(1-h4)*(1-b1)*b2*(1-b3)*b4
1 7 h1*(1-h2)*(1-h4)*(1-b1)*b2*b3*(1-b4)
1 8 h1*(1-h2)*(1-h4)*(1-b1)*b2*b3*b4
1 9 h1*(1-h2)*(1-h4)*b1*(1-b2)*(1-b3)*(1-b4)
1 10 h1*(1-h2)*(1-h4)*b1*(1-b2)*(1-b3)*b4
1 11 h1*(1-h2)*(1-h4)*b1*(1-b2)*b3*(1-b4)
1 12 h1*(1-h2)*(1-h4)*b1*(1-b2)*b3*b4
1 13 h1*(1-h2)*(1-h4)*b1*b2*(1-b3)*(1-b4)
1 14 h1*(1-h2)*(1-h4)*b1*b2*(1-b3)*b4
1 15 h1*(1-h2)*(1-h4)*b1*b2*b3*(1-b4)
1 16 h1*(1-h2)*(1-h4)*b1*b2*b3*b4
1 1 (1-h1)*h7*h5*a1*(1-b2)*(1-b3)*a4
1 2 (1-h1)*h7*h5*a1*(1-b2)*(1-b3)*(1-a4)
1 3 (1-h1)*h7*h5*a1*(1-b2)*b3*a4
1 4 (1-h1)*h7*h5*a1*(1-b2)*b3*(1-a4)
1 5 (1-h1)*h7*h5*a1*b2*(1-b3)*a4
1 6 (1-h1)*h7*h5*a1*b2*(1-b3)*(1-a4)
1 7 (1-h1)*h7*h5*a1*b2*b3*a4
1 8 (1-h1)*h7*h5*a1*b2*b3*(1-a4)
1 9 (1-h1)*h7*h5*(1-a1)*(1-b2)*(1-b3)*a4
1 10 (1-h1)*h7*h5*(1-a1)*(1-b2)*(1-b3)*(1-a4)
1 11 (1-h1)*h7*h5*(1-a1)*(1-b2)*b3*a4
1 12 (1-h1)*h7*h5*(1-a1)*(1-b2)*b3*(1-a4)
1 13 (1-h1)*h7*h5*(1-a1)*b2*(1-b3)*a4
1 14 (1-h1)*h7*h5*(1-a1)*b2*(1-b3)*(1-a4)
1 15 (1-h1)*h7*h5*(1-a1)*b2*b3*a4
1 16 (1-h1)*h7*h5*(1-a1)*b2*b3*(1-a4)
1 1 (1-h1)*h7*(1-h5)*(1-b1)*a2*a3*(1-b4)
1 2 (1-h1)*h7*(1-h5)*(1-b1)*a2*a3*b4
1 3 (1-h1)*h7*(1-h5)*(1-b1)*a2*(1-a3)*(1-b4)
1 4 (1-h1)*h7*(1-h5)*(1-b1)*a2*(1-a3)*b4
1 5 (1-h1)*h7*(1-h5)*(1-b1)*(1-a2)*a3*(1-b4)
1 6 (1-h1)*h7*(1-h5)*(1-b1)*(1-a2)*a3*b4
1 7 (1-h1)*h7*(1-h5)*(1-b1)*(1-a2)*(1-a3)*(1-b4)
1 8 (1-h1)*h7*(1-h5)*(1-b1)*(1-a2)*(1-a3)*b4
1 9 (1-h1)*h7*(1-h5)*b1*a2*a3*(1-b4)
1 10 (1-h1)*h7*(1-h5)*b1*a2*a3*b4
1 11 (1-h1)*h7*(1-h5)*b1*a2*(1-a3)*(1-b4)
1 12 (1-h1)*h7*(1-h5)*b1*a2*(1-a3)*b4
1 13 (1-h1)*h7*(1-h5)*b1*(1-a2)*a3*(1-b4)
1 14 (1-h1)*h7*(1-h5)*b1*(1-a2)*a3*b4
1 15 (1-h1)*h7*(1-h5)*b1*(1-a2)*(1-a3)*(1-b4)
1 16 (1-h1)*h7*(1-h5)*b1*(1-a2)*(1-a3)*b4
```

```
1 1 (1-h1)*(1-h7)*h6*a1*a2*(1-b3)*(1-b4)
1 2 (1-h1)*(1-h7)*h6*a1*a2*(1-b3)*b4
1 3 (1-h1)*(1-h7)*h6*a1*a2*b3*(1-b4)
1 4 (1-h1)*(1-h7)*h6*a1*a2*b3*b4
1 5 (1-h1)*(1-h7)*h6*a1*(1-a2)*(1-b3)*(1-b4)
1 6 (1-h1)*(1-h7)*h6*a1*(1-a2)*(1-b3)*b4
1 7 (1-h1)*(1-h7)*h6*a1*(1-a2)*b3*(1-b4)
1 8 (1-h1)*(1-h7)*h6*a1*(1-a2)*b3*b4
1 9 (1-h1)*(1-h7)*h6*(1-a1)*a2*(1-b3)*(1-b4)
1 10 (1-h1)*(1-h7)*h6*(1-a1)*a2*(1-b3)*b4
1 11 (1-h1)*(1-h7)*h6*(1-a1)*a2*b3*(1-b4)
1 12 (1-h1)*(1-h7)*h6*(1-a1)*a2*b3*b4
1 13 (1-h1)*(1-h7)*h6*(1-a1)*(1-a2)*(1-b3)*(1-b4)
1 14 (1-h1)*(1-h7)*h6*(1-a1)*(1-a2)*(1-b3)*b4
1 15 (1-h1)*(1-h7)*h6*(1-a1)*(1-a2)*b3*(1-b4)
1 16 (1-h1)*(1-h7)*h6*(1-a1)*(1-a2)*b3*b4
1 1 (1-h1)*(1-h7)*(1-h6)*(1-b1)*(1-b2)*a3*a4
1 2 (1-h1)*(1-h7)*(1-h6)*(1-b1)*(1-b2)*a3*(1-a4)
1 3 (1-h1)*(1-h7)*(1-h6)*(1-b1)*(1-b2)*(1-a3)*a4
1 4 (1-h1)*(1-h7)*(1-h6)*(1-b1)*(1-b2)*(1-a3)*(1-a4)
1 5 (1-h1)*(1-h7)*(1-h6)*(1-b1)*b2*a3*a4
1 6 (1-h1)*(1-h7)*(1-h6)*(1-b1)*b2*a3*(1-a4)
1 7 (1-h1)*(1-h7)*(1-h6)*(1-b1)*b2*(1-a3)*a4
1 8 (1-h1)*(1-h7)*(1-h6)*(1-b1)*b2*(1-a3)*(1-a4)
1 9 (1-h1)*(1-h7)*(1-h6)*b1*(1-b2)*a3*a4
1 10 (1-h1)*(1-h7)*(1-h6)*b1*(1-b2)*a3*(1-a4)
1 11 (1-h1)*(1-h7)*(1-h6)*b1*(1-b2)*(1-a3)*a4
1 12 (1-h1)*(1-h7)*(1-h6)*b1*(1-b2)*(1-a3)*(1-a4)
1 13 (1-h1)*(1-h7)*(1-h6)*b1*b2*a3*a4
1 14 (1-h1)*(1-h7)*(1-h6)*b1*b2*a3*(1-a4)
1 15 (1-h1)*(1-h7)*(1-h6)*b1*b2*(1-a3)*a4
1 16 (1-h1)*(1-h7)*(1-h6)*b1*b2*(1-a3)*(1-a4)
```

Die drei Datensätze zu den drei Blank-Trial-Sequenzen der in Abschnitt 6.3.2 referierten Untersuchung finden sich in folgender MDT-Datei:

```
Levine Daten Trier SS98, Item 2-5
1      0
2      1
3      1
4      15
5      2
6      28
7      9
8      2
9      0
10     4
11     1
12     3
13     1
14     2
15     0
16     1
==================================
Levine Daten Trier SS98, Item 7-10
1      0
2      0
3      1
4      0
5      1
6      43
7      9
8      1
9      0
10     1
11     4
12     2
13     6
```

```
14     0
15     0
16     1
==================================
Levine Daten Trier SS98, Item 12-15
1      1
2      0
3      2
4      5
5      2
6      3
7      42
8      1
9      0
10     3
11     1
12     1
13     2
14     2
15     2
16     1
==================================
```

Die Kategoriennumerierung folgt dem o.g. EQN-file zum LCA-Modell.

6) *Datei „LINDA4GR.EQN" zur Analyse des Binomialmischverteilungsmodells zum Konjunktionsfehler*

Das in Abschnitt 6.4.1 und 6.4.2 behandelte Modell zum Konjunktionsfehler kann mit der folgenden EQN-Datei analysiert werden. Der Modellbaum besteht aus 256 Ästen und ist deshalb mit dem DOS-Programm MBT sicher nicht mehr handhabbar. Genügend Arbeitsspeicher vorausgesetzt, kann dieses Modell aber mit AppleTree oder GPT analysiert werden.

Trotz der vielen Äste bezieht sich das Modell auf nur vier Bäume, die den vier Versuchsbedingungen der in Abschnitt 6.4.2 geschilderten Experimente entsprechen (Häufigkeits- versus Wahrscheinlichkeitsformat gekreuzt mit absoluter Schätzung versus Rangbildung der Antwortalternativen). Die Ereigniskategorien entsprechen der Anzahl der Konjunktionsfehler in den fünf Aufgaben in aufsteigender Reihenfolge. In der ersten Versuchsbedingung steht also beispielsweise die Kategorie 1 für „keine Konjunktionsfehler" und 6 für „fünf Konjunktionsfehler". Die Parameter p11 bis p22 repräsentieren die Größe der ersten latenten Klasse in den vier Versuchsbedingungen. Die Parameter s111 bis s122 stehen für die Konjunktionsfehlerwahrscheinlichkeiten in Klasse 1 und s211 bis s222 für die Konjunktionsfehlerwahrscheinlichkeiten in Klasse 2, wiederum über die vier Versuchsbedingungen hinweg. Die letzten beiden Ziffern stehen also immer für die Ausprägungen der beiden dichotomen unabhängigen Variablen des Experiments.

```
256
1 1 p11*(1-s111)*(1-s111)*(1-s111)*(1-s111)*(1-s111)
1 2 p11*(1-s111)*(1-s111)*(1-s111)*(1-s111)*s111
```

```
1 2 p11*(1-s111)*(1-s111)*(1-s111)*s111*(1-s111)
1 3 p11*(1-s111)*(1-s111)*(1-s111)*s111*s111
1 2 p11*(1-s111)*(1-s111)*s111*(1-s111)*(1-s111)
1 3 p11*(1-s111)*(1-s111)*s111*(1-s111)*s111
1 3 p11*(1-s111)*(1-s111)*s111*s111*(1-s111)
1 4 p11*(1-s111)*(1-s111)*s111*s111*s111
1 2 p11*(1-s111)*s111*(1-s111)*(1-s111)*(1-s111)
1 3 p11*(1-s111)*s111*(1-s111)*(1-s111)*s111
1 3 p11*(1-s111)*s111*(1-s111)*s111*(1-s111)
1 4 p11*(1-s111)*s111*(1-s111)*s111*s111
1 3 p11*(1-s111)*s111*s111*(1-s111)*(1-s111)
1 4 p11*(1-s111)*s111*s111*(1-s111)*s111
1 4 p11*(1-s111)*s111*s111*s111*(1-s111)
1 5 p11*(1-s111)*s111*s111*s111*s111
1 2 p11*s111*(1-s111)*(1-s111)*(1-s111)*(1-s111)
1 3 p11*s111*(1-s111)*(1-s111)*(1-s111)*s111
1 3 p11*s111*(1-s111)*(1-s111)*s111*(1-s111)
1 4 p11*s111*(1-s111)*(1-s111)*s111*s111
1 3 p11*s111*(1-s111)*s111*(1-s111)*(1-s111)
1 4 p11*s111*(1-s111)*s111*(1-s111)*s111
1 4 p11*s111*(1-s111)*s111*s111*(1-s111)
1 5 p11*s111*(1-s111)*s111*s111*s111
1 3 p11*s111*s111*(1-s111)*(1-s111)*(1-s111)
1 4 p11*s111*s111*(1-s111)*(1-s111)*s111
1 4 p11*s111*s111*(1-s111)*s111*(1-s111)
1 5 p11*s111*s111*(1-s111)*s111*s111
1 4 p11*s111*s111*s111*(1-s111)*(1-s111)
1 5 p11*s111*s111*s111*(1-s111)*s111
1 5 p11*s111*s111*s111*s111*(1-s111)
1 6 p11*s111*s111*s111*s111*s111
1 1 (1-p11)*(1-s112)*(1-s112)*(1-s112)*(1-s112)*(1-s112)
1 2 (1-p11)*(1-s112)*(1-s112)*(1-s112)*(1-s112)*s112
1 2 (1-p11)*(1-s112)*(1-s112)*(1-s112)*s112*(1-s112)
1 3 (1-p11)*(1-s112)*(1-s112)*(1-s112)*s112*s112
1 2 (1-p11)*(1-s112)*(1-s112)*s112*(1-s112)*(1-s112)
1 3 (1-p11)*(1-s112)*(1-s112)*s112*(1-s112)*s112
1 3 (1-p11)*(1-s112)*(1-s112)*s112*s112*(1-s112)
1 4 (1-p11)*(1-s112)*(1-s112)*s112*s112*s112
1 2 (1-p11)*(1-s112)*s112*(1-s112)*(1-s112)*(1-s112)
1 3 (1-p11)*(1-s112)*s112*(1-s112)*(1-s112)*s112
1 3 (1-p11)*(1-s112)*s112*(1-s112)*s112*(1-s112)
1 4 (1-p11)*(1-s112)*s112*(1-s112)*s112*s112
1 3 (1-p11)*(1-s112)*s112*s112*(1-s112)*(1-s112)
1 4 (1-p11)*(1-s112)*s112*s112*(1-s112)*s112
1 4 (1-p11)*(1-s112)*s112*s112*s112*(1-s112)
1 5 (1-p11)*(1-s112)*s112*s112*s112*s112
1 2 (1-p11)*s112*(1-s112)*(1-s112)*(1-s112)*(1-s112)
1 3 (1-p11)*s112*(1-s112)*(1-s112)*(1-s112)*s112
1 3 (1-p11)*s112*(1-s112)*(1-s112)*s112*(1-s112)
1 4 (1-p11)*s112*(1-s112)*(1-s112)*s112*s112
1 3 (1-p11)*s112*(1-s112)*s112*(1-s112)*(1-s112)
1 4 (1-p11)*s112*(1-s112)*s112*(1-s112)*s112
1 4 (1-p11)*s112*(1-s112)*s112*s112*(1-s112)
1 5 (1-p11)*s112*(1-s112)*s112*s112*s112
1 3 (1-p11)*s112*s112*(1-s112)*(1-s112)*(1-s112)
1 4 (1-p11)*s112*s112*(1-s112)*(1-s112)*s112
1 4 (1-p11)*s112*s112*(1-s112)*s112*(1-s112)
1 5 (1-p11)*s112*s112*(1-s112)*s112*s112
1 4 (1-p11)*s112*s112*s112*(1-s112)*(1-s112)
1 5 (1-p11)*s112*s112*s112*(1-s112)*s112
1 5 (1-p11)*s112*s112*s112*s112*(1-s112)
1 6 (1-p11)*s112*s112*s112*s112*s112
2 7 p12*(1-s121)*(1-s121)*(1-s121)*(1-s121)*(1-s121)
2 8 p12*(1-s121)*(1-s121)*(1-s121)*(1-s121)*s121
2 8 p12*(1-s121)*(1-s121)*(1-s121)*s121*(1-s121)
2 9 p12*(1-s121)*(1-s121)*(1-s121)*s121*s121
2 8 p12*(1-s121)*(1-s121)*s121*(1-s121)*(1-s121)
2 9 p12*(1-s121)*(1-s121)*s121*(1-s121)*s121
2 9 p12*(1-s121)*(1-s121)*s121*s121*(1-s121)
2 10 p12*(1-s121)*(1-s121)*s121*s121*s121
2 8 p12*(1-s121)*s121*(1-s121)*(1-s121)*(1-s121)
2 9 p12*(1-s121)*s121*(1-s121)*(1-s121)*s121
```

```
2 9 p12*(1-s121)*s121*(1-s121)*s121*(1-s121)
2 10 p12*(1-s121)*s121*(1-s121)*s121*s121
2 9 p12*(1-s121)*s121*s121*(1-s121)*(1-s121)
2 10 p12*(1-s121)*s121*s121*(1-s121)*s121
2 10 p12*(1-s121)*s121*s121*s121*(1-s121)
2 11 p12*(1-s121)*s121*s121*s121*s121
2 8 p12*s121*(1-s121)*(1-s121)*(1-s121)*(1-s121)
2 9 p12*s121*(1-s121)*(1-s121)*(1-s121)*s121
2 9 p12*s121*(1-s121)*(1-s121)*s121*(1-s121)
2 10 p12*s121*(1-s121)*(1-s121)*s121*s121
2 9 p12*s121*(1-s121)*s121*(1-s121)*(1-s121)
2 10 p12*s121*(1-s121)*s121*(1-s121)*s121
2 10 p12*s121*(1-s121)*s121*s121*(1-s121)
2 11 p12*s121*(1-s121)*s121*s121*s121
2 9 p12*s121*s121*(1-s121)*(1-s121)*(1-s121)
2 10 p12*s121*s121*(1-s121)*(1-s121)*s121
2 10 p12*s121*s121*(1-s121)*s121*(1-s121)
2 11 p12*s121*s121*(1-s121)*s121*s121
2 10 p12*s121*s121*s121*(1-s121)*(1-s121)
2 11 p12*s121*s121*s121*(1-s121)*s121
2 11 p12*s121*s121*s121*s121*(1-s121)
2 12 p12*s121*s121*s121*s121*s121
2 7 (1-p12)*(1-s122)*(1-s122)*(1-s122)*(1-s122)*(1-s122)
2 8 (1-p12)*(1-s122)*(1-s122)*(1-s122)*(1-s122)*s122
2 8 (1-p12)*(1-s122)*(1-s122)*(1-s122)*s122*(1-s122)
2 9 (1-p12)*(1-s122)*(1-s122)*(1-s122)*s122*s122
2 8 (1-p12)*(1-s122)*(1-s122)*s122*(1-s122)*(1-s122)
2 9 (1-p12)*(1-s122)*(1-s122)*s122*(1-s122)*s122
2 9 (1-p12)*(1-s122)*(1-s122)*s122*s122*(1-s122)
2 10 (1-p12)*(1-s122)*(1-s122)*s122*s122*s122
2 8 (1-p12)*(1-s122)*s122*(1-s122)*(1-s122)*(1-s122)
2 9 (1-p12)*(1-s122)*s122*(1-s122)*(1-s122)*s122
2 9 (1-p12)*(1-s122)*s122*(1-s122)*s122*(1-s122)
2 10 (1-p12)*(1-s122)*s122*(1-s122)*s122*s122
2 9 (1-p12)*(1-s122)*s122*s122*(1-s122)*(1-s122)
2 10 (1-p12)*(1-s122)*s122*s122*(1-s122)*s122
2 10 (1-p12)*(1-s122)*s122*s122*s122*(1-s122)
2 11 (1-p12)*(1-s122)*s122*s122*s122*s122
2 8 (1-p12)*s122*(1-s122)*(1-s122)*(1-s122)*(1-s122)
2 9 (1-p12)*s122*(1-s122)*(1-s122)*(1-s122)*s122
2 9 (1-p12)*s122*(1-s122)*(1-s122)*s122*(1-s122)
2 10 (1-p12)*s122*(1-s122)*(1-s122)*s122*s122
2 9 (1-p12)*s122*(1-s122)*s122*(1-s122)*(1-s122)
2 10 (1-p12)*s122*(1-s122)*s122*(1-s122)*s122
2 10 (1-p12)*s122*(1-s122)*s122*s122*(1-s122)
2 11 (1-p12)*s122*(1-s122)*s122*s122*s122
2 9 (1-p12)*s122*s122*(1-s122)*(1-s122)*(1-s122)
2 10 (1-p12)*s122*s122*(1-s122)*(1-s122)*s122
2 10 (1-p12)*s122*s122*(1-s122)*s122*(1-s122)
2 11 (1-p12)*s122*s122*(1-s122)*s122*s122
2 10 (1-p12)*s122*s122*s122*(1-s122)*(1-s122)
2 11 (1-p12)*s122*s122*s122*(1-s122)*s122
2 11 (1-p12)*s122*s122*s122*s122*(1-s122)
2 12 (1-p12)*s122*s122*s122*s122*s122
3 13 p21*(1-s211)*(1-s211)*(1-s211)*(1-s211)*(1-s211)
3 14 p21*(1-s211)*(1-s211)*(1-s211)*(1-s211)*s211
3 14 p21*(1-s211)*(1-s211)*(1-s211)*s211*(1-s211)
3 15 p21*(1-s211)*(1-s211)*(1-s211)*s211*s211
3 14 p21*(1-s211)*(1-s211)*s211*(1-s211)*(1-s211)
3 15 p21*(1-s211)*(1-s211)*s211*(1-s211)*s211
3 15 p21*(1-s211)*(1-s211)*s211*s211*(1-s211)
3 16 p21*(1-s211)*(1-s211)*s211*s211*s211
3 14 p21*(1-s211)*s211*(1-s211)*(1-s211)*(1-s211)
3 15 p21*(1-s211)*s211*(1-s211)*(1-s211)*s211
3 15 p21*(1-s211)*s211*(1-s211)*s211*(1-s211)
3 16 p21*(1-s211)*s211*(1-s211)*s211*s211
3 15 p21*(1-s211)*s211*s211*(1-s211)*(1-s211)
3 16 p21*(1-s211)*s211*s211*(1-s211)*s211
3 16 p21*(1-s211)*s211*s211*s211*(1-s211)
3 17 p21*(1-s211)*s211*s211*s211*s211
3 14 p21*s211*(1-s211)*(1-s211)*(1-s211)*(1-s211)
3 15 p21*s211*(1-s211)*(1-s211)*(1-s211)*s211
```

```
3 15 p21*s211*(1-s211)*(1-s211)*s211*(1-s211)
3 16 p21*s211*(1-s211)*(1-s211)*s211*s211
3 15 p21*s211*(1-s211)*s211*(1-s211)*(1-s211)
3 16 p21*s211*(1-s211)*s211*(1-s211)*s211
3 16 p21*s211*(1-s211)*s211*s211*(1-s211)
3 17 p21*s211*(1-s211)*s211*s211*s211
3 15 p21*s211*s211*(1-s211)*(1-s211)*(1-s211)
3 16 p21*s211*s211*(1-s211)*(1-s211)*s211
3 16 p21*s211*s211*(1-s211)*s211*(1-s211)
3 17 p21*s211*s211*(1-s211)*s211*s211
3 16 p21*s211*s211*s211*(1-s211)*(1-s211)
3 17 p21*s211*s211*s211*(1-s211)*s211
3 17 p21*s211*s211*s211*s211*(1-s211)
3 18 p21*s211*s211*s211*s211*s211
3 13 (1-p21)*(1-s212)*(1-s212)*(1-s212)*(1-s212)*(1-s212)
3 14 (1-p21)*(1-s212)*(1-s212)*(1-s212)*(1-s212)*s212
3 14 (1-p21)*(1-s212)*(1-s212)*(1-s212)*s212*(1-s212)
3 15 (1-p21)*(1-s212)*(1-s212)*(1-s212)*s212*s212
3 14 (1-p21)*(1-s212)*(1-s212)*s212*(1-s212)*(1-s212)
3 15 (1-p21)*(1-s212)*(1-s212)*s212*(1-s212)*s212
3 15 (1-p21)*(1-s212)*(1-s212)*s212*s212*(1-s212)
3 16 (1-p21)*(1-s212)*(1-s212)*s212*s212*s212
3 14 (1-p21)*(1-s212)*s212*(1-s212)*(1-s212)*(1-s212)
3 15 (1-p21)*(1-s212)*s212*(1-s212)*(1-s212)*s212
3 15 (1-p21)*(1-s212)*s212*(1-s212)*s212*(1-s212)
3 16 (1-p21)*(1-s212)*s212*(1-s212)*s212*s212
3 15 (1-p21)*(1-s212)*s212*s212*(1-s212)*(1-s212)
3 16 (1-p21)*(1-s212)*s212*s212*(1-s212)*s212
3 16 (1-p21)*(1-s212)*s212*s212*s212*(1-s212)
3 17 (1-p21)*(1-s212)*s212*s212*s212*s212
3 14 (1-p21)*s212*(1-s212)*(1-s212)*(1-s212)*(1-s212)
3 15 (1-p21)*s212*(1-s212)*(1-s212)*(1-s212)*s212
3 15 (1-p21)*s212*(1-s212)*(1-s212)*s212*(1-s212)
3 16 (1-p21)*s212*(1-s212)*(1-s212)*s212*s212
3 15 (1-p21)*s212*(1-s212)*s212*(1-s212)*(1-s212)
3 16 (1-p21)*s212*(1-s212)*s212*(1-s212)*s212
3 16 (1-p21)*s212*(1-s212)*s212*s212*(1-s212)
3 17 (1-p21)*s212*(1-s212)*s212*s212*s212
3 15 (1-p21)*s212*s212*(1-s212)*(1-s212)*(1-s212)
3 16 (1-p21)*s212*s212*(1-s212)*(1-s212)*s212
3 16 (1-p21)*s212*s212*(1-s212)*s212*(1-s212)
3 17 (1-p21)*s212*s212*(1-s212)*s212*s212
3 16 (1-p21)*s212*s212*s212*(1-s212)*(1-s212)
3 17 (1-p21)*s212*s212*s212*(1-s212)*s212
3 17 (1-p21)*s212*s212*s212*s212*(1-s212)
3 18 (1-p21)*s212*s212*s212*s212*s212
4 19 p22*(1-s221)*(1-s221)*(1-s221)*(1-s221)*(1-s221)
4 20 p22*(1-s221)*(1-s221)*(1-s221)*(1-s221)*s221
4 20 p22*(1-s221)*(1-s221)*(1-s221)*s221*(1-s221)
4 21 p22*(1-s221)*(1-s221)*(1-s221)*s221*s221
4 20 p22*(1-s221)*(1-s221)*s221*(1-s221)*(1-s221)
4 21 p22*(1-s221)*(1-s221)*s221*(1-s221)*s221
4 21 p22*(1-s221)*(1-s221)*s221*s221*(1-s221)
4 22 p22*(1-s221)*(1-s221)*s221*s221*s221
4 20 p22*(1-s221)*s221*(1-s221)*(1-s221)*(1-s221)
4 21 p22*(1-s221)*s221*(1-s221)*(1-s221)*s221
4 21 p22*(1-s221)*s221*(1-s221)*s221*(1-s221)
4 22 p22*(1-s221)*s221*(1-s221)*s221*s221
4 21 p22*(1-s221)*s221*s221*(1-s221)*(1-s221)
4 22 p22*(1-s221)*s221*s221*(1-s221)*s221
4 22 p22*(1-s221)*s221*s221*s221*(1-s221)
4 23 p22*(1-s221)*s221*s221*s221*s221
4 20 p22*s221*(1-s221)*(1-s221)*(1-s221)*(1-s221)
4 21 p22*s221*(1-s221)*(1-s221)*(1-s221)*s221
4 21 p22*s221*(1-s221)*(1-s221)*s221*(1-s221)
4 22 p22*s221*(1-s221)*(1-s221)*s221*s221
4 21 p22*s221*(1-s221)*s221*(1-s221)*(1-s221)
4 22 p22*s221*(1-s221)*s221*(1-s221)*s221
4 22 p22*s221*(1-s221)*s221*s221*(1-s221)
4 23 p22*s221*(1-s221)*s221*s221*s221
4 21 p22*s221*s221*(1-s221)*(1-s221)*(1-s221)
4 22 p22*s221*s221*(1-s221)*(1-s221)*s221
```

```
4 22 p22*s221*s221*(1-s221)*s221*(1-s221)
4 23 p22*s221*s221*(1-s221)*s221*s221
4 22 p22*s221*s221*s221*(1-s221)*(1-s221)
4 23 p22*s221*s221*s221*(1-s221)*s221
4 23 p22*s221*s221*s221*s221*(1-s221)
4 24 p22*s221*s221*s221*s221*s221
4 19 (1-p22)*(1-s222)*(1-s222)*(1-s222)*(1-s222)*(1-s222)
4 20 (1-p22)*(1-s222)*(1-s222)*(1-s222)*(1-s222)*s222
4 20 (1-p22)*(1-s222)*(1-s222)*(1-s222)*s222*(1-s222)
4 21 (1-p22)*(1-s222)*(1-s222)*(1-s222)*s222*s222
4 20 (1-p22)*(1-s222)*(1-s222)*s222*(1-s222)*(1-s222)
4 21 (1-p22)*(1-s222)*(1-s222)*s222*(1-s222)*s222
4 21 (1-p22)*(1-s222)*(1-s222)*s222*s222*(1-s222)
4 22 (1-p22)*(1-s222)*(1-s222)*s222*s222*s222
4 20 (1-p22)*(1-s222)*s222*(1-s222)*(1-s222)*(1-s222)
4 21 (1-p22)*(1-s222)*s222*(1-s222)*(1-s222)*s222
4 21 (1-p22)*(1-s222)*s222*(1-s222)*s222*(1-s222)
4 22 (1-p22)*(1-s222)*s222*(1-s222)*s222*s222
4 21 (1-p22)*(1-s222)*s222*s222*(1-s222)*(1-s222)
4 22 (1-p22)*(1-s222)*s222*s222*(1-s222)*s222
4 22 (1-p22)*(1-s222)*s222*s222*s222*(1-s222)
4 23 (1-p22)*(1-s222)*s222*s222*s222*s222
4 20 (1-p22)*s222*(1-s222)*(1-s222)*(1-s222)*(1-s222)
4 21 (1-p22)*s222*(1-s222)*(1-s222)*(1-s222)*s222
4 21 (1-p22)*s222*(1-s222)*(1-s222)*s222*(1-s222)
4 22 (1-p22)*s222*(1-s222)*(1-s222)*s222*s222
4 21 (1-p22)*s222*(1-s222)*s222*(1-s222)*(1-s222)
4 22 (1-p22)*s222*(1-s222)*s222*(1-s222)*s222
4 22 (1-p22)*s222*(1-s222)*s222*s222*(1-s222)
4 23 (1-p22)*s222*(1-s222)*s222*s222*s222
4 21 (1-p22)*s222*s222*(1-s222)*(1-s222)*(1-s222)
4 22 (1-p22)*s222*s222*(1-s222)*(1-s222)*s222
4 22 (1-p22)*s222*s222*(1-s222)*s222*(1-s222)
4 23 (1-p22)*s222*s222*(1-s222)*s222*s222
4 22 (1-p22)*s222*s222*s222*(1-s222)*(1-s222)
4 23 (1-p22)*s222*s222*s222*(1-s222)*s222
4 23 (1-p22)*s222*s222*s222*s222*(1-s222)
4 24 (1-p22)*s222*s222*s222*s222*s222
```

Die MDT-Datei zu obiger EQN-Datei, mit der die in Abschnitt 6.4.2 berichteten Ergebnisse ermittelt wurden, ist im folgenden wiedergegeben. Die Reihenfolge der vier Versuchsbedingungen ist immer in der Kopfzeile zu jedem der Experimente angegeben.

```
Linda-Experiment CF4 (Feb. 1998) Reihenfolge: Rank/Hauf, Rank/Wahr, Estim/hauf, Estim/Wahr
1    14
2     4
3     5
4     4
5     2
6     1
7     5
8     5
9     5
10     4
11     6
12     5
13     13
14     8
15     5
16     1
17     2
18     1
19     10
20     8
21     5
22     3
```

23 3
24 1
==========
Linda-Experiment CF5 (Jun. 1998) Reihenfolge: Rank/Hauf, Rank/Wahr, Estim/hauf, Estim/Wahr
1 3
2 4
3 4
4 4
5 3
6 2
7 1
8 6
9 3
10 2
11 6
12 2
13 13
14 3
15 1
16 1
17 2
18 0
19 4
20 2
21 6
22 4
23 2
24 2
==========
Linda-Experiment CF6 (Jan. 1999) Reihenfolge: Rank/Hauf, Rank/Wahr, Estim/hauf, Estim/Wahr
1 4
2 3
3 2
4 5
5 3
6 3
7 1
8 1
9 2
10 4
11 7
12 5
13 12
14 3
15 1
16 3
17 0
18 1
19 0
20 6
21 7
22 5
23 1
24 1
==========
Linda-Experiment CF5 und CF6 aggregiert Reihenfolge: Rank/Hauf, Rank/Wahr, Estim/hauf, Estim/Wahr
1 7
2 7
3 6
4 9
5 6
6 5
7 2
8 7
9 5
10 6
11 13
12 7
13 25
14 6
15 2

```
16      4
17      2
18      1
19      4
20      8
21      13
22      9
23      3
24      3
================================
```

7) *Datei „DUALCO2G.EQN" zur Analyse des Dual-Code-Markoff-Ketten-Modells aus Abschnitt 6.5 und die Datendatei „KLEIN89.MDT" mit den Daten von Klein (1989)*

Das in Abschnitt 6.5.1 beschriebene und in Abschnitt 6.5.2 analysierte Dual-Code-Markoff-Modell umfaßt zwei Ereignisbäume (für abstrakte und konkrete Wörter) mit insgesamt 280 Ästen. Die Bedeutung der Parameter wird in Abschnitt 6.5.2 behandelt. Die EQN-Datei lautet:

```
280
1       1               ai1*bi1*ri1*ri1*ri1
1       1               ai1*bi1*ri1*ri1*(1-ri1)*rv1
1       2               ai1*bi1*ri1*ri1*(1-ri1)*(1-rv1)
1       1               ai1*bi1*ri1*(1-ri1)*rv1*ri1
1       1               ai1*bi1*ri1*(1-ri1)*rv1*(1-ri1)*rv1
1       2               ai1*bi1*ri1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1       3               ai1*bi1*ri1*(1-ri1)*(1-rv1)*ri1
1       3               ai1*bi1*ri1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       4               ai1*bi1*ri1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1      1                ai1*bi1*(1-ri1)*rv1*ri1*ri1
1       1               ai1*bi1*(1-ri1)*rv1*ri1*(1-ri1)*rv1
1       2               ai1*bi1*(1-ri1)*rv1*ri1*(1-ri1)*(1-rv1)
1       1               ai1*bi1*(1-ri1)*rv1*(1-ri1)*rv1*ri1
1       1               ai1*bi1*(1-ri1)*rv1*(1-ri1)*rv1*(1-ri1)*rv1
1       2               ai1*bi1*(1-ri1)*rv1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1      3                ai1*bi1*(1-ri1)*rv1*(1-ri1)*(1-rv1)*ri1
1       3               ai1*bi1*(1-ri1)*rv1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       4               ai1*bi1*(1-ri1)*rv1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       5               ai1*bi1*(1-ri1)*(1-rv1)*ri1*ri1
1       5               ai1*bi1*(1-ri1)*(1-rv1)*ri1*(1-ri1)*rv1
1       6               ai1*bi1*(1-ri1)*(1-rv1)*ri1*(1-ri1)*(1-rv1)
1      5                ai1*bi1*(1-ri1)*(1-rv1)*(1-ri1)*rv1*ri1
1       5               ai1*bi1*(1-ri1)*(1-rv1)*(1-ri1)*rv1*(1-ri1)*rv1
1       6               ai1*bi1*(1-ri1)*(1-rv1)*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1      7                ai1*bi1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)*ri1
1       7               ai1*bi1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       8               ai1*bi1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       1               ai1*(1-bi1)*ri1*b1*ri1*ri1
1       1               ai1*(1-bi1)*ri1*b1*ri1*(1-ri1)*rv1
1       2               ai1*(1-bi1)*ri1*b1*ri1*(1-ri1)*(1-rv1)
1       1               ai1*(1-bi1)*ri1*b1*(1-ri1)*rv1*ri1
1       1               ai1*(1-bi1)*ri1*b1*(1-ri1)*rv1*(1-ri1)*rv1
1       2               ai1*(1-bi1)*ri1*b1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1       3               ai1*(1-bi1)*ri1*b1*(1-ri1)*(1-rv1)*ri1
1       3               ai1*(1-bi1)*ri1*b1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       4               ai1*(1-bi1)*ri1*b1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       1               ai1*(1-bi1)*ri1*(1-b1)*ri1*b1*ri1
1       1               ai1*(1-bi1)*ri1*(1-b1)*ri1*b1*(1-ri1)*rv1
1       2               ai1*(1-bi1)*ri1*(1-b1)*ri1*b1*(1-ri1)*(1-rv1)
1       1               ai1*(1-bi1)*ri1*(1-b1)*ri1*(1-b1)*ri1
1       2               ai1*(1-bi1)*ri1*(1-b1)*ri1*(1-b1)*(1-ri1)
```

```
1       3               ai1*(1-bi1)*ri1*(1-b1)*(1-ri1)*b1*ri1
1       3               ai1*(1-bi1)*ri1*(1-b1)*(1-ri1)*b1*(1-ri1)*rv1
1       4               ai1*(1-bi1)*ri1*(1-b1)*(1-ri1)*b1*(1-ri1)*(1-rv1)
1       3               ai1*(1-bi1)*ri1*(1-b1)*(1-ri1)*(1-b1)*ri1
1       4               ai1*(1-bi1)*ri1*(1-b1)*(1-ri1)*(1-b1)*(1-ri1)
1       5               ai1*(1-bi1)*(1-ri1)*b1*ri1*ri1
1       5               ai1*(1-bi1)*(1-ri1)*b1*ri1*(1-ri1)*rv1
1       6               ai1*(1-bi1)*(1-ri1)*b1*ri1*(1-ri1)*(1-rv1)
1       5               ai1*(1-bi1)*(1-ri1)*b1*(1-ri1)*rv1*ri1
1       5               ai1*(1-bi1)*(1-ri1)*b1*(1-ri1)*rv1*(1-ri1)*rv1
1       6               ai1*(1-bi1)*(1-ri1)*b1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1       7               ai1*(1-bi1)*(1-ri1)*b1*(1-ri1)*(1-rv1)*ri1
1       7               ai1*(1-bi1)*(1-ri1)*b1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       8               ai1*(1-bi1)*(1-ri1)*b1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       5               ai1*(1-bi1)*(1-ri1)*(1-b1)*ri1*b1*ri1
1       5               ai1*(1-bi1)*(1-ri1)*(1-b1)*ri1*b1*(1-ri1)*rv1
1       6               ai1*(1-bi1)*(1-ri1)*(1-b1)*ri1*b1*(1-ri1)*(1-rv1)
1       5               ai1*(1-bi1)*(1-ri1)*(1-b1)*ri1*(1-b1)*ri1
1       6               ai1*(1-bi1)*(1-ri1)*(1-b1)*ri1*(1-b1)*(1-ri1)
1       7               ai1*(1-bi1)*(1-ri1)*(1-b1)*(1-ri1)*b1*ri1
1       7               ai1*(1-bi1)*(1-ri1)*(1-b1)*(1-ri1)*b1*(1-ri1)*rv1
1       8               ai1*(1-bi1)*(1-ri1)*(1-b1)*(1-ri1)*b1*(1-ri1)*(1-rv1)
1       7               ai1*(1-bi1)*(1-ri1)*(1-b1)*(1-ri1)*(1-b1)*ri1
1       8               ai1*(1-bi1)*(1-ri1)*(1-b1)*(1-ri1)*(1-b1)*(1-ri1)
1       1               (1-ai1)*bi1*rv1*a1*ri1*ri1
1       1               (1-ai1)*bi1*rv1*a1*ri1*(1-ri1)*rv1
1       2               (1-ai1)*bi1*rv1*a1*ri1*(1-ri1)*(1-rv1)
1       1               (1-ai1)*bi1*rv1*a1*(1-ri1)*rv1*ri1
1       1               (1-ai1)*bi1*rv1*a1*(1-ri1)*rv1*(1-ri1)*rv1
1       2               (1-ai1)*bi1*rv1*a1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1       3               (1-ai1)*bi1*rv1*a1*(1-ri1)*(1-rv1)*ri1
1       3               (1-ai1)*bi1*rv1*a1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       4               (1-ai1)*bi1*rv1*a1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       1               (1-ai1)*bi1*rv1*(1-a1)*rv1*a1*ri1
1       1               (1-ai1)*bi1*rv1*(1-a1)*rv1*a1*(1-ri1)*rv1
1       2               (1-ai1)*bi1*rv1*(1-a1)*rv1*a1*(1-ri1)*(1-rv1)
1       1               (1-ai1)*bi1*rv1*(1-a1)*rv1*(1-a1)*rv1
1       2               (1-ai1)*bi1*rv1*(1-a1)*rv1*(1-a1)*(1-rv1)
1       3               (1-ai1)*bi1*rv1*(1-a1)*(1-rv1)*a1*ri1
1       3               (1-ai1)*bi1*rv1*(1-a1)*(1-rv1)*a1*(1-ri1)*rv1
1       4               (1-ai1)*bi1*rv1*(1-a1)*(1-rv1)*a1*(1-ri1)*(1-rv1)
1       3               (1-ai1)*bi1*rv1*(1-a1)*(1-rv1)*(1-a1)*rv1
1       4               (1-ai1)*bi1*rv1*(1-a1)*(1-rv1)*(1-a1)*(1-rv1)
1       5               (1-ai1)*bi1*(1-rv1)*a1*ri1*ri1
1       5               (1-ai1)*bi1*(1-rv1)*a1*ri1*(1-ri1)*rv1
1       6               (1-ai1)*bi1*(1-rv1)*a1*ri1*(1-ri1)*(1-rv1)
1       5               (1-ai1)*bi1*(1-rv1)*a1*(1-ri1)*rv1*ri1
1       5               (1-ai1)*bi1*(1-rv1)*a1*(1-ri1)*rv1*(1-ri1)*rv1
1       6               (1-ai1)*bi1*(1-rv1)*a1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1       7               (1-ai1)*bi1*(1-rv1)*a1*(1-ri1)*(1-rv1)*ri1
1       7               (1-ai1)*bi1*(1-rv1)*a1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       8               (1-ai1)*bi1*(1-rv1)*a1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       5               (1-ai1)*bi1*(1-rv1)*(1-a1)*rv1*a1*ri1
1       5               (1-ai1)*bi1*(1-rv1)*(1-a1)*rv1*a1*(1-ri1)*rv1
1       6               (1-ai1)*bi1*(1-rv1)*(1-a1)*rv1*a1*(1-ri1)*(1-rv1)
1       5               (1-ai1)*bi1*(1-rv1)*(1-a1)*rv1*(1-a1)*rv1
1       6               (1-ai1)*bi1*(1-rv1)*(1-a1)*rv1*(1-a1)*(1-rv1)
1       7               (1-ai1)*bi1*(1-rv1)*(1-a1)*(1-rv1)*a1*ri1
1       7               (1-ai1)*bi1*(1-rv1)*(1-a1)*(1-rv1)*a1*(1-ri1)*rv1
1       8               (1-ai1)*bi1*(1-rv1)*(1-a1)*(1-rv1)*a1*(1-ri1)*(1-rv1)
1       7               (1-ai1)*bi1*(1-rv1)*(1-a1)*(1-rv1)*(1-a1)*rv1
1       8               (1-ai1)*bi1*(1-rv1)*(1-a1)*(1-rv1)*(1-a1)*(1-rv1)
1       5               (1-ai1)*(1-bi1)*a1*b1*ri1*ri1
1       5               (1-ai1)*(1-bi1)*a1*b1*ri1*(1-ri1)*rv1
1       6               (1-ai1)*(1-bi1)*a1*b1*ri1*(1-ri1)*(1-rv1)
1       5               (1-ai1)*(1-bi1)*a1*b1*(1-ri1)*rv1*ri1
1       5               (1-ai1)*(1-bi1)*a1*b1*(1-ri1)*rv1*(1-ri1)*rv1
1       6               (1-ai1)*(1-bi1)*a1*b1*(1-ri1)*rv1*(1-ri1)*(1-rv1)
1       7               (1-ai1)*(1-bi1)*a1*b1*(1-ri1)*(1-rv1)*ri1
1       7               (1-ai1)*(1-bi1)*a1*b1*(1-ri1)*(1-rv1)*(1-ri1)*rv1
1       8               (1-ai1)*(1-bi1)*a1*b1*(1-ri1)*(1-rv1)*(1-ri1)*(1-rv1)
1       5               (1-ai1)*(1-bi1)*a1*(1-b1)*ri1*b1*ri1
```

```
1       5               (1-ai1)*(1-bi1)*a1*(1-b1)*ri1*b1*(1-ri1)*rv1
1       6               (1-ai1)*(1-bi1)*a1*(1-b1)*ri1*b1*(1-ri1)*(1-rv1)
1       5               (1-ai1)*(1-bi1)*a1*(1-b1)*ri1*(1-b1)*ri1
1       6               (1-ai1)*(1-bi1)*a1*(1-b1)*ri1*(1-b1)*(1-ri1)
1       7               (1-ai1)*(1-bi1)*a1*(1-b1)*(1-ri1)*b1*ri1
1       7               (1-ai1)*(1-bi1)*a1*(1-b1)*(1-ri1)*b1*(1-ri1)*rv1
1       8               (1-ai1)*(1-bi1)*a1*(1-b1)*(1-ri1)*b1*(1-ri1)*(1-rv1)
1       7               (1-ai1)*(1-bi1)*a1*(1-b1)*(1-ri1)*(1-b1)*ri1
1       8               (1-ai1)*(1-bi1)*a1*(1-b1)*(1-ri1)*(1-b1)*(1-ri1)
1       5               (1-ai1)*(1-bi1)*(1-a1)*b1*rv1*a1*ri1
1       5               (1-ai1)*(1-bi1)*(1-a1)*b1*rv1*a1*(1-ri1)*rv1
1       6               (1-ai1)*(1-bi1)*(1-a1)*b1*rv1*a1*(1-ri1)*(1-rv1)
1       5               (1-ai1)*(1-bi1)*(1-a1)*b1*rv1*(1-a1)*rv1
1       6               (1-ai1)*(1-bi1)*(1-a1)*b1*rv1*(1-a1)*(1-rv1)
1       7               (1-ai1)*(1-bi1)*(1-a1)*b1*(1-rv1)*a1*ri1
1       7               (1-ai1)*(1-bi1)*(1-a1)*b1*(1-rv1)*a1*(1-ri1)*rv1
1       8               (1-ai1)*(1-bi1)*(1-a1)*b1*(1-rv1)*a1*(1-ri1)*(1-rv1)
1       7               (1-ai1)*(1-bi1)*(1-a1)*b1*(1-rv1)*(1-a1)*rv1
1       8               (1-ai1)*(1-bi1)*(1-a1)*b1*(1-rv1)*(1-a1)*(1-rv1)
1       7               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*a1*b1*ri1
1       7               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*a1*b1*(1-ri1)*rv1
1       8               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*a1*b1*(1-ri1)*(1-rv1)
1       7               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*a1*(1-b1)*ri1
1       8               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*a1*(1-b1)*(1-ri1)
1       7               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*(1-a1)*b1*rv1
1       8               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*(1-a1)*b1*(1-rv1)
1       8               (1-ai1)*(1-bi1)*(1-a1)*(1-b1)*(1-a1)*(1-b1)
2       9               ai2*bi2*ri2*ri2*ri2
2       9               ai2*bi2*ri2*ri2*(1-ri2)*rv2
2       10              ai2*bi2*ri2*ri2*(1-ri2)*(1-rv2)
2       9               ai2*bi2*ri2*(1-ri2)*rv2*ri2
2       9               ai2*bi2*ri2*(1-ri2)*rv2*(1-ri2)*rv2
2       10              ai2*bi2*ri2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       11              ai2*bi2*ri2*(1-ri2)*(1-rv2)*ri2
2       11              ai2*bi2*ri2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       12              ai2*bi2*ri2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       9               ai2*bi2*(1-ri2)*rv2*ri2*ri2
2       9               ai2*bi2*(1-ri2)*rv2*ri2*(1-ri2)*rv2
2       10              ai2*bi2*(1-ri2)*rv2*ri2*(1-ri2)*(1-rv2)
2       9               ai2*bi2*(1-ri2)*rv2*(1-ri2)*rv2*ri2
2       9               ai2*bi2*(1-ri2)*rv2*(1-ri2)*rv2*(1-ri2)*rv2
2       10              ai2*bi2*(1-ri2)*rv2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       11              ai2*bi2*(1-ri2)*rv2*(1-ri2)*(1-rv2)*ri2
2       11              ai2*bi2*(1-ri2)*rv2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       12              ai2*bi2*(1-ri2)*rv2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       13              ai2*bi2*(1-ri2)*(1-rv2)*ri2*ri2
2       13              ai2*bi2*(1-ri2)*(1-rv2)*ri2*(1-ri2)*rv2
2       14              ai2*bi2*(1-ri2)*(1-rv2)*ri2*(1-ri2)*(1-rv2)
2       13              ai2*bi2*(1-ri2)*(1-rv2)*(1-ri2)*rv2*ri2
2       13              ai2*bi2*(1-ri2)*(1-rv2)*(1-ri2)*rv2*(1-ri2)*rv2
2       14              ai2*bi2*(1-ri2)*(1-rv2)*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       15              ai2*bi2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)*ri2
2       15              ai2*bi2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       16              ai2*bi2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       9               ai2*(1-bi2)*ri2*b2*ri2*ri2
2       9               ai2*(1-bi2)*ri2*b2*ri2*(1-ri2)*rv2
2       10              ai2*(1-bi2)*ri2*b2*ri2*(1-ri2)*(1-rv2)
2       9               ai2*(1-bi2)*ri2*b2*(1-ri2)*rv2*ri2
2       9               ai2*(1-bi2)*ri2*b2*(1-ri2)*rv2*(1-ri2)*rv2
2       10              ai2*(1-bi2)*ri2*b2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       11              ai2*(1-bi2)*ri2*b2*(1-ri2)*(1-rv2)*ri2
2       11              ai2*(1-bi2)*ri2*b2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       12              ai2*(1-bi2)*ri2*b2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       9               ai2*(1-bi2)*ri2*(1-b2)*ri2*b2*ri2
2       9               ai2*(1-bi2)*ri2*(1-b2)*ri2*b2*(1-ri2)*rv2
2       10              ai2*(1-bi2)*ri2*(1-b2)*ri2*b2*(1-ri2)*(1-rv2)
2       9               ai2*(1-bi2)*ri2*(1-b2)*ri2*(1-b2)*ri2
2       10              ai2*(1-bi2)*ri2*(1-b2)*ri2*(1-b2)*(1-ri2)
2       11              ai2*(1-bi2)*ri2*(1-b2)*(1-ri2)*b2*ri2
2       11              ai2*(1-bi2)*ri2*(1-b2)*(1-ri2)*b2*(1-ri2)*rv2
2       12              ai2*(1-bi2)*ri2*(1-b2)*(1-ri2)*b2*(1-ri2)*(1-rv2)
2       11              ai2*(1-bi2)*ri2*(1-b2)*(1-ri2)*(1-b2)*ri2
```

```
2       12              ai2*(1-bi2)*ri2*(1-b2)*(1-ri2)*(1-b2)*(1-ri2)
2       13              ai2*(1-bi2)*(1-ri2)*b2*ri2*ri2
2       13              ai2*(1-bi2)*(1-ri2)*b2*ri2*(1-ri2)*rv2
2       14              ai2*(1-bi2)*(1-ri2)*b2*ri2*(1-ri2)*(1-rv2)
2       13              ai2*(1-bi2)*(1-ri2)*b2*(1-ri2)*rv2*ri2
2       13              ai2*(1-bi2)*(1-ri2)*b2*(1-ri2)*rv2*(1-ri2)*rv2
2       14              ai2*(1-bi2)*(1-ri2)*b2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       15              ai2*(1-bi2)*(1-ri2)*b2*(1-ri2)*(1-rv2)*ri2
2       15              ai2*(1-bi2)*(1-ri2)*b2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       16              ai2*(1-bi2)*(1-ri2)*b2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       13              ai2*(1-bi2)*(1-ri2)*(1-b2)*ri2*b2*ri2
2       13              ai2*(1-bi2)*(1-ri2)*(1-b2)*ri2*b2*(1-ri2)*rv2
2       14              ai2*(1-bi2)*(1-ri2)*(1-b2)*ri2*b2*(1-ri2)*(1-rv2)
2       13              ai2*(1-bi2)*(1-ri2)*(1-b2)*ri2*(1-b2)*ri2
2       14              ai2*(1-bi2)*(1-ri2)*(1-b2)*ri2*(1-b2)*(1-ri2)
2       15              ai2*(1-bi2)*(1-ri2)*(1-b2)*(1-ri2)*b2*ri2
2       15              ai2*(1-bi2)*(1-ri2)*(1-b2)*(1-ri2)*b2*(1-ri2)*rv2
2       16              ai2*(1-bi2)*(1-ri2)*(1-b2)*(1-ri2)*b2*(1-ri2)*(1-rv2)
2       15              ai2*(1-bi2)*(1-ri2)*(1-b2)*(1-ri2)*(1-b2)*ri2
2       16              ai2*(1-bi2)*(1-ri2)*(1-b2)*(1-ri2)*(1-b2)*(1-ri2)
2       9               (1-ai2)*bi2*rv2*a2*ri2*ri2
2       9               (1-ai2)*bi2*rv2*a2*ri2*(1-ri2)*rv2
2       10              (1-ai2)*bi2*rv2*a2*ri2*(1-ri2)*(1-rv2)
2       9               (1-ai2)*bi2*rv2*a2*(1-ri2)*rv2*ri2
2       9               (1-ai2)*bi2*rv2*a2*(1-ri2)*rv2*(1-ri2)*rv2
2       10              (1-ai2)*bi2*rv2*a2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       11              (1-ai2)*bi2*rv2*a2*(1-ri2)*(1-rv2)*ri2
2       11              (1-ai2)*bi2*rv2*a2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       12              (1-ai2)*bi2*rv2*a2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       9               (1-ai2)*bi2*rv2*(1-a2)*rv2*a2*ri2
2       9               (1-ai2)*bi2*rv2*(1-a2)*rv2*a2*(1-ri2)*rv2
2       10              (1-ai2)*bi2*rv2*(1-a2)*rv2*a2*(1-ri2)*(1-rv2)
2       9               (1-ai2)*bi2*rv2*(1-a2)*rv2*(1-a2)*rv2
2       10              (1-ai2)*bi2*rv2*(1-a2)*rv2*(1-a2)*(1-rv2)
2       11              (1-ai2)*bi2*rv2*(1-a2)*(1-rv2)*a2*ri2
2       11              (1-ai2)*bi2*rv2*(1-a2)*(1-rv2)*a2*(1-ri2)*rv2
2       12              (1-ai2)*bi2*rv2*(1-a2)*(1-rv2)*a2*(1-ri2)*(1-rv2)
2       11              (1-ai2)*bi2*rv2*(1-a2)*(1-rv2)*(1-a2)*rv2
2       12              (1-ai2)*bi2*rv2*(1-a2)*(1-rv2)*(1-a2)*(1-rv2)
2       13              (1-ai2)*bi2*(1-rv2)*a2*ri2*ri2
2       13              (1-ai2)*bi2*(1-rv2)*a2*ri2*(1-ri2)*rv2
2       14              (1-ai2)*bi2*(1-rv2)*a2*ri2*(1-ri2)*(1-rv2)
2       13              (1-ai2)*bi2*(1-rv2)*a2*(1-ri2)*rv2*ri2
2       13              (1-ai2)*bi2*(1-rv2)*a2*(1-ri2)*rv2*(1-ri2)*rv2
2       14              (1-ai2)*bi2*(1-rv2)*a2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       15              (1-ai2)*bi2*(1-rv2)*a2*(1-ri2)*(1-rv2)*ri2
2       15              (1-ai2)*bi2*(1-rv2)*a2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       16              (1-ai2)*bi2*(1-rv2)*a2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       13              (1-ai2)*bi2*(1-rv2)*(1-a2)*rv2*a2*ri2
2       13              (1-ai2)*bi2*(1-rv2)*(1-a2)*rv2*a2*(1-ri2)*rv2
2       14              (1-ai2)*bi2*(1-rv2)*(1-a2)*rv2*a2*(1-ri2)*(1-rv2)
2       13              (1-ai2)*bi2*(1-rv2)*(1-a2)*rv2*(1-a2)*rv2
2       14              (1-ai2)*bi2*(1-rv2)*(1-a2)*rv2*(1-a2)*(1-rv2)
2       15              (1-ai2)*bi2*(1-rv2)*(1-a2)*(1-rv2)*a2*ri2
2       15              (1-ai2)*bi2*(1-rv2)*(1-a2)*(1-rv2)*a2*(1-ri2)*rv2
2       16              (1-ai2)*bi2*(1-rv2)*(1-a2)*(1-rv2)*a2*(1-ri2)*(1-rv2)
2       15              (1-ai2)*bi2*(1-rv2)*(1-a2)*(1-rv2)*(1-a2)*rv2
2       16              (1-ai2)*bi2*(1-rv2)*(1-a2)*(1-rv2)*(1-a2)*(1-rv2)
2       13              (1-ai2)*(1-bi2)*a2*b2*ri2*ri2
2       13              (1-ai2)*(1-bi2)*a2*b2*ri2*(1-ri2)*rv2
2       14              (1-ai2)*(1-bi2)*a2*b2*ri2*(1-ri2)*(1-rv2)
2       13              (1-ai2)*(1-bi2)*a2*b2*(1-ri2)*rv2*ri2
2       13              (1-ai2)*(1-bi2)*a2*b2*(1-ri2)*rv2*(1-ri2)*rv2
2       14              (1-ai2)*(1-bi2)*a2*b2*(1-ri2)*rv2*(1-ri2)*(1-rv2)
2       15              (1-ai2)*(1-bi2)*a2*b2*(1-ri2)*(1-rv2)*ri2
2       15              (1-ai2)*(1-bi2)*a2*b2*(1-ri2)*(1-rv2)*(1-ri2)*rv2
2       16              (1-ai2)*(1-bi2)*a2*b2*(1-ri2)*(1-rv2)*(1-ri2)*(1-rv2)
2       13              (1-ai2)*(1-bi2)*a2*(1-b2)*ri2*b2*ri2
2       13              (1-ai2)*(1-bi2)*a2*(1-b2)*ri2*b2*(1-ri2)*rv2
2       14              (1-ai2)*(1-bi2)*a2*(1-b2)*ri2*b2*(1-ri2)*(1-rv2)
2       13              (1-ai2)*(1-bi2)*a2*(1-b2)*ri2*(1-b2)*ri2
2       14              (1-ai2)*(1-bi2)*a2*(1-b2)*ri2*(1-b2)*(1-ri2)
```

```
2      15          (1-ai2)*(1-bi2)*a2*(1-b2)*(1-ri2)*b2*ri2
2      15          (1-ai2)*(1-bi2)*a2*(1-b2)*(1-ri2)*b2*(1-ri2)*rv2
2      16          (1-ai2)*(1-bi2)*a2*(1-b2)*(1-ri2)*b2*(1-ri2)*(1-rv2)
2      15          (1-ai2)*(1-bi2)*a2*(1-b2)*(1-ri2)*(1-b2)*ri2
2      16          (1-ai2)*(1-bi2)*a2*(1-b2)*(1-ri2)*(1-b2)*(1-ri2)
2      13          (1-ai2)*(1-bi2)*(1-a2)*b2*rv2*a2*ri2
2      13          (1-ai2)*(1-bi2)*(1-a2)*b2*rv2*a2*(1-ri2)*rv2
2      14          (1-ai2)*(1-bi2)*(1-a2)*b2*rv2*a2*(1-ri2)*(1-rv2)
2      13          (1-ai2)*(1-bi2)*(1-a2)*b2*rv2*(1-a2)*rv2
2      14          (1-ai2)*(1-bi2)*(1-a2)*b2*rv2*(1-a2)*(1-rv2)
2      15          (1-ai2)*(1-bi2)*(1-a2)*b2*(1-rv2)*a2*ri2
2      15          (1-ai2)*(1-bi2)*(1-a2)*b2*(1-rv2)*a2*(1-ri2)*rv2
2      16          (1-ai2)*(1-bi2)*(1-a2)*b2*(1-rv2)*a2*(1-ri2)*(1-rv2)
2      15          (1-ai2)*(1-bi2)*(1-a2)*b2*(1-rv2)*(1-a2)*rv2
2      16          (1-ai2)*(1-bi2)*(1-a2)*b2*(1-rv2)*(1-a2)*(1-rv2)
2      15          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*a2*b2*ri2
2      15          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*a2*b2*(1-ri2)*rv2
2      16          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*a2*b2*(1-ri2)*(1-rv2)
2      15          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*a2*(1-b2)*ri2
2      16          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*a2*(1-b2)*(1-ri2)
2      15          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*(1-a2)*b2*rv2
2      16          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*(1-a2)*b2*(1-rv2)
2      16          (1-ai2)*(1-bi2)*(1-a2)*(1-b2)*(1-a2)*(1-b2)
```

Die MDT-Datei mit den Daten von Klein (1989) ist lexikographisch aufgebaut. Die Reihenfolge der Antwortmuster ist +++, ++–, +–+, +––, usw. bis hin zu –––. Hierbei ist + eine korrekte und – eine inkorrekte Reproduktion. Erst werden die Häufigkeiten der acht Antwortmuster zu abstrakten Wörtern (Zeile 1 bis 8), danach die Antworthäufigkeiten zu konkreten Wörtern aufgeführt (Zeile 9 bis 16). Die Datei sieht wie folgt aus:

```
KMK 1989, Gesunde, erst abstrakt, dann konkret
1    15
2    0
3    3
4    2
5    35
6    8
7    47
8    226
9    112
10   2
11   7
12   2
13   96
14   4
15   71
16   42
====
KMK 1989, Linksfokale Pat., erst abstrakt, dann konkret
1    1
2    0
3    1
4    1
5    2
6    1
7    5
8    178
9    23
10   1
11   1
12   0
13   38
14   6
15   37
```

```
16   83
====
KMK 1989, Rechtsfokale Pat., erst abstrakt, dann konkret
1    1
2    0
3    2
4    0
5    2
6    1
7    8
8    217
9    26
10   2
11   2
12   0
13   61
14   7
15   52
16   81
====
```

8) *Datei „FASTGUES.EQN" zur Analyse des Fast-Guess-Modells aus Abschnitt 6. und die Datei „FASTGUES.MDT" mit fiktiven modellkonformen Daten*

Das Fast-Guess-Modell aus Abschnitt 6.8.1 kann für den Fall einer Aufteilung der Antwortzeitverteilungen in vier Kategorien (mit drei Grenzwerten c_1, c_2, c_3) mit folgender EQN-Datei analysiert werden. Die Beobachtungskategorien 1 bis 4 stehen für richtige Antworten, deren Antwortzeiten in die Zeitkategorien 1 bis 4 fallen; analog stehen die Beobachtungskategorien 5 bis 8 für die vier Zeitkategorien bei falschen Antworten. Die Parameter *r* und *b* entsprechen den Rekognitions- bzw. Antworttendenzparameter des Hochschwellen-Modells; die t_r-Parameter kennzeichnen die Antwortzeitverteilung rekognizierter Antworten, die t_b-Parameter analog die Antwortzeitverteilung geratener Antworten. Die Parameter t_{ri} und t_{bi}, i = 1, ..., 3, geben jeweils den Anteil der Antwortzeitverteilung an, der unter dem i-ten Grenzwert c_i liegt. Ohne weitere Parameterrestriktionen ist das Modell nicht identifizierbar. In Abschnitt 6.8.1 wird dargestellt, wie das Modell durch geeignete Wahl der Zeitgrenzwerte c_i identifizierbar gemacht werden kann.

```
12
1      1      r*tr2*tr1
1      2      r*tr2*(1-tr1)
1      3      r*(1-tr2)*tr3
1      4      r*(1-tr2)*(1-tr3)
1      1      (1-r)*b*tb2*tb1
1      2      (1-r)*b*tb2*(1-tb1)
1      3      (1-r)*b*(1-tb2)*tb3
1      4      (1-r)*b*(1-tb2)*(1-tb3)
1      5      (1-r)*(1-b)*tb2*tb1
1      6      (1-r)*(1-b)*tb2*(1-tb1)
1      7      (1-r)*(1-b)*(1-tb2)*tb3
1      8      (1-r)*(1-b)*(1-tb2)*(1-tb3)
```

Die folgende MDT-Datei enthält fiktive modellkonforme Daten zur obigen EQN-Datei mit den Parametern $r = b = t_{r1} = t_{r2} = t_{r3} = .50$ und $t_{b1} = t_{b2} = t_{b3} = .7$. Die Ant-

wortzeitverteilung der geraten Antworten (t_b-Parameter) ist also relativ zur Antwortzeitverteilung der rekognizierten Antworten (t_r-Parameter) in den Bereich kurzer Reaktionszeiten verschoben.

```
Modellkonforme Daten zu 'FASTGUES.EQN'
1       112
2       97
3       97
4       81
5       67
6       29
7       29
8       13
=====
```

9) *Datei „KNOSPACE.EQN" zur Analyse des Wissensraumes aus Abschnitt 7.1.4 und die Datei „KNOSPACE.MDT" mit fiktiven modellkonformen Daten*

Der in Abschnitt 7.1.4 beschriebene Wissensraum zu einer Wissensdomäne mit drei Items A, B und C (vgl. Abbildung 7.1) läßt sich als binäres MVB-Modell mit zwei Ereignisbäumen (für jede der beiden Probandengruppen einen Baum) und insgesamt 112 Ästen darstellen. Obwohl äquivalente Modelle mit weniger Ästen existieren, wird hier die komplette EQN-Datei wiedergegeben, die der Abbildung 7.1 (erweitert auf 2 Gruppen) korrespondiert. Die Parameter des Modells werden in Abschnitt 7.1.4 erläutert. Parameterrestriktionen, die das Modell identifizierbar machen, werden ebenfalls in Abschnitt 7.1.4 behandelt.

```
112
1       1       (1-p01)*a1*pA1*bA*(1-vB)*(1-vC)
1       2       (1-p01)*a1*pA1*bA*(1-vB)*vC
1       3       (1-p01)*a1*pA1*bA*vB*(1-vC)
1       4       (1-p01)*a1*pA1*bA*vB*vC
1       5       (1-p01)*a1*pA1*(1-bA)*(1-vB)*(1-vC)
1       6       (1-p01)*a1*pA1*(1-bA)*(1-vB)*vC
1       7       (1-p01)*a1*pA1*(1-bA)*vB*(1-vC)
1       8       (1-p01)*a1*pA1*(1-bA)*vB*vC
1       1       (1-p01)*a1*(1-pA1)*pAB1*bA*bB*(1-vC)
1       2       (1-p01)*a1*(1-pA1)*pAB1*bA*bB*vC
1       3       (1-p01)*a1*(1-pA1)*pAB1*bA*(1-bB)*(1-vC)
1       4       (1-p01)*a1*(1-pA1)*pAB1*bA*(1-bB)*vC
1       5       (1-p01)*a1*(1-pA1)*pAB1*(1-bA)*bB*(1-vC)
1       6       (1-p01)*a1*(1-pA1)*pAB1*(1-bA)*bB*vC
1       7       (1-p01)*a1*(1-pA1)*pAB1*(1-bA)*(1-bB)*(1-vC)
1       8       (1-p01)*a1*(1-pA1)*pAB1*(1-bA)*(1-bB)*vC
1       1       (1-p01)*a1*(1-pA1)*(1-pAB1)*bA*bB*bC
1       2       (1-p01)*a1*(1-pA1)*(1-pAB1)*bA*bB*(1-bC)
1       3       (1-p01)*a1*(1-pA1)*(1-pAB1)*bA*(1-bB)*bC
1       4       (1-p01)*a1*(1-pA1)*(1-pAB1)*bA*(1-bB)*(1-bC)
1       5       (1-p01)*a1*(1-pA1)*(1-pAB1)*(1-bA)*bB*bC
1       6       (1-p01)*a1*(1-pA1)*(1-pAB1)*(1-bA)*bB*(1-bC)
1       7       (1-p01)*a1*(1-pA1)*(1-pAB1)*(1-bA)*(1-bB)*bC
1       8       (1-p01)*a1*(1-pA1)*(1-pAB1)*(1-bA)*(1-bB)*(1-bC)
1       1       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*bA*bB*bC
1       2       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*bA*bB*(1-bC)
1       3       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*bA*(1-bB)*bC
```

```
1       4       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*bA*(1-bB)*(1-bC)
1       5       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*(1-bA)*bB*bC
1       6       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*(1-bA)*bB*(1-bC)
1       7       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*(1-bA)*(1-bB)*bC
1       8       (1-p01)*(1-a1)*(1-pB1)*(1-pAB1)*(1-bA)*(1-bB)*(1-bC)
1       1       (1-p01)*(1-a1)*(1-pB1)*pAB1*bA*bB*(1-vC)
1       2       (1-p01)*(1-a1)*(1-pB1)*pAB1*bA*bB*vC
1       3       (1-p01)*(1-a1)*(1-pB1)*pAB1*bA*(1-bB)*(1-vC)
1       4       (1-p01)*(1-a1)*(1-pB1)*pAB1*bA*(1-bB)*vC
1       5       (1-p01)*(1-a1)*(1-pB1)*pAB1*(1-bA)*bB*(1-vC)
1       6       (1-p01)*(1-a1)*(1-pB1)*pAB1*(1-bA)*bB*vC
1       7       (1-p01)*(1-a1)*(1-pB1)*pAB1*(1-bA)*(1-bB)*(1-vC)
1       8       (1-p01)*(1-a1)*(1-pB1)*pAB1*(1-bA)*(1-bB)*vC
1       1       (1-p01)*(1-a1)*pB1*(1-vA)*bB*(1-vC)
1       2       (1-p01)*(1-a1)*pB1*(1-vA)*bB*vC
1       3       (1-p01)*(1-a1)*pB1*(1-vA)*(1-bB)*(1-vC)
1       4       (1-p01)*(1-a1)*pB1*(1-vA)*(1-bB)*vC
1       5       (1-p01)*(1-a1)*pB1*vA*bB*(1-vC)
1       6       (1-p01)*(1-a1)*pB1*vA*bB*vC
1       7       (1-p01)*(1-a1)*pB1*vA*(1-bB)*(1-vC)
1       8       (1-p01)*(1-a1)*pB1*vA*(1-bB)*vC
1       1       p01*(1-vA)*(1-vB)*(1-vC)
1       2       p01*(1-vA)*(1-vB)*vC
1       3       p01*(1-vA)*vB*(1-vC)
1       4       p01*(1-vA)*vB*vC
1       5       p01*vA*(1-vB)*(1-vC)
1       6       p01*vA*(1-vB)*vC
1       7       p01*vA*vB*(1-vC)
1       8       p01*vA*vB*vC
2       9       (1-p02)*a2*pA2*bA*(1-vB)*(1-vC)
2       10      (1-p02)*a2*pA2*bA*(1-vB)*vC
2       11      (1-p02)*a2*pA2*bA*vB*(1-vC)
2       12      (1-p02)*a2*pA2*bA*vB*vC
2       13      (1-p02)*a2*pA2*(1-bA)*(1-vB)*(1-vC)
2       14      (1-p02)*a2*pA2*(1-bA)*(1-vB)*vC
2       15      (1-p02)*a2*pA2*(1-bA)*vB*(1-vC)
2       16      (1-p02)*a2*pA2*(1-bA)*vB*vC
2       9       (1-p02)*a2*(1-pA2)*pAB2*bA*bB*(1-vC)
2       10      (1-p02)*a2*(1-pA2)*pAB2*bA*bB*vC
2       11      (1-p02)*a2*(1-pA2)*pAB2*bA*(1-bB)*(1-vC)
2       12      (1-p02)*a2*(1-pA2)*pAB2*bA*(1-bB)*vC
2       13      (1-p02)*a2*(1-pA2)*pAB2*(1-bA)*bB*(1-vC)
2       14      (1-p02)*a2*(1-pA2)*pAB2*(1-bA)*bB*vC
2       15      (1-p02)*a2*(1-pA2)*pAB2*(1-bA)*(1-bB)*(1-vC)
2       16      (1-p02)*a2*(1-pA2)*pAB2*(1-bA)*(1-bB)*vC
2       9       (1-p02)*a2*(1-pA2)*(1-pAB2)*bA*bB*bC
2       10      (1-p02)*a2*(1-pA2)*(1-pAB2)*bA*bB*(1-bC)
2       11      (1-p02)*a2*(1-pA2)*(1-pAB2)*bA*(1-bB)*bC
2       12      (1-p02)*a2*(1-pA2)*(1-pAB2)*bA*(1-bB)*(1-bC)
2       13      (1-p02)*a2*(1-pA2)*(1-pAB2)*(1-bA)*bB*bC
2       14      (1-p02)*a2*(1-pA2)*(1-pAB2)*(1-bA)*bB*(1-bC)
2       15      (1-p02)*a2*(1-pA2)*(1-pAB2)*(1-bA)*(1-bB)*bC
2       16      (1-p02)*a2*(1-pA2)*(1-pAB2)*(1-bA)*(1-bB)*(1-bC)
2       9       (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*bA*bB*bC
2       10      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*bA*bB*(1-bC)
2       11      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*bA*(1-bB)*bC
2       12      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*bA*(1-bB)*(1-bC)
2       13      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*(1-bA)*bB*bC
2       14      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*(1-bA)*bB*(1-bC)
2       15      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*(1-bA)*(1-bB)*bC
2       16      (1-p02)*(1-a2)*(1-pB2)*(1-pAB2)*(1-bA)*(1-bB)*(1-bC)
2       9       (1-p02)*(1-a2)*(1-pB2)*pAB2*bA*bB*(1-vC)
2       10      (1-p02)*(1-a2)*(1-pB2)*pAB2*bA*bB*vC
2       11      (1-p02)*(1-a2)*(1-pB2)*pAB2*bA*(1-bB)*(1-vC)
2       12      (1-p02)*(1-a2)*(1-pB2)*pAB2*bA*(1-bB)*vC
2       13      (1-p02)*(1-a2)*(1-pB2)*pAB2*(1-bA)*bB*(1-vC)
2       14      (1-p02)*(1-a2)*(1-pB2)*pAB2*(1-bA)*bB*vC
2       15      (1-p02)*(1-a2)*(1-pB2)*pAB2*(1-bA)*(1-bB)*(1-vC)
2       16      (1-p02)*(1-a2)*(1-pB2)*pAB2*(1-bA)*(1-bB)*vC
2       9       (1-p02)*(1-a2)*pB2*(1-vA)*bB*(1-vC)
2       10      (1-p02)*(1-a2)*pB2*(1-vA)*bB*vC
2       11      (1-p02)*(1-a2)*pB2*(1-vA)*(1-bB)*(1-vC)
```

```
2       12      (1-p02)*(1-a2)*pB2*(1-vA)*(1-bB)*vC
2       13      (1-p02)*(1-a2)*pB2*vA*bB*(1-vC)
2       14      (1-p02)*(1-a2)*pB2*vA*bB*vC
2       15      (1-p02)*(1-a2)*pB2*vA*(1-bB)*(1-vC)
2       16      (1-p02)*(1-a2)*pB2*vA*(1-bB)*vC
2       9       p02*(1-vA)*(1-vB)*(1-vC)
2       10      p02*(1-vA)*(1-vB)*vC
2       11      p02*(1-vA)*vB*(1-vC)
2       12      p02*(1-vA)*vB*vC
2       13      p02*vA*(1-vB)*(1-vC)
2       14      p02*vA*(1-vB)*vC
2       15      p02*vA*vB*(1-vC)
2       16      p02*vA*vB*vC
```

Die folgende MDT-Datei gibt nahezu perfekt modellkonforme Daten zu obiger EQN-Datei an, wenn man für beide Gruppen und alle Items Fehlerparameter (*v*- und β-Parameter) von vA = vB = vC = bA = bB = bC = .1 und einen Lernpfadparameter (*a*-Parameter) von a1 = a2 = .30 zugrunde legt. Alle *p*-Parameter (bedingte Wahrscheinlichkeiten des Verbleibens in einem Wissenszustand) wurden für Gruppe 1 auf p01 = pA1 = pB1 = pAB1 = .5 festgesetzt und für Gruppe 2 auf p02 = pA2 = pB2 = pAB2 = .3.

In der MDT-Datei müssen die Häufigkeiten aller acht Antwortmuster für die erste Gruppe in den Zeilen 1 bis 8 und für die zweiten Gruppe in den folgenden Zeilen 9 bis 16 aufgeführt werden. Die Reihenfolge der Antwortmuster ist pro Gruppe jeweils lexikographisch, dieses Mal jedoch beginnend von −−− (Zeile 1 bzw. 9) über −−+, −+−,−++ usw. bis hin zu +++ (Zeile 8 bzw. 16). Erneut bedeutet + eine korrekte und − eine inkorrekte Antwort auf das Item.

```
Ideale Daten zum Knowledge Space Modell aus Abschnitt 7.2.4'
1       309
2       35
3       144
4       24
5       86
6       18
7       101
8       83
9       190
10      24
11      118
12      35
13      69
14      30
15      124
16      211
===
```

10) *Datei „MARSHICK.EQN" zur gemeinsamen Analyse eines Drei-Alternativen- und zweier Ja-Nein-Quellengedächtnistests sowie die Datei „MARSHICK.MDT" mit den rekonstruierten Daten zu Experiment 1 und 2 von Marsh und Hicks (1998).*

Die folgende EQN-Datei ermöglicht die simultane Analyse eines gewöhnlichen Drei-Alternativen-Quellengedächtnistests (Antwortalternativen: Item stammt aus Quelle A, aus Quelle B, Item ist neu) und zweier Ja-Nein-Gedächtnistests (Stammt das Item aus Quelle A? Stammt das Item aus Quelle B?). Die Ereignisbäume 1 bis 3 entsprechen dem 2HTSM-Modell (Bayen et al., 1996), die Bäume 4 bis 6 dem Modell zum Ja-Nein-Test zu Quelle A und die Bäume 7 bis 9 dem Modell zum Ja-Nein-Test zu Quelle B wie in Abschnitt 8.1 erläutert. Die Bäume 1, 4 und 7 gehören zu Items aus Quelle A, 2, 5 und 8 zu Items aus Quelle B und 3, 6 und 9 zu neuen Items. Die Antwortkategorien 1, 4 und 7 entsprechen einer Quelle-A-Antwort, 2, 5 und 8 einer Quelle-B-Antwort und 3, 6 und 9 einer Neu-Antwort. Die folgenden geradzahligen Antwortkategorien (10, 12 usw.) gehören zu Ja-Antworten, die ungeradzahligen (11, 13 usw.) zu Nein-Antworten in den Ja-Nein-Tests. Die Modellparameter D1, D2 und D3 entsprechen den Itemgedächtnisparameter für Items aus Quelle A und Quelle B sowie der Distraktorentdeckungsrate im Drei-Alternativen-Test. D1a, D2a und D3a sowie D1b, D2b und D3b sind die korrespondierenden Parameter für die Ja-Nein-Tests zu Quelle A und Quelle B. Die Quellengedächtnisparameter zu den beiden Quellen und den drei Tests werden entsprechen mit d1 und d2, d1a und d2a sowie d1b und d2b notiert. werden in Kapitel 8.1 erläutert. Die restlichen Parameter sind testspezifische Antworttendenz- bzw. Rateparameter, die das Antwortverhalten für den Fall beschreiben, daß das Item- oder Quellengedächtnis versagt (vgl. Abschnitt 8.1)

```
42
1       1       D1*d1
1       1       D1*(1-d1)*a
1       2       D1*(1-d1)*(1-a)
1       1       (1-D1)*b*g
1       2       (1-D1)*b*(1-g)
1       3       (1-D1)*(1-b)
2       5       D2*d2
2       4       D2*(1-d2)*a
2       5       D2*(1-d2)*(1-a)
2       4       (1-D2)*b*g
2       5       (1-D2)*b*(1-g)
2       6       (1-D2)*(1-b)
3       9       D3
3       7       (1-D3)*b*g
3       8       (1-D3)*b*(1-g)
3       9       (1-D3)*(1-b)
4       10      D1a*d1a
4       10      D1a*(1-d1a)*aa
4       11      D1a*(1-d1a)*(1-aa)
4       10      (1-D1a)*ga
4       11      (1-D1a)*(1-ga)
```

```
5       13      D2a*d2a
5       12      D2a*(1-d2a)*aa
5       13      D2a*(1-d2a)*(1-aa)
5       12      (1-D2a)*ga
5       13      (1-D2a)*(1-ga)
6       15      D3a
6       14      (1-D3a)*ga
6       15      (1-D3a)*(1-ga)
7       17      D1b*d1b
7       16      D1b*(1-d1b)*ab
7       17      D1b*(1-d1b)*(1-ab)
7       16      (1-D1b)*gb
7       17      (1-D1b)*(1-gb)
8       18      D2b*d2b
8       18      D2b*(1-d2b)*ab
8       19      D2b*(1-d2b)*(1-ab)
8       18      (1-D2b)*gb
8       19      (1-D2b)*(1-gb)
9       21      D3b
9       20      (1-D3b)*gb
9       21      (1-D3b)*(1-gb)
```

Die Kategoriehäufigkeiten in der folgenden MDT-Datei wurden aus den Tabellen 1 (zu Experiment 1) und 2 (zu Experiment 2) bei Marsh und Hicks (1998, s. auch Tabelle 8.1 in der vorliegenden Arbeit) durch Multiplikation der dort angegebenen relativen Häufigkeiten mit der Vp-Anzahl und der Itemzahl pro Vp geschätzt. Die Nummern der Antwortkategorien (1 bis 16) wurden oben bereits erläutert.

```
Marsh und Hicks 1998, Exp. 1
1       700
2       158
3       475
4       251
5       937
6       132
7       264
8       92
9       964
10      492
11      289
12      265
13      515
14      257
15      523
16      148
17      632
18      647
19      133
20      62
21      718
===
Marsh und Hicks 1998, Exp. 2
1       720
2       204
3       276
4       240
5       708
6       252
7       156
8       180
9       876
10      448
11      252
12      189
13      511
14      147
```

```
15      553
16      224
17      476
18      490
19      210
20      266
21      434
===
```

11) *Modellbeschreibungsdateien „BEV2HTSM.EQN", „BEV1HTSM.EQN" und „BEVLTSM.EQN" und Antworthäufigkeitendatei „BEVRAHD.MDT" zur Analyse der Prozeßdissoziationsdaten von Buchner et al. (1995) mit Zwei-Hochschwellen,- Ein-Hochschwellen- und Niedrigschwellen-Modellen des Quellengedächtnisses*

Die in diesem Abschnitt aufgeführten EQN-Dateien beziehen sich auf die vollständigen Prozeßdissoziationsdaten von Buchner et al. (1995, Experiment 1 bis 3). In jedem der drei Experimente wurden Antworttendenzen in zwei Stufen manipuliert (laxes vs. strenges Antwortkriterium). Die ersten acht Ereignisbäume jedes der drei Modelle beziehen sich auf die erste Antworttendenzbedingung, die Ereignisbäume 9 bis 16 jeweils auf die zweite Antworttendenzbedingung. Alle Parameter der ersten Bedingung beginnen mit dem Buchstaben „a", alle Parameter der zweiten Bedingung mit dem Buchstaben „b". Innerhalb jeder Antworttendenzbedingung wird zunächst die Inklusionsbedingung (Bäume 1 bis 4 bzw. 9 bis 12), danach die Exklusionsbedingung modelliert (Bäume 5 bis 8 bzw. 13 bis 16). Der erste Baum in jeder Bedingungskombination betrifft jeweils gelesene Items aus Phase 1, der zweite Baum generierte Anagramm-Items aus Phase 1. Baum 3 repräsentiert gehörte Items aus Phase 2 des Experiments, Baum 4 neue Distraktoritems. Die Antwortkategorien innerhalb eines Baumes folgen immer dem gleichen Muster: Die erste Kategorie repräsentiert die „Ja"-Antworten, die zweite die „Nein"-Antworten unter der betreffenden Versuchsbedingung. Die Parameter mit den Endungen D1r, D1a, D2 und D3 repräsentieren die Itemrekognitions- bzw. Detektionswahrscheinlichkeiten gelesener, generierter, gehörter und neuer Items. Die Parameter mit den Endungen d1r, d1a und d2 repräsentieren die Wahrscheinlichkeit, die Quelle eines gelesenen, generierten bzw. gehörten Items zu erinnern, wenn es als alt rekogniziert wurde. Die Parameter mit den Endungen b, a und g repräsentieren Ratewahrscheinlichkeiten, wobei b für eine „Alt"-Entscheidung bei Rekognitionsunsicherheit steht, a für eine Phase-1-Entscheidung im Falle von Quellenunsicherheit mit Itemrekognition und g für eine Phase-1-Entscheidung im Falle von Quellenunsicherheit verbunden mit Rekognitionsunsicherheit.

Für das auf Inklusions-Exklusions-Daten adaptierte 2HTSM-Modell sehen die Modellgleichungen dann wie folgt aus:

```
68
1 1 aD1r
1 1 (1-aD1r)*ab
1 2 (1-aD1r)*(1-ab)
2 3 aD1a
2 3 (1-aD1a)*ab
2 4 (1-aD1a)*(1-ab)
3 5 aD2
3 5 (1-aD2)*ab
3 6 (1-aD2)*(1-ab)
4 7 (1-aD3)*ab
4 8 (1-aD3)*(1-ab)
4 8 aD3
5 10 aD1r*ad1r
5 10 aD1r*(1-ad1r)*aa
5 9 aD1r*(1-ad1r)*(1-aa)
5 10 (1-aD1r)*ab*ag
5 9 (1-aD1r)*ab*(1-ag)
5 10 (1-aD1r)*(1-ab)
6 12 aD1a*ad1a
6 12 aD1a*(1-ad1a)*aa
6 11 aD1a*(1-ad1a)*(1-aa)
6 12 (1-aD1a)*ab*ag
6 11 (1-aD1a)*ab*(1-ag)
6 12 (1-aD1a)*(1-ab)
7 13 aD2*ad2
7 14 aD2*(1-ad2)*aa
7 13 aD2*(1-ad2)*(1-aa)
7 14 (1-aD2)*ab*ag
7 13 (1-aD2)*ab*(1-ag)
7 14 (1-aD2)*(1-ab)
8 16 aD3
8 16 (1-aD3)*ab*ag
8 15 (1-aD3)*ab*(1-ag)
8 16 (1-aD3)*(1-ab)
9 17 bD1r
9 17 (1-bD1r)*bb
9 18 (1-bD1r)*(1-bb)
10 19 bD1a
10 19 (1-bD1a)*bb
10 20 (1-bD1a)*(1-bb)
11 21 bD2
11 21 (1-bD2)*bb
11 22 (1-bD2)*(1-bb)
12 23 (1-bD3)*bb
12 24 (1-bD3)*(1-bb)
12 24 bD3
13 26 bD1r*bd1r
13 26 bD1r*(1-bd1r)*ba
13 25 bD1r*(1-bd1r)*(1-ba)
13 26 (1-bD1r)*bb*bg
13 25 (1-bD1r)*bb*(1-bg)
13 26 (1-bD1r)*(1-bb)
14 28 bD1a*bd1a
14 28 bD1a*(1-bd1a)*ba
14 27 bD1a*(1-bd1a)*(1-ba)
14 28 (1-bD1a)*bb*bg
14 27 (1-bD1a)*bb*(1-bg)
14 28 (1-bD1a)*(1-bb)
15 29 bD2*bd2
15 30 bD2*(1-bd2)*ba
15 29 bD2*(1-bd2)*(1-ba)
15 30 (1-bD2)*bb*bg
15 29 (1-bD2)*bb*(1-bg)
15 30 (1-bD2)*(1-bb)
16 32 bD3
16 32 (1-bD3)*bb*bg
16 31 (1-bD3)*bb*(1-bg)
16 32 (1-bD3)*(1-bb)
```

Ganz analog ist die EQN-Datei zum 1HTSM-Modell aufgebaut. Sie ergibt sich aus der o.g. Datei, wenn man überall D3 = 0 setzt. Hierdurch fallen vier Äste fort.

```
64
1 1 aD1r
1 1 (1-aD1r)*ab
1 2 (1-aD1r)*(1-ab)
2 3 aD1a
2 3 (1-aD1a)*ab
2 4 (1-aD1a)*(1-ab)
3 5 aD2
3 5 (1-aD2)*ab
3 6 (1-aD2)*(1-ab)
4 7 ab
4 8 (1-ab)
5 10 aD1r*ad1r
5 10 aD1r*(1-ad1r)*aa
5 9 aD1r*(1-ad1r)*(1-aa)
5 10 (1-aD1r)*ab*ag
5 9 (1-aD1r)*ab*(1-ag)
5 10 (1-aD1r)*(1-ab)
6 12 aD1a*ad1a
6 12 aD1a*(1-ad1a)*aa
6 11 aD1a*(1-ad1a)*(1-aa)
6 12 (1-aD1a)*ab*ag
6 11 (1-aD1a)*ab*(1-ag)
6 12 (1-aD1a)*(1-ab)
7 13 aD2*ad2
7 14 aD2*(1-ad2)*aa
7 13 aD2*(1-ad2)*(1-aa)
7 14 (1-aD2)*ab*ag
7 13 (1-aD2)*ab*(1-ag)
7 14 (1-aD2)*(1-ab)
8 16 ab*ag
8 15 ab*(1-ag)
8 16 (1-ab)
9 17 bD1r
9 17 (1-bD1r)*bb
9 18 (1-bD1r)*(1-bb)
10 19 bD1a
10 19 (1-bD1a)*bb
10 20 (1-bD1a)*(1-bb)
11 21 bD2
11 21 (1-bD2)*bb
11 22 (1-bD2)*(1-bb)
12 23 bb
12 24 (1-bb)
13 26 bD1r*bd1r
13 26 bD1r*(1-bd1r)*ba
13 25 bD1r*(1-bd1r)*(1-ba)
13 26 (1-bD1r)*bb*bg
13 25 (1-bD1r)*bb*(1-bg)
13 26 (1-bD1r)*(1-bb)
14 28 bD1a*bd1a
14 28 bD1a*(1-bd1a)*ba
14 27 bD1a*(1-bd1a)*(1-ba)
14 28 (1-bD1a)*bb*bg
14 27 (1-bD1a)*bb*(1-bg)
14 28 (1-bD1a)*(1-bb)
15 29 bD2*bd2
15 30 bD2*(1-bd2)*ba
15 29 bD2*(1-bd2)*(1-ba)
15 30 (1-bD2)*bb*bg
15 29 (1-bD2)*bb*(1-bg)
15 30 (1-bD2)*(1-bb)
16 32 bb*bg
16 31 bb*(1-bg)
16 32 (1-bb)
```

Die EQN-Datei zum Niedrigschwellen-Modell, das analog zu dem von Batchelder et al. (1994) vorgeschlagenen Modell konstruiert wurde, sieht wie folgt aus:

```
46
1 1 aD1r
1 2 (1-aD1r)
2 3 aD1a
2 4 (1-aD1a)
3 5 aD2
3 6 (1-aD2)
4 7 ab
4 8 (1-ab)
5 10 aD1r*ad1r
5 10 aD1r*(1-ad1r)*aa
5 9 aD1r*(1-ad1r)*(1-aa)
5 10 (1-aD1r)
6 12 aD1a*ad1a
6 12 aD1a*(1-ad1a)*aa
6 11 aD1a*(1-ad1a)*(1-aa)
6 12 (1-aD1a)
7 13 aD2*ad2
7 14 aD2*(1-ad2)*aa
7 13 aD2*(1-ad2)*(1-aa)
7 14 (1-aD2)
8 16 ab*ag
8 15 ab*(1-ag)
8 16 (1-ab)
9 17 bD1r
9 18 (1-bD1r)
10 19 bD1a
10 20 (1-bD1a)
11 21 bD2
11 22 (1-bD2)
12 23 bb
12 24 (1-bb)
13 26 bD1r*bd1r
13 26 bD1r*(1-bd1r)*ba
13 25 bD1r*(1-bd1r)*(1-ba)
13 26 (1-bD1r)
14 28 bD1a*bd1a
14 28 bD1a*(1-bd1a)*ba
14 27 bD1a*(1-bd1a)*(1-ba)
14 28 (1-bD1a)
15 29 bD2*bd2
15 30 bD2*(1-bd2)*ba
15 29 bD2*(1-bd2)*(1-ba)
15 30 (1-bD2)
16 32 bb*bg
16 31 bb*(1-bg)
16 32 (1-bb)
```

Die Kategoriehäufigkeiten der drei Experimente von Buchner et al. (1995) findet man in der folgenden MDT-Datei. Die hier angegebenen Häufigkeiten entsprechen den in Tabelle A1 bei Buchner et al. (1995) angegebenen Häufigkeiten, sind hier aber hinsichtlich der gehörten Items aus Phase 2 vervollständigt.

```
E1 Read Anagramm Phase 2 DISTRACTOR SET-Manipulation
1 285
2 115
3 354
4 46
5 360
6 40
7 121
8 279
9 77
```

10 323
11 33
12 367
13 242
14 158
15 33
16 367
17 238
18 162
19 329
20 71
21 290
22 110
23 99
24 301
25 71
26 329
27 35
28 365
29 253
30 147
31 20
32 380
==========
E2 Read Anagram Phase 2 Items AM-Manipulation
1 276
2 124
3 340
4 60
5 327
6 73
7 101
8 299
9 70
10 330
11 30
12 370
13 229
14 171
15 13
16 387
17 301
18 99
19 370
20 30
21 358
22 42
23 142
24 258
25 111
26 289
27 69
28 331
29 273
30 127
31 61
32 339
==========
E3 Read Anagramm Phase 2 Items INSTRUKTION-Manipulation
1 233
2 167
3 313
4 87
5 312
6 88
7 66
8 334
9 106
10 294
11 72
12 328
13 282

```
14 118
15 32
16 368
17 292
18 108
19 344
20 56
21 348
22 52
23 137
24 263
25 74
26 326
27 59
28 341
29 248
30 152
31 21
32 379
================================================
```

12) *Modellbeschreibungsdatei „RKPR2GR.EQN" und Antworthäufigkeitendatei „RKPR2GR.MDT" zur Analyse der Daten von Strack und Förster (1995), Hirshman und Henzler (1988), Hicks und Marsh (1998) sowie Erdfelder (1999) mit dem diskreten Zustandsmodell zur einfachen RK-Prozedur (vgl. Abschnitt 9.1 und dessen Unterabschnitte).*

Die folgende EQN-Datei entspricht dem diskreten Zustandsmodell zur einfachen RK-Prozedur wie es in Abschnitt 9.1.2 beschrieben wurde. In der hier spezifizierten Form ist die Datei zur simultanen Analyse zweier Versuchsbedingungen (z.B. unterschiedliche Antworttendenzbedingungen) geeignet. Die Ereignisbäume 1 und 2 repräsentieren die Zielitems und Distraktoren der ersten Versuchsbedingung, die Ereignisbäume 3 und 4 entsprechend die Zielitems und Distraktoren der zweiten Versuchsbedingung. Die Beobachtungskategorien sind immer in der Reihenfolgen „remember" (R), „know" (K) und „new" (N) durchnumeriert, d.h. die Kategorien 1, 4, 7 und 10 entsprechend einer R-Antwort, 2, 5, 8 und 11 einer K-Antwort und 3, 6, 9 und 12 einer N-Antwort. Der Parameter r steht für die Wahrscheinlichkeit einer bewußten Erinnerung, k entsprechend für die Wahrscheinlichkeit eines „Alt"-Urteils aufgrund von Vertrautheit. Die Parameter gr und gk sind die bedingten Ratewahrscheinlichkeiten für die R- und die K-Antwortkategorien, d ist die Distraktorentdeckungswahrscheinlichkeit. Die Ziffern hinter der Parameterbezeichnung geben jeweils die Versuchsbedingung an.

```
18
1 1 r1
1 2 (1-r1)*k1
1 1 (1-r1)*(1-k1)*gr1
1 2 (1-r1)*(1-k1)*(1-gr1)*gk1
1 3 (1-r1)*(1-k1)*(1-gr1)*(1-gk1)
2 6 d1
2 4 (1-d1)*gr1
2 5 (1-d1)*(1-gr1)*gk1
```

```
2 6 (1-d1)*(1-gr1)*(1-gk1)
3 7 r2
3 8 (1-r2)*k2
3 7 (1-r2)*(1-k2)*gr2
3 8 (1-r2)*(1-k2)*(1-gr2)*gk2
3 9 (1-r2)*(1-k2)*(1-gr2)*(1-gk2)
4 12 d2
4 10 (1-d2)*gr2
4 11 (1-d2)*(1-gr2)*gk2
4 12 (1-d2)*(1-gr2)*(1-gk2)
```

Die Daten, anhand derer das diskrete Zustandsmodell und das Kontinuum-Modell in Abschnitt 9.1.1.3 bzw. 9.1.2 evaluiert wurden, sind in der folgenden Häufigkeitendatei „RKPR2GR.MDT" zusammengefaßt. Die Nummern der Ereigniskategorien wurden so gewählt, daß sie zu der obigen EQN-Datei passen. Jeder Datensatz umfaßt zwei Versuchsbedingungen, die sich in bezug auf induzierte Antworttendenzen unterscheiden: Die ersten sechs Häufigkeiten repräsentieren R-, K- und N-Antworten auf Zielitems und Distraktoren bei liberaler Antworttendenz, die folgenden sechs bei konservativer Antworttendenz. Jede der vier Untersuchungen von Strack und Förster (1995), Hirshman und Henzler (1998), Hicks und Marsh (1999) und Erdfelder (1999) wurde in zwei Datensätze aufgesplittet. Der erste Datensatz repräsentiert jeweils ungünstige Enkodierungs- bzw. Abrufbedingungen (häufige Worte, kurze Darbietungszeiten, „flache" Enkodierung), der zweite Datensatz günstige Bedingungen (seltene Worte, lange Darbietungszeiten, „tiefe" Enkodierung).

```
Remember-Know Daten von Strack und Förster (1995) Häufige Items
1       76
2       96
3       103
4       6
5       66
6       203
7       67
8      73
9      160
10      6
11      22
12      272
=================
Remember-Know Daten von Strack und Förster (1995) Seltene Items
1       127
2       84
3       64
4       8
5       61
6       206
7       114
8       84
9       102
10      10
11      37
12      253
=================
Remember-Know Daten von Hirshman & Henzler (1998) 500 msec
1       134
2       275
3       230
4       141
```

```
5       448
6       692
7       70
8       160
9       410
10      64
11      166
12      1050
=================
Remember-Know Daten von Hirshman & Henzler (1998) 2 s
1       224
2       269
3       147
4      141
5      448
6      692
7      173
8      211
9      256
10      64
11      166
12      1050
=================
Remember-Know Daten von Hicks & Marsh (1999) (1 sec)
1       1039
2       470
3       451
4       196
5       412
6       1352
7      804
8      588
9      568
10      98
11      314
12      1548
=================
Remember-Know Daten von Hicks & Marsh (1999) (4,5 sec)
1       1059
2       588
3       314
4      118
5      274
6      1568
7      960
8      588
9      412
10      39
11      196
12      1725
=================
Remember-Know Daten von Erdfelder (1999) (flache Verarbeitung)
1       166
2       184
3       421
4       279
5       344
6       920
7      141
8      162
9      514
10      220
11      282
12      1130
=================
Remember-Know Daten von Erdfelder (1999) (tiefe Verarbeitung)
1       325
2       146
3       295
4      279
5      344
6      920
```

```
7      325
8      133
9      365
10      220
11      282
12      1130
================
```

13) *Modellbeschreibungsdatei „RKGPR2GR.EQN" und Antworthäufigkeitendatei „RKGPR2GR.MDT" zur Analyse der Daten von Gardiner et al. (1997) und Erdfelder (1999) mit dem diskreten RKG-Zustandsmodell zur erweiterten RKG-Prozedur (vgl. Abschnitt 9.2.2).*

Eine Modellbeschreibungsdatei zum diskreten RKG-Kontinuum-Modell, das in Abschnitt 9.2.2. besprochen wird, erhält durch minimale Ergänzungen der EQN-Datei zum RKG-Zustandsmodell aus Abschnitt 12 des Anhangs. In der hier spezifizierten Form ist die Datei „RKGPR2GR.EQN" zur simultanen Analyse zweier Versuchsbedingungen geeignet. Die Ereignisbäume 1 und 2 repräsentieren die Zielitems und Distraktoren der ersten Versuchsbedingung, die Ereignisbäume 3 und 4 entsprechend die Zielitems und Distraktoren der zweiten Versuchsbedingung. Die Beobachtungskategorien sind immer in der Reihenfolge „remember" (R), „know" (K), „guess" (G) und „new" (N) durchnumeriert, d.h. die Kategorien 1, 5, 9 und 13 entsprechend einer R-Antwort, 2, 6, 10 und 14 einer K-Antwort, 3, 7, 11 und 15 einer G-Antwort und 4, 8, 12 und 16 einer N-Antwort. Die Parameter haben dieselbe Bedeutung wie im Falle des diskreten Zustandsmodells zur einfachen RK-Prozedur. Es kommt lediglich ein Parameter gg hinzu, der für die (bedingte) Ratewahrscheinlichkeit der G-Antwortkategorie steht. Die Ziffern hinter der Parameterbezeichnung geben wiederum die Versuchsbedingung an.

```
22
1 1 r1
1 2 (1-r1)*k1
1 1 (1-r1)*(1-k1)*gr1
1 2 (1-r1)*(1-k1)*(1-gr1)*gk1
1 3 (1-r1)*(1-k1)*(1-gr1)*(1-gk1)*gg1
1 4 (1-r1)*(1-k1)*(1-gr1)*(1-gk1)*(1-gg1)
2 8 d1
2 5  (1-d1)*gr1
2 6 (1-d1)*(1-gr1)*gk1
2 7 (1-d1)*(1-gr1)*(1-gk1)*gg1
2 8 (1-d1)*(1-gr1)*(1-gk1)*(1-gg1)
3 9 r2
3 10 (1-r2)*k2
3 9 (1-r2)*(1-k2)*gr2
3 10 (1-r2)*(1-k2)*(1-gr2)*gk2
3 11 (1-r2)*(1-k2)*(1-gr2)*(1-gk2)*gg2
3 12 (1-r2)*(1-k2)*(1-gr2)*(1-gk2)*(1-gg2)
4 16 d2
4 13 (1-d2)*gr2
4 14 (1-d2)*(1-gr2)*gk2
4 15 (1-d2)*(1-gr2)*(1-gk2)*gg2
4 16 (1-d2)*(1-gr2)*(1-gk2)*(1-gg2)
```

Die Evaluation des diskreten RKG-Zustandsmodells erfolgte in Abschnitt 9.2.2 anhand der folgenden Häufigkeitendatei „RKGPR2GR.MDT". Die Nummern der Ereigniskategorien wurden so gewählt, daß sie zu der obigen EQN-Datei passen. Jeder Datensatz umfaßt zwei Versuchsbedingungen, die sich in bezug auf induzierte Antworttendenzen unterscheiden: Die ersten acht Häufigkeiten repräsentieren R-, K-, G- und N-Antworten auf Zielitems und Distraktoren bei liberaler Antworttendenz, die folgenden sechs bei konservativer Antworttendenz. Jede der beiden Untersuchungen zur RKG-Prozedur von Gardiner et al. (1997) und Erdfelder (1999) wurde in zwei Datensätze aufgesplittet. Der erste Datensatz repräsentiert jeweils ungünstige Enkodierungs- bzw. Abrufbedingungen (häufige Worte bzw. „flache" Enkodierung), der zweite Datensatz günstige Bedingungen (seltene Worte bzw. „tiefe" Enkodierung).

```
Remember-Know-Guess-Daten von Gardiner 1997 (häufige Wörter)
1       173
2       127
3       69
4       207
5       17
6       52
7       104
8       403
9       150
10      121
11      69
12      236
13      12
14      63
15      58
16      443
================
Remember-Know-Guess-Daten von Gardiner 1997 (seltene Wörter)
1       236
2       132
3       52
4       156
5       12
6       52
7       63
8       449
9       213
10      127
11      52
12      184
13      17
14      69
15      63
16      427
================
Remember-Know-Guess-Daten von Erdfelder 1999 (flach verarbeitet)
1       184
2       120
3       142
4       450
5       348
6       198
7       265
8       977
9       95
10      58
```

```
11      54
12      437
13      160
14      95
15      80
16      956
================
Remember-Know-Guess-Daten von Erdfelder 1999 (tief verarbeitet)
1       409
2       83
3       99
4       304
5       348
6       198
7       265
8       977
9       233
10      72
11      46
12      291
13      160
14      95
15      80
16      956
================
```

14) *Modellbeschreibungsdatei „RKGESAMT.EQN" und Antworthäufigkeitendatei „RKGESAMT.MDT" zur simultanen Analyse der RK- und der RKG-Daten von Erdfelder (1999) mit dem diskreten Zustandsmodell und dem diskreten RKG-Zustandsmodell (vgl. Abschnitt 9.2.2).*

Die in den letzten beiden Abschnitten aufgeführten EQN- und MDT-Dateien eignen sich zur Analyse von RK- bzw. RKG-Datensätzen für zwei Versuchsbedingungen. Sollen alle Daten von Erdfelder (1999) aus Tabelle 9.4 und 9.8 simultan analysiert werden um beispielsweise Parameterübereinstimmung zwischen der RK- und der RKG-Prozedur zu prüfen, so muß die folgende EQN-Datei verwendet werden. Die Ereignisbäume 1 bis 6 repräsentieren die RKG-Häufigkeiten. Baum 1 bis 3 stehen für die Versuchsbedingungen mit liberaler Antworttendenz in der Reihenfolge tiefe verarbeitete alte Items, flach verarbeitete alte Items und Distraktoren. Baum 4 bis 6 sind die entsprechenden Bäume für die konservative Antworttendenzbedingung. Die Bäume 7 bis 12 repräsentieren die RK-Häufigkeiten in dergleichen Reihenfolge der Versuchsbedingungen. Die Ereigniskategorien innerhalb der Bäume sind immer in der Reihenfolge R, K, G und N bzw. R, K und N durchnumeriert. Parameterbezeichnung, die mit „a" beginnen, stehen für Parameter des RKG-Modells und Parameterbezeichnung, die mit „b" beginnen, analog für Parameter des RKG-Modells. Der letzte Buchstabe gibt an, ob sich der betreffende Parameter auf Versuchsbedingungen mit liberaler (a) oder konservativer (n) Antworttendenz bezieht. Die *r*- und die *k*-Parameterbezeichnungen beinhalten die Verarbeitungsbedingung als dritten Buchstaben: „t" steht für die tiefe Verarbeitungsbedingung, „f" für die flache.

Parametergleichheiten zwischen den Testprozeduren lassen sich also z.B. durch Gleichsetzungrestriktionen wie arfa = brfa, arta = brta usw. prüfen.

```
62
1 1 arfa
1 2 (1-arfa)*akfa
1 1 (1-arfa)*(1-akfa)*agra
1 2 (1-arfa)*(1-akfa)*(1-agra)*agka
1 3 (1-arfa)*(1-akfa)*(1-agra)*(1-agka)*agga
1 4 (1-arfa)*(1-akfa)*(1-agra)*(1-agka)*(1-agga)
2 5 arta
2 6 (1-arta)*akta
2 5 (1-arta)*(1-akta)*agra
2 6 (1-arta)*(1-akta)*(1-agra)*agka
2 7 (1-arta)*(1-akta)*(1-agra)*(1-agka)*agga
2 8 (1-arta)*(1-akta)*(1-agra)*(1-agka)*(1-agga)
3 12 ada
3 9 (1-ada)*agra
3 10 (1-ada)*(1-agra)*agka
3 11 (1-ada)*(1-agra)*(1-agka)*agga
3 12 (1-ada)*(1-agra)*(1-agka)*(1-agga)
4 13 arfn
4 14 (1-arfn)*akfn
4 13 (1-arfn)*(1-akfn)*agrn
4 14 (1-arfn)*(1-akfn)*(1-agrn)*agkn
4 15 (1-arfn)*(1-akfn)*(1-agrn)*(1-agkn)*aggn
4 16 (1-arfn)*(1-akfn)*(1-agrn)*(1-agkn)*(1-aggn)
5 17 artn
5 18 (1-artn)*aktn
5 17 (1-artn)*(1-aktn)*agrn
5 18 (1-artn)*(1-aktn)*(1-agrn)*agkn
5 19 (1-artn)*(1-aktn)*(1-agrn)*(1-agkn)*aggn
5 20 (1-artn)*(1-aktn)*(1-agrn)*(1-agkn)*(1-aggn)
6 24 adn
6 21 (1-adn)*agra
6 22 (1-adn)*(1-agra)*agkn
6 23 (1-adn)*(1-agra)*(1-agkn)*aggn
6 24 (1-adn)*(1-agra)*(1-agkn)*(1-aggn)
7 25 brfa
7 26 (1-brfa)*bkfa
7 25 (1-brfa)*(1-bkfa)*bgra
7 26 (1-brfa)*(1-bkfa)*(1-bgra)*bgka
7 27 (1-brfa)*(1-bkfa)*(1-bgra)*(1-bgka)
8 28 brta
8 29 (1-brta)*bkta
8 28 (1-brta)*(1-bkta)*bgra
8 29 (1-brta)*(1-bkta)*(1-bgra)*bgka
8 30 (1-brta)*(1-bkta)*(1-bgra)*(1-bgka)
9 33 bda
9 31 (1-bda)*bgra
9 32 (1-bda)*(1-bgra)*bgka
9 33 (1-bda)*(1-bgra)*(1-bgka)
10 34 brfn
10 35 (1-brfn)*bkfn
10 34 (1-brfn)*(1-bkfn)*bgrn
10 35 (1-brfn)*(1-bkfn)*(1-bgrn)*bgkn
10 36 (1-brfn)*(1-bkfn)*(1-bgrn)*(1-bgkn)
11 37 brtn
11 38 (1-brtn)*bktn
11 37 (1-brtn)*(1-bktn)*bgrn
11 38 (1-brtn)*(1-bktn)*(1-bgrn)*bgkn
11 39 (1-brtn)*(1-bktn)*(1-bgrn)*(1-bgkn)
12 42 bdn
12 40 (1-bdn)*bgra
12 41 (1-bdn)*(1-bgra)*bgkn
12 42 (1-bdn)*(1-bgra)*(1-bgkn)
```

Die Datendatei, die auf die obige EQN-Datei zugeschnitten ist, läßt sich aus den in den in letzten beiden Abschnitten angegebenen Häufigkeiten (bzw. aus Tabelle 9.4 und 9.8) leicht ableiten:

```
Remember Know Guess Daten von Erdfelder 1999 (RKG-flach-proalt-targets zuerst)
1       184
2       120
3       142
4       450
5       409
6       83
7       99
8       304
9       348
10      198
11      265
12      977
13      95
14      58
15      54
16      437
17      233
18      72
19      46
20      291
21      160
22      95
23      80
24      956
25      166
26      184
27      421
28      325
29      146
30      295
31      279
32      344
33      920
34      141
35      162
36      514
37      325
38      133
39      365
40      220
41      282
42      1130
===============
```

15) *Datei „WAGBOER1.EQN" zur „destructive-updating theory" aus Wagenaar und Boer (1987, Abbildung 2, S. 296).*

Das in Kapitel 10 angesprochene Modell besteht aus drei Bäumen mit insgesamt 35 Ästen für die Bedingungen „nachträgliche Falschinformation" (Baum 1), „konsistente Nachinformation" (Baum 2) und „neutrale Nachinformation" (Baum 3). Pro Baum sind die empirischen Kategorien in der Reihenfolge

— OI „Ampel" rekogniziert und OI-Detail „Ampelfarbe" rekogniziert (1);

— OI „Ampel" rekogniziert, OI-Detail „Ampelfarbe" nicht rekogniziert (2);

— OI „Ampel" nicht rekogniziert, OI-Detail „Ampelfarbe" rekogniziert (3);
— weder OI „Ampel" noch OI-Detail „Ampelfarbe" rekogniziert (4)
durchnumeriert worden, so daß sich insgesamt 3·4 = 12 Kategorien ergeben. Die vier Parameter *p*, *c*, *q* und *d* entsprechen den bei Wagenaar und Boer (1987) angegebenen Wahrscheinlichkeiten für die Enkodierung der OI „Ampel" (*p*), die Enkodierung des OI-Detail „Ampelfarbe" (*c*), die Enkodierung der NI (*q*) und einer Beeinträchtigung der OI-Gedächtnisspur durch falsche NI (*d*). Sowohl h_1 als auch h_2 sind Hilfsparameter für die Ratewahrscheinlichkeiten; sie sind bei der Modellanpassung auf $h_1 = .333$ und $h_2 = .5$ zu fixieren.

Es folgt der Inhalt der EQN-Datei:

```
35
1 3 p*c*q*d*h1
1 4 p*c*q*d*(1-h1)
1 1 p*c*q*(1-d)
1 1 p*c*(1-q)
1 3 p*(1-c)*q*d*h1
1 4 p*(1-c)*q*d*(1-h1)
1 1 p*(1-c)*q*(1-d)*h1
1 2 p*(1-c)*q*(1-d)*(1-h1)
1 1 p*(1-c)*(1-q)*h1
1 2 p*(1-c)*(1-q)*(1-h1)
1 3 (1-p)*q*h1
1 4 (1-p)*q*(1-h1)
1 1 (1-p)*(1-q)*h2*h1
1 2 (1-p)*(1-q)*h2*(1-h1)
1 3 (1-p)*(1-q)*(1-h2)*h1
1 4 (1-p)*(1-q)*(1-h2)*(1-h1)
2 5 p*c*q
2 5 p*c*(1-q)
2 5 p*(1-c)*q*h1
2 6 p*(1-c)*q*(1-h1)
2 5 p*(1-c)*(1-q)*h1
2 6 p*(1-c)*(1-q)*(1-h1)
2 5 (1-p)*q*h1
2 6 (1-p)*q*(1-h1)
2 5 (1-p)*(1-q)*h2*h1
2 6 (1-p)*(1-q)*h2*(1-h1)
2 7 (1-p)*(1-q)*(1-h2)*h1
2 8 (1-p)*(1-q)*(1-h2)*(1-h1)
3 9 p*c
3 9 p*(1-c)*h1
3 10 p*(1-c)*(1-h1)
3 9 (1-p)*h2*h1
3 10 (1-p)*h2*(1-h1)
3 11 (1-p)*(1-h2)*h1
3 12 (1-p)*(1-h2)*(1-h1)
```

16) *Datei „WAGBOER2.EQN" zur „coexistence theory" aus Wagenaar und Boer (1987, Abbildung 3, S. 298).*

Die folgende EQN-Datei bezieht sich auf die gleiche Baum- und Kategorienstruktur wie WAGBOER1.EQN. Die Parametrisierung entspricht hier jedoch der Koexistenztheorie, so daß der *d*-Parameter fortfällt und stattdessen der *s*-Parameter (Abrufbeeinträchtigung für die OI durch die NI) hinzukommt.

```
34
1 3 p*c*q*s
1 1 p*c*q*(1-s)
1 1 p*c*(1-q)
1 3 p*(1-c)*q*s*h1
1 4 p*(1-c)*q*s*(1-h1)
1 1 p*(1-c)*q*(1-s)*h1
1 2 p*(1-c)*q*(1-s)*(1-h1)
1 1 p*(1-c)*(1-q)*h1
1 2 p*(1-c)*(1-q)*(1-h1)
1 3 (1-p)*q*h1
1 4 (1-p)*q*(1-h1)
1 1 (1-p)*(1-q)*h2*h1
1 2 (1-p)*(1-q)*h2*(1-h1)
1 3 (1-p)*(1-q)*(1-h2)*h1
1 4 (1-p)*(1-q)*(1-h2)*(1-h1)
2 5 p*c*q
2 5 p*c*(1-q)
2 5 p*(1-c)*q*h1
2 6 p*(1-c)*q*(1-h1)
2 5 p*(1-c)*(1-q)*h1
2 6 p*(1-c)*(1-q)*(1-h1)
2 5 (1-p)*q*h1
2 6 (1-p)*q*(1-h1)
2 5 (1-p)*(1-q)*h2*h1
2 6 (1-p)*(1-q)*h2*(1-h1)
2 7 (1-p)*(1-q)*(1-h2)*h1
2 8 (1-p)*(1-q)*(1-h2)*(1-h1)
3 9 p*c
3 9 p*(1-c)*h1
3 10 p*(1-c)*(1-h1)
3 9 (1-p)*h2*h1
3 10 (1-p)*h2*(1-h1)
3 11 (1-p)*(1-h2)*h1
3 12 (1-p)*(1-h2)*(1-h1)
```

17) *Datei „WAGBOER3.EQN" zur „no-conflict theory" aus Wagenaar und Boer (1987, Abbildung 4, S. 299).*

Die folgende EQN-Datei beschreibt ein Submodell der beiden zuvor dargestellten Modelle. Man erhält die folgende EQN-Datei, wenn man in WAGBOER1.EQN $d = 0$ oder in WAGBOER2.EQN $s = 0$ setzt.

```
31
1 1 p*c*q
1 1 p*c*(1-q)
1 1 p*(1-c)*q*h1
1 2 p*(1-c)*q*(1-h1)
1 1 p*(1-c)*(1-q)*h1
1 2 p*(1-c)*(1-q)*(1-h1)
1 3 (1-p)*q*h1
1 4 (1-p)*q*(1-h1)
1 1 (1-p)*(1-q)*h2*h1
1 2 (1-p)*(1-q)*h2*(1-h1)
1 3 (1-p)*(1-q)*(1-h2)*h1
1 4 (1-p)*(1-q)*(1-h2)*(1-h1)
2 5 p*c*q
2 5 p*c*(1-q)
2 5 p*(1-c)*q*h1
2 6 p*(1-c)*q*(1-h1)
2 5 p*(1-c)*(1-q)*h1
2 6 p*(1-c)*(1-q)*(1-h1)
2 5 (1-p)*q*h1
```

```
2 6 (1-p)*q*(1-h1)
2 5 (1-p)*(1-q)*h2*h1
2 6 (1-p)*(1-q)*h2*(1-h1)
2 7 (1-p)*(1-q)*(1-h2)*h1
2 8 (1-p)*(1-q)*(1-h2)*(1-h1)
3 9 p*c
3 9 p*(1-c)*h1
3 10 p*(1-c)*(1-h1)
3 9 (1-p)*h2*h1
3 10 (1-p)*h2*(1-h1)
3 11 (1-p)*(1-h2)*h1
3 12 (1-p)*(1-h2)*(1-h1)
```

18) *Datei „WAGBOER.MDT" mit den Daten aus Wagenaar und Boer (1987, Tabelle 4, S. 303) mit der Kategoriereihenfolge gemäß WAGBOER1.EQN, WAGBOER2.EQN und WAGBOER3.EQN.*

Die folgende MDT-Datei kann mit einer der drei zuvor aufgeführten EQN-Dateien (WAGBOER1.EQN, WAGBOER2.EQN und WAGBOER3.EQN) analysiert werden. Es handelt sich um empirische Kategoriehäufigkeiten, die aus den in Tabelle 4 bei Wagenaar und Boer (1987, S. 303) angegebenen Prozentangaben und den dort ebenfalls angegebenen Stichprobenumfängen pro Baum des Modells errechnet wurden.

```
Wagenaar & Boer (1987) Daten (inkonsistent/konsistent/neutral)
1      103
2      55
3      40
4      52
5      78
6      70
7      7
8      15
9      63
10     45
11     13
12     21
==================================
```

19) *Datei „LMBOUR9G.EQN" zur Analyse von Experiment 3 aus Loftus, Miller und Burns (1978).*

Diese EQN-Datei beschreibt das in Abschnitt 2.6 vorgestellte Modell zum 2AFC-Rekognitionstest von Loftus, Miller und Burns (1978) für insgesamt neun Versuchsgruppen bzw. -bedingungen. Es umfaßt insgesamt 18 Bäume. Die ungeraden Baumnummern stehen jeweils für die Experimentalbedingungen mit falscher Nachinformation, die geraden Baumnummern für Kontrollbedingungen mit neutraler (oder keiner) Nachinformation. Jeder Baum umfaßt zwei Kategorien, nämlich die Wahl der richtigen (OI-konformen) Alternative im 2AFC-Test (ungerade

Kategorienummern) und die Wahl der falschen (NI-konformen) Alternative. Insgesamt bezieht sich das Modell also auf 2·18 = 36 Kategorien. Für jede der neun Versuchsbedingungen ist ein *r*-Parameter, ein *b*-Parameter und ein *m*-Parameter vorgesehen. Parameter *g* ist ein globaler Hilfsparameter, der bei der Modellanpassung auf *g* = 0.5 fixiert werden muß.

```
72
1 1 r1*(1-m1)
1 1 r1*m1*(1-b1)
1 2 r1*m1*b1
1 1 (1-r1)*(1-b1)
1 2 (1-r1)*b1
2 3 r1
2 3 (1-r1)*g
2 4 (1-r1)*(1-g)
3 5 r2*(1-m2)
3 5 r2*m2*(1-b2)
3 6 r2*m2*b2
3 5 (1-r2)*(1-b2)
3 6 (1-r2)*b2
4 7 r2
4 7 (1-r2)*g
4 8 (1-r2)*(1-g)
5 9 r3*(1-m3)
5 9 r3*m3*(1-b3)
5 10 r3*m3*b3
5 9 (1-r3)*(1-b3)
5 10 (1-r3)*b3
6 11 r3
6 11 (1-r3)*g
6 12 (1-r3)*(1-g)
7 13 r4*(1-m4)
7 13 r4*m4*(1-b4)
7 14 r4*m4*b4
7 13 (1-r4)*(1-b4)
7 14 (1-r4)*b4
8 15 r4
8 15 (1-r4)*g
8 16 (1-r4)*(1-g)
9 17 r5*(1-m5)
9 17 r5*m5*(1-b5)
9 18 r5*m5*b5
9 17 (1-r5)*(1-b5)
9 18 (1-r5)*b5
10 19 r5
10 19 (1-r5)*g
10 20 (1-r5)*(1-g)
11 21 r6*(1-m6)
11 21 r6*m6*(1-b6)
11 22 r6*m6*b6
11 21 (1-r6)*(1-b6)
11 22 (1-r6)*b6
12 23 r6
12 23 (1-r6)*g
12 24 (1-r6)*(1-g)
13 25 r7*(1-m7)
13 25 r7*m7*(1-b7)
13 26 r7*m7*b7
13 25 (1-r7)*(1-b7)
13 26 (1-r7)*b7
14 27 r7
14 27 (1-r7)*g
14 28 (1-r7)*(1-g)
15 29 r8*(1-m8)
15 29 r8*m8*(1-b8)
15 30 r8*m8*b8
15 29 (1-r8)*(1-b8)
```

```
15 30 (1-r8)*b8
16 31 r8
16 31 (1-r8)*g
16 32 (1-r8)*(1-g)
17 33 r9*(1-m9)
17 33 r9*m9*(1-b9)
17 34 r9*m9*b9
17 33 (1-r9)*(1-b9)
17 34 (1-r9)*b9
18 35 r9
18 35 (1-r9)*g
18 36 (1-r9)*(1-g)
```

20) *Datei „LMBMUR9G.EQN" zur Analyse von Experiment 3 aus Loftus, Miller und Burns (1978).*

Die folgende EQN-Datei ist mit der im letzten Abschnitt angegebenen Datei identisch, abgesehen von der Ungleichheitsrestriktion $b_5 \geq .5$ in Versuchsgruppe 5. Zu diesem Zweck wurde statt b_5 der Hilfsparameter h_5 in das Modell aufgenommen, aus dem b_5 gemäß $b_{(1\ \text{Woche})} = .5 + .5 \cdot h_5$ errechnet werden kann.

```
74
1 1 r1*(1-m1)
1 1 r1*m1*(1-b1)
1 2 r1*m1*b1
1 1 (1-r1)*(1-b1)
1 2 (1-r1)*b1
2 3 r1
2 3 (1-r1)*g
2 4 (1-r1)*(1-g)
3 5 r2*(1-m2)
3 5 r2*m2*(1-b2)
3 6 r2*m2*b2
3 5 (1-r2)*(1-b2)
3 6 (1-r2)*b2
4 7 r2
4 7 (1-r2)*g
4 8 (1-r2)*(1-g)
5 9 r3*(1-m3)
5 9 r3*m3*(1-b3)
5 10 r3*m3*b3
5 9 (1-r3)*(1-b3)
5 10 (1-r3)*b3
6 11 r3
6 11 (1-r3)*g
6 12 (1-r3)*(1-g)
7 13 r4*(1-m4)
7 13 r4*m4*(1-b4)
7 14 r4*m4*b4
7 13 (1-r4)*(1-b4)
7 14 (1-r4)*b4
8 15 r4
8 15 (1-r4)*g
8 16 (1-r4)*(1-g)
9 17 r5*(1-m5)
9 18 r5*m5*g
9 17 r5*m5*(1-g)*(1-h5)
9 18 r5*m5*(1-g)*h5
9 18 (1-r5)*g
9 17 (1-r5)*(1-g)*(1-h5)
9 18 (1-r5)*(1-g)*h5
10 19 r5
10 19 (1-r5)*g
10 20 (1-r5)*(1-g)
```

```
11 21 r6*(1-m6)
11 21 r6*m6*(1-b6)
11 22 r6*m6*b6
11 21 (1-r6)*(1-b6)
11 22 (1-r6)*b6
12 23 r6
12 23 (1-r6)*g
12 24 (1-r6)*(1-g)
13 25 r7*(1-m7)
13 25 r7*m7*(1-b7)
13 26 r7*m7*b7
13 25 (1-r7)*(1-b7)
13 26 (1-r7)*b7
14 27 r7
14 27 (1-r7)*g
14 28 (1-r7)*(1-g)
15 29 r8*(1-m8)
15 29 r8*m8*(1-b8)
15 30 r8*m8*b8
15 29 (1-r8)*(1-b8)
15 30 (1-r8)*b8
16 31 r8
16 31 (1-r8)*g
16 32 (1-r8)*(1-g)
17 33 r9*(1-m9)
17 33 r9*m9*(1-b9)
17 34 r9*m9*b9
17 33 (1-r9)*(1-b9)
17 34 (1-r9)*b9
18 35 r9
18 35 (1-r9)*g
18 36 (1-r9)*(1-g)
```

21) *Datei „LMB789G.MDT" mit den Ergebnissen von Experiment 3 aus Loftus, Miller und Burns (1978).*

Diese MDT-Datei enthält die aus Tabelle 8.1 (Abschnitt 8.1) geschätzten absoluten Häufigkeiten für richtige und falsche Antworten in Experiment 3 von Loftus, Miller und Burns (1978). Die in Tabelle 8.1 angegebenen relativen Häufigkeiten wurden mit 24 (Stichprobenumfang) multipliziert, und das Ergebnis wurde auf die nächste natürliche Zahl gerundet.
Die Kategoriennumerierung ist auf die in Abschnitt 6 und 7 angegebenen EQN-Dateien abgestimmt. Die Kategorien 1 bis 20 stehen für die fünf Versuchsbedingungen mit unmittelbarer NI (Retentionsintervalle in der Reihenfolge 0 Minuten bis 1 Woche), die Kategorien 21 bis 38 für die vier Versuchsbedingungen mit verzögerter NI (Retentionsintervalle in der Reihenfolge 20 Minuten bis eine Woche).

```
Loftus Miller Burns (1978, Experiment 3)
1 9
2 15
3 21
4 3
5 10
6 14
7 19
8 5
9 11
10 13
11 17
```

```
12 7
13 10
14 14
15 14
16 10
17 13
18 11
19 12
20 12
21 10
22 14
23 14
24 10
25 9
26 15
27 15
28 9
29 7
30 17
31 11
32 13
33 4
34 20
35 12
36 12
==========================================
```

22) *Datei „FI1GR.EQN" zum Falschinformationsmodell für Quellendiskriminationstests aus Abschnitt 10.2. und die Daten aus Experiment 1 (FI-EXP1.MDT)*

Die folgende Datei „FI1GR.EQN" beschreibt das Falschinformationsmodell für Quellendiskriminationstests in der in Abschnitt 10.2 behandelten Form. Das Modell umfaßt insgesamt 49 Äste und neun Ereignisbäume, die hier in derselben Reihenfolge wie in Tabelle 10.4 und Abbildung 10.4a) bis 10.4d) durchnumeriert wurden. Die ersten fünf Ereignisbäume korrespondieren alten Items aus einer oder beiden Quellen (OI = Originalinformation oder NI = Nachinformation), die Ereignisbäume 6 bis 9 den vier verschiedenen Typen von Distraktoritems (vgl. Tabelle 10.4).

Die vier Antwortkategorien pro Bedingung wurden pro Baum in der Reihenfolge „OI-Antwort", „NI-Antwort", „OI&NI-Antwort" und „Neu-Antwort" durchnumeriert. Die Parameterbezeichnungen entsprechenden weitgehend denen in Abschnitt 10.2: Gedächtnisparameter beginnen mit dem Buchstaben „r", Urteils- bzw. Rateparameter mit „g". Danach folgt die Kennzeichnung der Quelle (OI bzw. NI) und am Schluß ein Buchstabe, der die Beziehung zwischen OI und NI (bei Gedächtnisparametern) bzw. die Beziehung zwischen Testitem und faktischen Item in der betreffenden Quelle (bei Rateparametern) angibt: „k" steht für konsistent, „w" für widersprüchlich und „n" für neutral (bei Gedächtnisparametern) bzw. nicht widersprüchlich (bei Rateparametern). Man beachte, daß die Rateparameter gOIn, gNIn1 und gNIn2 in Kapitel 10 mit $g(\mathrm{OI})_{\mathrm{nw}}$, $g_1(\mathrm{NI})_{\mathrm{nw}}$ und $g_2(\mathrm{NI})_{\mathrm{nw}}$ notiert

werden; analog werden gOIw, gNIw1 und gNIw2 in Kapitel 10 mit $g(\mathrm{OI})_\mathrm{w}$, $g_1(\mathrm{NI})_\mathrm{w}$ und $g_2(\mathrm{NI})_\mathrm{w}$ bezeichnet.

```
49
1  3 rOIk*rNIk
1  3 rOIk*(1-rNIk)*gNIn1
1  1 rOIk*(1-rNIk)*(1-gNIn1)
1  3 (1-rOIk)*rNIk*gOIn
1  2 (1-rOIk)*rNIk*(1-gOIn)
1  3 (1-rOIk)*(1-rNIk)*gOIn*gNIn1
1  1 (1-rOIk)*(1-rNIk)*gOIn*(1-gNIn1)
1  2 (1-rOIk)*(1-rNIk)*(1-gOIn)*gNIn2
1  4 (1-rOIk)*(1-rNIk)*(1-gOIn)*(1-gNIn2)
2  7 rOIn*gNIn1
2  5 rOIn*(1-gNIn1)
2  7 (1-rOIn)*gOIn*gNIn1
2  5 (1-rOIn)*gOIn*(1-gNIn1)
2  6 (1-rOIn)*(1-gOIn)*gNIn2
2  8 (1-rOIn)*(1-gOIn)*(1-gNIn2)
3 11 rOIw*gNIw1
3  9 rOIw*(1-gNIw1)
3 11 (1-rOIw)*gOIn*gNIw1
3  9 (1-rOIw)*gOIn*(1-gNIw1)
3 10 (1-rOIw)*(1-gOIn)*gNIw2
3 12 (1-rOIw)*(1-gOIn)*(1-gNIw2)
4 15 rNIn*gOIn
4 14 rNIn*(1-gOIn)
4 15 (1-rNIn)*gOIn*gNIn1
4 13 (1-rNIn)*gOIn*(1-gNIn1)
4 14 (1-rNIn)*(1-gOIn)*gNIn2
4 16 (1-rNIn)*(1-gOIn)*(1-gNIn2)
5 19 rNIw*gOIw
5 18 rNIw*(1-gOIw)
5 19 (1-rNIw)*gOIw*gNIn1
5 17 (1-rNIw)*gOIw*(1-gNIn1)
5 18 (1-rNIw)*(1-gOIw)*gNIn2
5 20 (1-rNIw)*(1-gOIw)*(1-gNIn2)
6 23 gOIn*gNIn1
6 21 gOIn*(1-gNIn1)
6 22 (1-gOIn)*gNIn2
6 24 (1-gOIn)*(1-gNIn2)
7 27 gOIn*gNIw1
7 25 gOIn*(1-gNIw1)
7 26 (1-gOIn)*gNIw2
7 28 (1-gOIn)*(1-gNIw2)
8 31 gOIw*gNIn1
8 29 gOIw*(1-gNIn1)
8 30 (1-gOIw)*gNIn2
8 32 (1-gOIw)*(1-gNIn2)
9 35 gOIw*gNIw1
9 33 gOIw*(1-gNIw1)
9 34 (1-gOIw)*gNIw2
9 36 (1-gOIw)*(1-gNIw2)
```

Die folgende Datei „FI-EXP1.MDT” enthält die Antworthäufigkeiten zu Experiment 1 aus Abschnitt 10.2.1 in der Reihenfolge, die die obige EQN-Datei verlangt. Die Häufigkeiten decken sich mit den Angaben in Tabelle 10.5.

```
Experiment 1 zum FI-Modell, SS92
1     10
2     22
3     41
4      9
5     46
6      2
```

```
7      3
8     31
9     36
10     6
11     6
12    34
13     4
14    59
15     4
16    15
17     6
18    43
19    17
20    16
21     4
22     4
23     0
24    74
25     4
26    14
27     0
28    64
29    16
30     2
31     4
32    60
33     7
34    10
35     7
36    58
================
```

23) *Datei „FI2GR.EQN" zum Falschinformationsmodell für Quellendiskriminationstests aus den Abschnitten10.2.2 und 10.2.3und die zugehörige Datendatei en(FI-EXP2.MDT und FI-EXP3.MDT)*

Die folgende Modellbeschreibungsdatei „FI2GR.EQN" erlaubt Analysen zum Falschinformationsmodell aus Abschnitt 10.2 für zwei verschiedene experimentelle Bedingungen oder Probandengruppen. Sie bezieht sich folglich auf Häufigkeitsdaten für alle neun Versuchsbedingungen des FI-Modells, erhoben für zwei Gruppen A und B von Versuchspersonen, insgesamt also $9 \cdot 4 \cdot 2 = 72$ Häufigkeiten. Die Ereignisbäume 1 bis 9 des folgenden Modells repräsentieren Gruppe A, die Bäume 10 bis 18 Gruppe B. Die ersten 49 Äste sind mit denen der Datei „FI1GR.EQN" aus dem letzten Abschnitt identisch. Allen Parameterbezeichnungen wurde hier der Buchstabe „a" vorangestellt, um deutlich zu machen, daß sich diese Parameter auf die Vpn-Gruppe A beziehen. Die folgenden Äste 50 bis 98 stellen dasselbe Modell für die Probanden der Gruppe B dar, weshalb hier allen Parameterbezeichnungen ein „b" vorangestellt ist. Durch Gleichsetzungsrestriktionen können dann Nullhypothesen der Gleichheit korrespondierender Parameter in Gruppe A und Gruppe B geprüft werden (vgl. dazu die Abschnitte 10.2.2 und 10.2.3).

```
98
1  3 arOIk*arNIk
1  3 arOIk*(1-arNIk)*agNIn1
1  1 arOIk*(1-arNIk)*(1-agNIn1)
1  3 (1-arOIk)*arNIk*agOIn
1  2 (1-arOIk)*arNIk*(1-agOIn)
1  3 (1-arOIk)*(1-arNIk)*agOIn*agNIn1
1  1 (1-arOIk)*(1-arNIk)*agOIn*(1-agNIn1)
1  2 (1-arOIk)*(1-arNIk)*(1-agOIn)*agNIn2
1  4 (1-arOIk)*(1-arNIk)*(1-agOIn)*(1-agNIn2)
2  7 arOIn*agNIn1
2  5 arOIn*(1-agNIn1)
2  7 (1-arOIn)*agOIn*agNIn1
2  5 (1-arOIn)*agOIn*(1-agNIn1)
2  6 (1-arOIn)*(1-agOIn)*agNIn2
2  8 (1-arOIn)*(1-agOIn)*(1-agNIn2)
3 11 arOIw*agNIw1
3  9 arOIw*(1-agNIw1)
3 11 (1-arOIw)*agOIn*agNIw1
3  9 (1-arOIw)*agOIn*(1-agNIw1)
3 10 (1-arOIw)*(1-agOIn)*agNIw2
3 12 (1-arOIw)*(1-agOIn)*(1-agNIw2)
4 15 arNIn*agOIn
4 14 arNIn*(1-agOIn)
4 15 (1-arNIn)*agOIn*agNIn1
4 13 (1-arNIn)*agOIn*(1-agNIn1)
4 14 (1-arNIn)*(1-agOIn)*agNIn2
4 16 (1-arNIn)*(1-agOIn)*(1-agNIn2)
5 19 arNIw*agOIw
5 18 arNIw*(1-agOIw)
5 19 (1-arNIw)*agOIw*agNIn1
5 17 (1-arNIw)*agOIw*(1-agNIn1)
5 18 (1-arNIw)*(1-agOIw)*agNIn2
5 20 (1-arNIw)*(1-agOIw)*(1-agNIn2)
6 23 agOIn*agNIn1
6 21 agOIn*(1-agNIn1)
6 22 (1-agOIn)*agNIn2
6 24 (1-agOIn)*(1-agNIn2)
7 27 agOIn*agNIw1
7 25 agOIn*(1-agNIw1)
7 26 (1-agOIn)*agNIw2
7 28 (1-agOIn)*(1-agNIw2)
8 31 agOIw*agNIn1
8 29 agOIw*(1-agNIn1)
8 30 (1-agOIw)*agNIn2
8 32 (1-agOIw)*(1-agNIn2)
9 35 agOIw*agNIw1
9 33 agOIw*(1-agNIw1)
9 34 (1-agOIw)*agNIw2
9 36 (1-agOIw)*(1-agNIw2)
10 39 brOIk*brNIk
10 39 brOIk*(1-brNIk)*bgNIn1
10 37 brOIk*(1-brNIk)*(1-bgNIn1)
10 39 (1-brOIk)*brNIk*bgOIn
10 38 (1-brOIk)*brNIk*(1-bgOIn)
10 39 (1-brOIk)*(1-brNIk)*bgOIn*bgNIn1
10 37 (1-brOIk)*(1-brNIk)*bgOIn*(1-bgNIn1)
10 38 (1-brOIk)*(1-brNIk)*(1-bgOIn)*bgNIn2
10 40 (1-brOIk)*(1-brNIk)*(1-bgOIn)*(1-bgNIn2)
11 43 brOIn*bgNIn1
11 41 brOIn*(1-bgNIn1)
11 43 (1-brOIn)*bgOIn*bgNIn1
11 41 (1-brOIn)*bgOIn*(1-bgNIn1)
11 42 (1-brOIn)*(1-bgOIn)*bgNIn2
11 44 (1-brOIn)*(1-bgOIn)*(1-bgNIn2)
12 47 brOIw*bgNIw1
12 45 brOIw*(1-bgNIw1)
12 47 (1-brOIw)*bgOIn*bgNIw1
12 45 (1-brOIw)*bgOIn*(1-bgNIw1)
12 46 (1-brOIw)*(1-bgOIn)*bgNIw2
12 48 (1-brOIw)*(1-bgOIn)*(1-bgNIw2)
```

```
13 51 brNIn*bgOIn
13 50 brNIn*(1-bgOIn)
13 51 (1-brNIn)*bgOIn*bgNIn1
13 49 (1-brNIn)*bgOIn*(1-bgNIn1)
13 50 (1-brNIn)*(1-bgOIn)*bgNIn2
13 52 (1-brNIn)*(1-bgOIn)*(1-bgNIn2)
14 55 brNIw*bgOIw
14 54 brNIw*(1-bgOIw)
14 55 (1-brNIw)*bgOIw*bgNIn1
14 53 (1-brNIw)*bgOIw*(1-bgNIn1)
14 54 (1-brNIw)*(1-bgOIw)*bgNIn2
14 56 (1-brNIw)*(1-bgOIw)*(1-bgNIn2)
15 59 bgOIn*bgNIn1
15 57 bgOIn*(1-bgNIn1)
15 58 (1-bgOIn)*bgNIn2
15 60 (1-bgOIn)*(1-bgNIn2)
16 63 bgOIn*bgNIw1
16 61 bgOIn*(1-bgNIw1)
16 62 (1-bgOIn)*bgNIw2
16 64 (1-bgOIn)*(1-bgNIw2)
17 67 bgOIw*bgNIn1
17 65 bgOIw*(1-bgNIn1)
17 66 (1-bgOIw)*bgNIn2
17 68 (1-bgOIw)*(1-bgNIn2)
18 71 bgOIw*bgNIw1
18 69 bgOIw*(1-bgNIw1)
18 70 (1-bgOIw)*bgNIw2
18 72 (1-bgOIw)*(1-bgNIw2)
```

Die folgende Datei „FI-EXP2.MDT” beinhaltet die Antworthäufigkeiten zu Experiment 2 aus Abschnitt 10.2.2 in der von obiger EQN-Datei geforderten Reihenfolge (vgl. auch Tabelle 10.6).

```
FI-Experiment 2 SS93 (erst Standardbedingung (bis Zeile 36), dann Warnbedingung)
1   16
2   21
3   35
4    8
5   37
6    4
7    8
8   31
9   41
10   3
11  12
12  22
13   2
14  54
15   4
16  19
17   5
18  54
19   8
20  11
21   3
22   7
23   1
24  69
25   7
26  15
27   1
28  57
29   8
30   7
31   3
32  62
33  10
34  10
```

```
35   1
36  58
37  13
38  14
39  27
40   4
41  29
42   3
43   1
44  25
45  25
46   1
47   5
48  27
49   1
50  45
51   2
52   9
53   3
54  36
55   9
56  10
57   1
58   2
59   1
60  53
61   2
62   9
63   2
64  44
65   5
66   1
67   1
68  51
69   8
70   6
71   3
72  40
=====================================================
```

Die folgende Datei „FI-EXP3.MDT” beinhaltet die Antworthäufigkeiten zu Experiment 3 aus Abschnitt 10.2.3 in der von „FI2GR.EQN” geforderten Reihenfolge (vgl. Tabelle 10.7).

```
FI-Experiment 3 SS94 (erst BILD-BILD bis Zeile 36, dann BILD-TEXT)
1     7
2     9
3    19
4    42
5    18
6     5
7     5
8    49
9    19
10    5
11    7
12   45
13    2
14   16
15   10
16   45
17    6
18   21
19    8
20   38
21    5
22    1
23    4
```

```
24   63
25    5
26   12
27    7
28   54
29    9
30    3
31    4
32   43
33    8
34    8
35    6
36   51
37    5
38   15
39   10
40   13
41   11
42    3
43    2
44   29
45    7
46    3
47    5
48   28
49    2
50   17
51    4
52   21
53    5
54   20
55    4
56   14
57    1
58    4
59    1
60   37
61    1
62   10
63    2
64   30
65    5
66    3
67    0
68   35
69    6
70    9
71    4
72   24
==================================
```