Zur Psychometrie der Mathematik am Ende der Sekundarstufe I

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften
der Universität Mannheim

vorgelegt von
Dipl.-Psych. Fabian Jasper

Universität Mannheim
Fakultät für Sozialwissenschaften

Oktober 2009

Dekan der Fakultät für Sozialwissenschaften:

Prof. Dr. Berthold Rittberger

Gutachter:

Prof. Dr. Werner W. Wittmann (Emeritus, Universität Mannheim)

Prof. Dr. Manfred Hofer (Emeritus, Universität Mannheim)

Tag der Disputation:

14.12.2009

Vorwort

Diese Arbeit widmet sich der Psychometrie der Mathematik aus den Blickwinkeln der psychologischen Diagnostik und – zu etwas geringerem Ausmaß – der pädagogischen Psychologie. Trotz zahlreicher Internationaler Vergleichsstudien in etwa 15 Jahren wurde die Frage nach dem Konstrukt der Mathematikfähigkeit nur selten gestellt. Etwas nachdenklich hat mich als Autor gemacht, dass auch für Mathematiktests, die ab den 90iger Jahren entwickelt wurden, entweder keinerlei Datensätze mehr vorhanden waren, oder man sich weigerte, sie zu wissenschaftlichen Zwecken zur Verfügung zu stellen. Dies erinnert ein wenig an das, zwar methodisch begründete, aber aus wissenschaftlicher Sicht fragliche Vorgehen bei PISA, selbst 2009, also bald 10 Jahre nach der ersten Untersuchungswelle, nicht alle ursprünglichen Aufgaben zu veröffentlichen. Meine Hoffnung ist mit dieser Arbeit die Forschung zur Psychometrie der Mathematik ein wenig voranzutreiben, damit sich vielleicht, wie in der Intelligenz- oder Persönlichkeitsdiagnostik, in einigen Jahren ein Grundmodell etabliert und so die Diagnostik in diesem Bereich objektiver wird.

Mein ausdrücklicher Dank gilt an dieser Stelle Prof. Werner W. Wittmann, von dem ich in Veranstaltungen und Gesprächen viel lernen konnte, und der es mir überhaupt erst ermöglicht hat, diese Arbeit zu schreiben. Gleiches gilt für meinem geschätzten Kollegen und früheren Lehrer Dr. Dietrich Wagener, der mir ermöglichte, in eine Zusammenarbeit mit dem Hogrefe-Verlag einzusteigen. Auch sei Prof. Manfred Hofer genannt, an dessen Lehrstuhl ich erste Erfahrungen im Bereich der Wissenschaft sammeln konnte und der bereit war, als Zweitkorrektor der Arbeit zu fungieren. An dieser Stelle muss auch Prof. Liepmann von der Uni Berlin genannt werden, der sich stets sehr kooperativ zeigte und auf meine Wünsche im Rahmen der Normierung des Tests einging. Auch Dr. Wolfgang Conrad möchte ich nicht unerwähnt lassen, der mir wertvolle Tipps gab, um diese Arbeit besser zu machen. Darüber hinaus möchte ich mich auch bei meinen Kollegen David Kriz und Andrés Steffanowski bedanken, die mich 2007 an die Arbeit in bestehenden Projekten herangeführt haben und von denen ich ebenfalls viel lernen konnte. Nicht zuletzt waren es auch die vielen Schüler, Studenten und natürlich Lehrer, die ihre Zeit geopfert haben, um dieses Projekt überhaupt erst möglich zu machen.

Natürlich möchte ich mich auch bei meiner Freundin Kristine bedanken, die mich in guten wie in schlechten Zeiten stets vollstens unterstützt hat.

Mannheim, den 24.10.2009 Fabian Jasper

INHALTSVERZEICHNIS

1 EINLEITUNG 15

1.1 PSYCHOMETRIE 16
1.2 LEITFRAGEN DIESER ARBEIT 16
1.3 BEDEUTUNG VON MULTIDIMENSIONALITÄT FÜR DIE PRAKTISCHE DIAGNOSTIK 17

2 AKTUELLER FORSCHUNGSSTAND 18

2.1 VERFÜGBARE KOMMERZIELLE MATHEMATIKTESTS 19
2.1.1 Analyse des Rechentest 8+ 19
2.1.2 Analyse des MTAS 20
2.1.3 Analyse des Berufsbezogenen Rechentests 21
2.1.4 Analyse des Rechentests 9+ 21
2.1.5 Analyse des Mathematiktest – Grundkenntnisse für Lehre und Beruf 22
2.1.6 Schlussfolgerung aus Sichtung aktueller Mathetests 23
2.2 MATHEMATIK IN INTERNATIONALEN VERGLEICHSSTUDIEN 24
2.2.1 Third International Mathematics and Science Study 24
2.2.2 Programme for International Student Assessment 26
2.2.3 Schlussfolgerung aus der Betrachtung der TIMSS und PISA-Studien für eine psychometrische Ordnung29

3 THEORETISCHE STRUKTURIERUNG VON MATHEMATIKFÄHIGKEIT 31

3.1 INTELLIGENZDIAGNOSTISCHE ÜBERLEGUNGEN ZUR ORDNUNG VON MATHEMATIK 31
3.1.1 Thurstones primary abilities 33
3.1.2 Cattells Theorie fluider und kristaliner Intelligenz 34
3.1.3 Jägers Facettenmodell 34
3.1.4 Die Zwei-Faktoren-Theorie und integrative Modelle 36
3.1.5 Verbreitete Intelligenztests 39
3.1.6 Schlussfolgerung aus Betrachtung von Intelligenztests und Konzepten: Skalenkonzeption 41
3.1.6.1 Verbale Mathematikfähigkeit: Mathematische Literalität 42
3.1.6.2 Figurale Mathematikfähigkeit: Geometrie und grafische Funktionen 43
3.1.6.3 Numerische Mathematikfähigkeit I: Prozedurales Rechnen 43
3.1.6.4 Numerische Mathematikfähigkeit II: Komplexes Rechnen 44
3.2 TAXONOMIEN ZUR ORDNUNG VON MATHEMATIKFÄHIGKEIT 44
3.2.1 Bloom et al. (1956) 45
3.2.1.1 Empirische Bewährung der Taxonomie 47
3.2.2 A revised taxonomy: Anderson und Krathwohl (2001) 48
3.2.2.1 Zur kognitiven Dimension 49
3.2.2.2 Zur Wissensdimension 51
3.2.2.3 Empirische Bewährung der Taxonomie 51
3.2.3 Wilson (1970) 52
3.2.4 Components Display Theory (CDT) 53
3.2.5 Ein integratives Modell 54
3.2.6 Schlussfolgerung 55
3.3 ERWEITERTE INTEGRATION: EIN KOGNITIVES PROZESS X INHALTE –MODELL 56

4 VORPRÜFUNGEN ZU DEN BISHERIGEN THEORETISCHEN ÜBERLEGUNGEN 58

4.1 HYPOTHESEN I 59
4.2 BESTIMMUNG DER N-DIMENSIONALITÄT EINES TESTS 59
4.2.1 Begriffklärung: Unidimensionalität 60
4.2.2 Antwortpattern 61
4.2.3 Reliabilität 63
4.2.4 Faktorenanalyse 64
4.2.5 Latent Trait Modell-Indizes 66
4.2.6 Nonlineare Faktorenanalyse 67

4.2.7 Die DIMTEST-Prozedur ... 69
4.2.8 Clusteranalyse ... 74
4.2.9 Schlussfolgerungen für diese Arbeit ... 74
4.3 Die Bedeutung der Itemschwierigkeit für Strukturanalysen ... 76
4.3.1 Parceling ... 77
4.3.2 Alternative SEM-Schätzverfahren ... 78
4.3.3 Law of diminishing returns ... 80
4.4 Reanalyse eines an der Uni Mannheim entwickelten Tests ... 81
4.4.1 Testaufbau ... 82
4.4.2 Klassische Itemkennwerte ... 83
4.4.2.1 DIMTEST und DETECT ... 84
4.4.2.2 HCA/CCPROX ... 86
4.4.2.3 NOHARM ... 88
4.5 Schlussfolgerungen ... 90

5 ERSTELLUNG EINER NEUEN TESTVORFORM ... 91

5.1 Geltungsbereich und Zielgruppe ... 91
5.2 Bildungsstandards und Lehrpläne ... 92
5.2.1 Bildungsstandards für den Hauptschulabschluss (Mathematik) ... 93
5.2.2 Bildungsstandards für den mittleren Schulabschluss (Mathematik) ... 95
5.2.3 Fazit zu den Bildungsstandards Mathematik ... 97
5.3 Exemplarische Betrachtung vorhandener Curricula ... 98
5.3.1 Lehrpläne Nordrhein-Westfalens ... 99
5.3.2 Lehrpläne Niedersachsens ... 100
5.3.3 Lehrpläne Baden-Württembergs ... 103
5.3.4 Lehrpläne Bayerns ... 105
5.3.5 Fazit zu den Lehrplänen ... 106
5.4 Technische Konstruktionsprinzipien ... 107
5.4.1 Item-Benennungen in dieser Arbeit ... 108
5.4.2 Antwortformat ... 108
5.5 Generierung der Testaufgaben ... 108
5.6 Zusammenstellung zweier Testvorformen ... 109

6 ZUSAMMENSTELLUNG DER ENDFORM ... 110

6.1 Stichprobe ... 110
6.2 Zusammenstellung der Skalen der Endform ... 111
6.2.1 Auswahl von Items für Geometrie und grafische Funktionen ... 111
6.2.2 Auswahl von Items für prozedurales Rechnen ... 113
6.2.3 Auswahl von Items für komplexes Rechnen ... 115
6.2.4 Auswahl von Items für mathematische Literalität ... 116
6.3 Weitere Veränderungen bis zur Endform ... 117

7 PASSUNG DER ENDFORM GEMÄß KLASSISCHER TESTTHEORIE ... 117

7.1 Wieso klassische Testtheorie? ... 117
7.2 Testanalyse ... 118
7.2.1 Stichprobe ... 119
7.2.2 Reliabilitätsschätzungen ... 119
7.2.3 Abschließende Itemselektionen ... 122
7.3 Hypothesen II ... 122
7.4 Konstruktvalidität des Mathematiktests ... 123
7.4.1 Zusammenhänge zwischen den Skalen der Endform ... 123
7.4.2 Zusammenhänge mit Trait-State-Angst ... 124
7.4.3 Verbale Intelligenz ... 126
7.4.4 Numerische Intelligenz ... 126
7.4.5 Schulnoten ... 128
7.5 Schlussfolgerungen ... 129

8 KONFIRMATORISCHE PRÜFUNG DER THEORETISCHEN ANNAHMEN ... 129

8.1 HYPOTHESEN III 129
8.1.1 N-Dimensionalität der Inhalte: DIMTEST – DETECT 130
8.2 STRUKTURANALYSEN DER INHALTSFACETTEN AUF ITEMEBENE 132
8.2.1 Faktorenanalyse 133
8.2.2 NOHARM 136
8.2.3 Allgemeine Schlussfolgerungen aus der NOHARM-Lösung 138
8.3 STRUKTURANALYSEN DER INHALTSFACETTEN AUF PARCEL-EBENE 138
8.3.1 Faktorenanalyse 139
8.3.2 Strukturgleichungsmodelle 141
8.3.2.1 Sinn und Nutzen von Cut-Off Kriterien 141
8.3.2.2 Modelle mit einem G-Faktor 144
8.3.2.3 Modell mit drei Inhaltsfaktoren 145
8.3.2.4 Modell mit 4 Inhaltsfaktoren 145
8.4 TAXONOMISCHE PASSUNG DER ENDFORM 147
8.4.1 Rekrutierung 147
8.4.2 Durchführung 148
8.4.3 Ergebnisse 148
8.4.3.1 Stichprobe 149
8.4.3.2 Zusammenhang von Einschätzung und Itemschwierigkeit 151
8.4.3.3 Bedeutung der 6 Taxonomiestufen 151
8.4.3.4 Rater-Übereinstimmung 152
8.4.3.5 Taxonomielevel des Mathematiktests 153
8.5 SCHLUSSFOLGERUNG 155

9 WEITERFÜHRENDE BETRACHTUNGEN 156

9.1 EIN SCHMID-LEIMAN MODELL 156
9.2 SCHMID-LEIMAN-MODELL VERSUS OBLIQUE-MODELLE 158
9.3 TRENNBARKEIT DER SKALEN PROZEDURALES- UND KOMPLEXES RECHNEN 162
9.3.1 Faktorenanalytisch 162
9.3.2 Diskrininanzanalyse 163
9.4 GESCHLECHTERUNTERSCHEIDE 165
9.5 PROFILDIAGNOSTIK IM EINZEL- UND GRUPPENFALL 167
9.6 MULTIDIMENSIONAL RANDOM COEFFICIENT MULTINOMIAL LOGIT MODEL 171
9.6.1 Das Rasch-Modell als Spezialfall 171
9.6.2 Within und between Item-Multidimensionalität 173
9.6.3 Modelltests 175
9.6.3.1 Conquest: 3- und 4 Faktormodelle 175
9.6.3.2 Conquest: 3- und 4 Faktor SL-Modelle 176
9.7 STRUKTURELLE TRENNBARKEIT DER TAXONOMIESTUFEN 177
9.8 ZUSAMMENHANG VON TAXONOMIELEVEL UND SKALENZUGEHÖRIGKEIT 180

10 GESAMTDISKUSSION UND AUSBLICK 182

11 LITERATUR 186

12 ANHANG 209

12.1 REANALYSE DES EXPRA-TESTS 209
12.1.1 Klassische Kennwerte aller Items 209
12.1.2 NOHARM Lösung 3-Faktoren, explorativ 210
12.2 ITEMBENNENUNGEN IN ALLEN TESTFORMEN 213
12.3 SPSS-SKRIPT ZUM VERGLEICH ABHÄNGIGER KORRELATIONEN 217
12.4 LADUNGEN EINER DREIFAKTORIELLEN MPLUS-ML LÖSUNG DER ENDFORM 218
12.5 4-FAKTORIELLE SL-LÖSUNG DER ENDFORM MIT WLSMV-SCHÄTZUNG 219
12.6 KENNWERTE FÜR DIE SKALEN DER VORFORM A, VOR JEGLICHER ITEMSELEKTION 220
12.6.1 Geometrie und grafische Funktionen 220
12.6.2 Komplexes Rechnen 220
12.6.3 Mathematische Literalität 221
12.6.4 Prozedurales Rechnen 221
12.7 KENNWERTE DER SKALEN DER VORFORM B, VOR JEGLICHER ITEMSELEKTION 223

12.7.1 Geometrie und Grafische Funktionen 223
12.7.2 Mathematische Literalität 224
12.7.3 Prozedurales Rechnen 224

ABBILDUNGSVERZEICHNIS

Abbildung 1 Ausmaß in dem die 4 Inhaltsdimensionen der TIMSS 2007-Untersuchung im Test enthalten sind. 25
Abbildung 2 Organisation der Mathematikdomäne in PISA 2003, nach OECD (2003, S. 28) 28
Abbildung 3 Facettenmodell der Intelligenz nach Jäger (1982) 35
Abbildung 4 Intelligenzmodell nach Spearman (1904). 37
Abbildung 5 Lernzieltaxonomie nach Bloom et al. (1956). 46
Abbildung 6 Veränderung von der alten (Bloom et al., 1956, rechts) zur neuen (Anderson & Krathwohl, 2001, links) Taxonomie. 49
Abbildung 7 Zusammenfassung von Wilsons (1970) Modell 53
Abbildung 8 Modell zur Ordnung der Mathematik auf Basis einer Kognitive Prozesse x Inhalte-Matrix. 57
Abbildung 9 Guttman-Pattern mit einer Abweichung (VP5). 62
Abbildung 10 Logik der Aufteilung in AT und PT. 70
Abbildung 11 Veranschaulichung der Logik hinter DETECT, nach Zhang & Stout (1999, S. 218) 72
Abbildung 12 Ablaufschema zur Prüfung der N-Dimensionalität eines Tests. 75
Abbildung 13 Notwendige Stichprobengröße im ADF-Verfahren. 79
Abbildung 14 Von DETECT vorgeschlagene Cluster. Die Verbindungslinien zwischen Clustern verdeutlichen, dass sie nicht unabhängig sind LIT (mathematische Literalität), PROZ (prozedurales Rechnen), GEO (Geometrie und grafische Funktionen) 85
Abbildung 15 Beispielaufgabe der Bildungsstandards (Hauptschule). 94
Abbildung 16 Kompetenzen die Schüler zum Ende der Hauptschule (9. Klasse) erworben haben sollten (links) und Kompetenzen die Schüler mit dem mittleren Schulabschluss erworben haben sollten (rechts). Quelle: KMK (2004a, 2005a). 95
Abbildung 17 Aufgabenbeispiel 7 aus den Bildungsstandards für den mittleren Schulabschluss. Quelle: (KMK, 2004a, S. 25). 96
Abbildung 18 Verteilung der Schüler allgemeinbildender- und Berufsschulen auf die Bundesländer. Stand 2007, N ≈ 12,1 Millionen. Quelle: Statistisches Bundesamt. 99
Abbildung 20 Prototypisches Item der Skala Geometrie und grafische Funktionen 113
Abbildung 21 Prototypische Aufgabe der Skala prozedurales Rechnen A1a-d. 115
Abbildung 22 Prototypische Items der Skala komplexes Rechnen, A16a, A16b 116
Abbildung 23 Exploratorische DETECT-Lösung der Endform, $N = 1554$. 131
Abbildung 24 Mittelwerte und Konfidenzintervalle (95%) zur Einschätzung der Wichtigkeit der 6 kognitiven Prozesse nach Anderson und Krathwohl (2001) von 17 Realschullehrern. 152
Abbildung 25 Mittlere Anzahl von Ratings für eine der 6 Taxonomiestufen einschließlich Standardfehler (95%), die Mittelwerte summieren sich zur Anzahl der Items (77) 154
Abbildung 26 Strukturmodell zweier korrelierter Faktoren (Inhalt A und Inhalt B). 156
Abbildung 27 Schmid-Leiman Transformation des Modells gemäß Abbildung 26. 157
Abbildung 28 Darstellung des, finalen SL–Modells. Jeder manifesten Variable ist ein Messfehler zugeordnet, der aus Platzgründen nicht in der Abbildung

aufgeführt ist. LIT = mathematische Literalität, PROZ = prozedurales Rechnen, KOMPL = komplexes Rechnen, GEO = Geometrie und grafische Fkt. 159
Abbildung 29. Unterschiede in den Mittelwerten aller Skalen getrennt für Männer und Frauen. N = 1554. 166
Abbildung 30 Standardwerte von 3 Personen der Normgruppe Gymnasial,über 20 Jahre alt. Alle Personen weisen denselben Gesamtscore auf (Z = 100, Rohwert = 50). Die kritische Differenz und die Normtabellen sind Jasper und Wagener (in Druck) zu entnehmen. 168
Abbildung 31 Tilt-Maß (gestrichelte Linie; größer Null: tilt Richtung mathematische Literalität) und Standardisierter Gesamtscore, getrennt nach bisher erreichtem Abschluss, *N* = 1554. Standardfehler sind aufgrund der großen Stichprobe irreführend und wurden daher nicht abgetragen. 170
Abbildung 32 Verdeutlichung des Prinzips der *within-item* Multidimensionalität (linke Seite) und *between item* Multidimensionalität (rechts), angelehnt an Adams et al. (1997, S. 9). 173
Abbildung 33 Schematische Darstellung des hierarchischen Aufbaus der ersten 4 Taxonomiestufen. 180

TABELLENVERZEICHNIS

Tabelle 1 Deutschsprachige Mathematiktests für Schüler ab Klasse 8 und Erwachsene. .. 19
Tabelle 2 Interkorrelationen zwischen den PISA-Skalen (overarching Ideas) der Studie von 2003 (OECD, 2005, S. 190) 27
Tabelle 3 Auszug zur kognitiven Dimension nach Anderson und Krathwohl (2001) 50
Tabelle 4 Auszug eines Lernzieltaxonomien-Vergleichs nach Reigeluth und Moore (1999, S. 54). 54
Tabelle 5 Aufgaben 1d (mathematisches Grundwissen) und 7d (kaufmännisches Rechnen) des Studententests (Jung, Kempf & Seggewiß, 2007; Orth, 2006). 82
Tabelle 6 Nummerierung der AT-Test Items in DIMTEST, Benennung im Test und Trennschärfen. 84
Tabelle 7 Dendrogramm. Clusterbildungen ab zwei Objekten wurden grau hinterlegt. Die oberste Zeile zeigt den Schritt an. 87
Tabelle 8 Fit-Indizes der exploratorischen NOHARM-Lösungen für ein bis 5-faktorielle Modelle. 88
Tabelle 9 Konfirmatorische, dreifaktorielle NOHARM-Lösung des Expra-Tests. 89
Tabelle 10 Auszüge zweier mathematischer Kompetenzbereiche aus den Bildungsstandards für Mathematik (KMK, 2005a) 93
Tabelle 11 Auszug der Kernkompetenzumschreibung für den Prozessbereich *Argumentieren*, Unterkategorie: hinterfragen mathematischer Aussagen (NK, 2006a, S. 18) 101
Tabelle 12 Auszug der Kernkompetenzumschreibung für den Inhaltsbereich *Daten und Zufall* (NK, 2006, S. 18), Unterkategorie interpretieren Daten (NK, 2006, S. 34). 102
Tabelle 13 Leitidee Daten und Zufall für Haupt-, Werkreal- und Realschule in Baden-Württemberg. Nach BW (2004a, 2004b, 2004c). 104
Tabelle 14 Passung von Hauptschullehrplan und KMK-Bildungsstandards für die Leitidee Raum und Form laut ISB (2005, S. 28) 105
Tabelle 15 Testformen, Klassenstufen und Anzahl von Personen. 110
Tabelle 16 Verbliebene Items Geometrie und grafische Funktionen 112
Tabelle 17 Verbliebene Items prozedurales Rechnen. 114
Tabelle 18 Verbliebene Items der Skala komplexes Rechnen, getrennt für Form A. 115
Tabelle 19 Items der Skala mathematische Literalität. 116
Tabelle 20 Maximal erreichter Schulabschluss der Probanden der Normstichprobe. 119
Tabelle 21 Reliabilitätsschätzungen für Skalen und Gesamtwert in der Gesamt-Stichprobe und getrennt nach Geschlecht. 120
Tabelle 22 Itemanalyse Gesamtstichprobe (N = 1554). 121
Tabelle 23 Interkorrelationen der Mathetest-Skalen in der Gesamtstichprobe (N = 1554). 124
Tabelle 24 Zusammenhang von der Mathematikskalen mit State und Trait-Angst (N = 79). 125
Tabelle 25 Korrelation von VKI mit den Mathetestskalen und Gesmtscore , N = 58. 126
Tabelle 26 Korrelationen von Mathetestskalen mit IST-Subtests Rechenaufgaben, Zahlenreihen und deren Summe. 127
Tabelle 27 Zusammenhang der Skalen des Mathetests mit dem Mittel der letzten beiden Deutsch- und Mathenoten. 128
Tabelle 28 Oblimin Pattern-Matrix und Itemschwierigkeiten der Endform (N = 1554). .. 134

Tabelle 29 Tanaka (GFI)-Index und RMSR Fit-Index für 1, 3 und 4-Faktorielle Lösungen (*N* = 1554). 136
Tabelle 30 Faktorladungen der konfirmatorischen, obliquen NOHARM Lösung (*N* = 1554)..... 137
Tabelle 31 Intekorrelationen der 4 Faktoren einer obliquen NOHARM-Lösung (*N* = 1554). 138
Tabelle 32 Schwierigkeitsbasierte Parcelbildung der Endform auf Skalenebene. 139
Tabelle 33 Pattern Matrix einer schiefwinkligen Faktorenanalyse der Parcels 140
Tabelle 34 Standardisierte Pfadkoeffizienten der 4 Faktorlösung. 146
Tabelle 35 Interkorrelationen der 4 Faktoren. 147
Tabelle 36 Minimal- und Maximalwert, Mittelwert, Streuung und Anzahl der Einschätzung aller Aufgaben durch Realschullehrer auf einer Skala von 1 = erinnern bis 6 = kreieren. 149
Tabelle 37 Gegenüberstellung von bereits getesteten obliquen-Modellen, einem G-Faktor Modell und zwei Schmid-Leiman Modellen (*N* = 1554). 158
Tabelle 38 Standardisierte Pfadkoeffizienten der 4 Faktor-SL-Lösung der Mathetest-Parcel. 161
Tabelle 39 Pattern Matrix der Mathetest-Parcels für die schlechtere Hälfte der Stichprobe (Gesamtscore < 39, *N* = 787). 163
Tabelle 40 Trennbarkeit von Personen mit und ohne Abitur anhand hierachischer Diskriminanzanalyse. 164
Tabelle 41 Mittelwerte und Mittelwertsunterschiede für Männer (*N* = 1048) und Frauen (*N* = 482) der Stichprobe. 166
Tabelle 42 Varimax-Rotierte Faktorladungsmatrix der Parcels für komplexes Rechnen und mathematische Literalität 169
Tabelle 43 Korrelationen zwischen den IRT-basierten Dimensionen nach Conquest (*N* = 1554)..... 176
Tabelle 44 Alle Items, die gemäß dem Modus der Kategoriezuordnung durch 17 Rater den ersten 4 Stufen der Lernzieltaxonomie zugeordnet wurden. 178
Tabelle 45 Häufigkeiten mit denen von den 17 Ratern Items der 4 Mathetestskalen den ersten 4 Taxonomiestufen zugeordnet wurden. 181
Tabelle 46 Standardisierte Residuen bei Annahme von Unabhängigkeit der Zuordnung Taxonomiestufe x Skalenzugehörigkeit. 182
Tabelle 47 Klassische Kennwerte aller Items des Expra-Tests, N = 182. 209
Tabelle 48 Für jede Aufgabe ist abgetragen in welchem Test sie auftaucht und wie sie dort heißt. 213
Tabelle 49 Dreifaktorielle Lösung korrelierter Faktoren N = 1554..... 218
Tabelle 50 4-Faktorielle Schmid-Leiman-Lösung der Endform mit WLSMV-Schätzung. 219
Tabelle 51 Klassische Kennwerte vor Itemselektion, Form A (N = 73). 220
Tabelle 52 Klassische Kennwerte vor Itemselektion, Form A (N = 73). 220
Tabelle 53 Klassische Kennwerte vor Itemselektion, Form A (N = 73). 221
Tabelle 54 Klassische Kennwerte vor Itemselektion, Form A (N = 73). 221
Tabelle 55 Klassische Kennwerte vor Itemselektion, Form B (N = 76)..... 223
Tabelle 56 Klassische Kennwerte vor Itemselektion, Form B (N = 76)..... 224
Tabelle 57 Klassische Kennwerte vor Itemselektion, Form B (N = 76)..... 224

ABKÜRZUNGSVERZEICHNIS

ADF = Asymptotic distribution free

ANOVA = Analysis of Variance

AIC = Akaike's Information Criterion

APA = American Psychological Association

AT = Assessment test

BW = Baden-Württemberg

CFI = Comparative fit index

GFI = General fit index

HCA/CCPROX = hierarchical cluster analysis / conditional covariance proximities

HPI = Hierarchisches Rahmen bzw. Protomodell der Intelligenzstrukturforschung

ICC = Intra-Klassen-Korrelation

IRT = Item response theory

ISB = Institut für Schulqualität und Bildungsforschung

JM = Joint moment

KTT = Klassische Testtheorie

ML = Maximum Likelihood

MATLUB = Mathematiktest für Lehre und Beruf

MRCML-Modell = Multidimensional Random Coefficient Multinominal Logit-Modell

MTMM = Multi trait multi method

NFI = Normed fit index

NK = Niedersächsisches Kultusministerium

NNFI = Non-normed fit index

NOHARM = Normal Ogive by Harmonic Analysis

NRW = Nordrhein-Westfalen

OECD = Organisation for economic co-operation and development

PISA = Program for International Student Assessment

PK = Pisa Konsortium

PTT = Probabilistische Testtheorie

PT = partitioning test

RMSR = Root mean square redsiduals

RMSEA = Root Mean Square Error of Approximation

RT 9+ = Rechentest 9+

SEM = Structural Equation Modeling

SL = Schmid-Leiman

STAI = State-Trait-Angstinventar

TIMSS = Third International Mathematics and Science Study

UPGMA = Unweighted pair group method with arithmetic mean

VKI = Verbaler Kurzintelligenztest

WLSMV = weighted least squares with mean and variance adjusted

I THEORETISCHER TEIL

1 Einleitung

Das Bundesministerium für Bildung und Forschung veranstaltet seit dem Jahr 2000 die so genannten Wissenschaftsjahre. Im Jahr 2008 war es so weit: Nach dem Einsteinjahr, dem Informatikjahr und dem Jahr der Geisteswissenschaften ist nun die Mathematik auserkoren worden (Bundesministerium für Bildung und Forschung [BFBUF], 2008). Dass gute Mathematikkenntnisse für eine Ausbildung, ebenso wie für ein Studium, von hoher Wichtigkeit sind und eine Bedingung für das Verständnis unserer Lebenswelt darstellen, ist unstrittig (BFBUF, 2008, S. 17). Sehr zeitnah war es die Deutsche Industrie und Handelskammer, die in einer Online-Befragung an über 10.000 Betrieben feststellte, dass 52% aller Betriebe als Ausbildungshemmnisse bei der Annahme neuer Lehrlinge unzureichende Ausbildungsreife der Schulabgänger angeben (Borstel, 2008; Deutsche Industrie und Handelskammer [DIHK], 2006). Genauer betrachtet, sind es vor allem die mangelnden Mathematikfähigkeiten (elementare Rechenfertigkeiten), die mit über 50% als großes Ausbildungshemmnis identifiziert wurden (Neuman, 2006).

Sicherlich haben auch die Ergebnisse der PISA-Studien, in denen die deutschen Schüler im internationalen Vergleich eher mittelmäßige Leistungen erbrachten (OECD, 2006), trotz zunehmender Kritik an dem Vorgehen bei PISA im Allgemeinen (Hopmann, Brinek & Retzl, 2007; Jahnke & Meyerhöfer, 2006), die Aufmerksamkeit der Öffentlichkeit auf diesen Themenbereich gelenkt.

Umso wichtiger wird daher die Leistungsdiagnostik in eben diesem Bereich. Anhand von reliablen, validen und objektiven Instrumenten wird es für Unternehmen möglich, in der Personalauswahl und Entwicklung die besten Entscheidungen zu treffen. Schuler, Hell, Trapmann, Schaar und Boramir (2007, S. 65) untersuchten vor kurzem die Bedeutung verschiedener Methoden zur Personalauswahl und kamen unter anderem zu dem Schluss, dass bei der Gruppe der Auszubildenden mit über 30% Anteil Leistungstests am häufigsten eingesetzt werden.

Derzeit existieren viele Verfahren für die Mathematikdiagnostik bei jüngeren Schülern, (z. B. DEMAT 1+, DBZ, SR1-3, HRT 1-4, DEMAT 2+, MT 2, ZAREKI, DEMAT 3+, DRE 3, DEMAT 4, SR 4-6, BDT 6 für die Klassen 2 bis 6), wohingegen es für die Diagnostik am Ende der Sekundarstufe I wenig etablierte Verfahren gibt (Hofe, Michael, Blum & Pekrun, 2005). Diese Arbeit widmet sich der Psychometrie der Mathematik am Ende der Sekundarstufe I. Ziel ist es, den aktuellen Forschungsstand zu deuten und zu interpretieren.

Ferner soll ein eigener Beitrag zur Strukturierung und Mehrung des vorhandenen Wissens in diesem Bereich geleistet werden. Dazu wird – dies sei vorweggenommen - ein eigener Mathematiktest entwickelt, mit dem theoretische Annahmen umgesetzt und geprüft werden. Erfreulicher Weise suchte der Hogrefe-Verlag zur Konstruktion eines neuen Mathematiktests - der Verlag führt derzeit kein aktuelles Verfahren für Probanden am Ende der Sekundarstufe I – Autoren und wandte sich an die Universität Mannheim. Dies ermöglichte erst die finanziell aufwändige Erhebung der Normstichprobe. Zur Veranschaulichung werden viele Aufgaben präsentiert, die sich bis auf Oberflächenmerkmale nicht von den Testaufgaben der Endform unterscheiden. Bei Interesse an der Endform und um Nachprüfungen zu ermöglichen, bitte ich um Kontaktaufnahme.

1.1 Psychometrie

Bereits in der ersten Ausgabe der Zeitschrift *Psychometrika* (1936) heißt es im Untertitel*: A Journal devoted to the development of Psychology as a Quantitative Rational Science*. Interessant ist in diesem Zusammenhang auch die Antwort von Paul Kline auf die Frage, was Psychometrie darstellt. So schreibt er in seinem letzten Werk *A Psychometrics Primer* (Kline, 2000, S. 1) „Psychometrics refers to all those aspects of psychology which are concerned with psychologial testing, both the methods of testing and the substantive findings“. Paul Horst (1971) weist ferner darauf hin, dass die Psychometrie nicht als Teilgebiet der Psychologie anzusehen ist, sondern in allen Bereichen der Psychologie von Bedeutung ist. Deshalb wurde im Titel dieser Arbeit auch eine präzise Eingrenzung, nämlich auf die Mathematik am Ende der Sekundarstufe I, vorgenommen. Hier geht es also darum, wie bislang in der Psychologie in diesem Gegenstandsbereich, quantitative Messungen vorgenommen wurden, die bisherigen Ansätze zu bewerten und - wenn nötig – einen neuen Ansatz zu entwickeln.

1.2 Leitfragen dieser Arbeit

In dieser Arbeit wird versucht, anhand der modernsten derzeit verfügbaren statistischen Methoden die Forschung zur Psychometrie der Mathematik am Ende der Sekundarstufe I voranzutreiben. Innerhalb der letzten Jahrzehnte haben sich die Möglichkeiten im Bereich der Methodenlehre – vor allem durch Einzug der EDV – rapide entwickelt. Leider ist dementsprechend ein zunehmendes Auseinanderdriften von mathematischer Psychologie

und anderen Bereichen psychologischer Forschung absehbar. In einem Beitrag zum 40 jährigen Bestehen des *Journals of Mathematic Psychology* zeigt Falmagne (2005, S. 437) auf, dass in den ersten Ausgaben etwa 3% der Beiträge (5 von 148 in 5 Jahren) der Kategorie Mathematik und Methodologie zuzuordnen waren und dieser Anteil für die Jahre 2000 bis 2004 auf 38% (66 von 174) angestiegen ist. Da nicht erwartetet werden kann, dass jeder Psychologe mit jeder neuen Methode vertraut ist, werden in dieser Arbeit Ansätze, die nicht zum Standardrepertoire (z.B. ANOVA, lineare Faktorenanalyse etc.) gehören, ausführlich genug dargestellt, um das Vorgehen nachvollziehbar zu machen. Der Nutzen komplexer Verfahren wird hierbei stets herausgestellt.

Ein Ziel dieser Arbeit ist es, sowohl für mehr praktisch orientierte Diagnostiker, als auch für Forscher im Bereich der diagnostischen Psychologie einen Mehrwert zu erbringen. Deshalb wird im Laufe dieser Arbeit neben einem durch die Intelligenzdiagnostik geprägtem Blick auf die Erfassung von Mathematik auch eine Testanalyse anhand einer Lernzieltaxonomie durchgeführt.

Während dieser Arbeit ergab sich die problematische Situation, dass für aktuelle Testverfahren Normierungsdaten fehlen oder nicht zur Verfügung gestellt wurden (vgl Abschnitt 2.1). Dies schränkt eines der wesentlichsten Prinzipien jeder Forschung massiv ein, die Nachprüfbarkeit (bzw. Falsifizierbarkeit, Chalmers, 2007). Deshalb wurden die in dieser Arbeit angefallenen Daten mehrfach gesichert und sind auf Nachfrage zu wissenschaftlichen Zwecken beim Autor dieser Arbeit erhältlich.

1.3 Bedeutung von Multidimensionalität für die praktische Diagnostik

Einen wichtigen Aspekt dieser Arbeit stellt die Bestimmung der Dimensionalität von Mathematiktests dar. Abschnitt 4.2 widmet sich vor allem den technischen Aspekten der Dimensionalitätsbestimmung. Ebenfalls behandelt wird die Frage nach der praktischen Bedeutung von Multi-(Dimensionalität), schließlich sind für alle im folgenden vorgestellten kommerziellen Mathematiktests aus unterschiedlichen (teils auch nachvollziehbaren) Gründen kaum elaborierte Analysen der Dimensionalität vorgenommen worden. Die Kosten von Fehlentscheidungen bei Stellenbesetzungen (Lorenz & Rohrschneider, 2009, S. 10 ff.), die durch mangelnde Prüfung eines Fähigkeitsprofils entstehen können, können enorm sein und es überrascht, dass die empirische Prüfung der Dimensionalität anscheinend nach wie vor häufig vernachlässigt

wird. Ein Grund dafür könnte darin bestehen, dass bereits vor über 90 Jahren, damals beim Stanford-Binet-Intelligenztest, nachweislich versucht wurde Tests derart zu konstruieren, dass sie keine Unterschiede zwischen Gruppen mit nachweislich verschiedenen Fähigkeitsprofilen aufweisen (Ackermann, 2002). Eine ähnliche Situation findet sich auch heute im Kontext der PISA-Untersuchungen. Wie Wittmann (2004) zeigte verdoppeln sich die Unterschiede zwischen den Geschlechtern in den Bereichen Lesen, Mathematik und Wissenschaft, wenn anstelle von PISA item response theory scores die Faktorwerte einer Hauptkomponentenanalyse gewählt werden (vgl. auch Abschnitt 9.5). Demnach ist es der komplexe Aufbau der Items, der die Unterschiede zwischen diesen Gruppen in PISA-Berichten (in beide Richtungen) eher abmildert (vgl. Abschnitt 9.4). Das bedeutet, dass durch Heterogenität der Bestandteile von Testaufgaben die Möglichkeit trennbare Skalen zu erhalten sinkt.

Die Bedeutung von Profilunterschieden, die Analysen zur Dimensionalität zwingend voraussetzen (sollten), zeigt sich auch eindrucksvoll in einer Längsschnittuntersuchung von Lubinski et al. (2001). Sie fanden heraus, dass Jugendliche mit hohen und gleichzeitig stark divergierenden Fähigkeitsprofilen (Mathematik versus Sprachen) 10 Jahre nach der Testung nicht nur sehr erfolgreich waren, sondern auch in Schule und Hochschule klar entsprechend ihrem Profil zu verbal- oder mathematiklastigen Wissenschaftsrichtungen tendierten.

In dieser Arbeit wird versucht die Inhalte von Items die zu einer Skala gehören möglichst rein zu halten und eine Vermischung der Inhalte innerhalb von Items (z.B. viel Text in Geometrieaufgaben) zu verhindern

2 Aktueller Forschungsstand

Im folgenden wird versucht, einen Überblick bezüglich derzeit verfügbaren kommerziellen Mathematiktests und der Rolle von Mathematik in internationalen Vergleichsstudien zu geben.

2.1 Verfügbare kommerzielle Mathematiktests

Die Anzahl der für den deutschen Sprachraum verfügbaren reinen Mathetests, die sich für Testanden zum Ende der Sekundarstufe I eignen, ist relativ überschaubar. Tabelle 1 fasst einige der bekanntesten Tests zusammen.

Tabelle 1 Deutschsprachige Mathematiktests für Schüler ab Klasse 8 und Erwachsene.

Testverfahren	Autoren	Erscheinungsjahr
Rechentest 8+	Fisch, Hylla & Süllwold	1965
Mathematiktest für Abiturienten und Studienanfänger (MTAS)	Lienert & Hofer	1972
Berufsbezogener Rechentest	Balser, Ringsdorf & Traxler	1986
Rechentest 9+	Bremm & Kühn	1992
Mathematiktest für Lehre und Beruf	Ibrahimovic´& Bulheller	2005

Erste einigermaßen brauchbare Verfahren zur Leistungsdiagnostik im Bereich der Mathematik existierten bereits in den 1950er Jahren, z.B. mit dem Frankfurter Rechentest für das 8. Schuljahr von 1959 (Ingenkamp, 1962, S. 153). Während diese ersten Verfahren zwar als anwendbar gelten, jedoch kaum auf pädagogischen Konzepten fußten (Ingenkamp, 1964, S. 137), wurde mit dem Rechentest RT 8+ von Fisch, Hylla und Süllwold (1965) erstmals ein durchdachtes Verfahren für den deutschen Raum vorgestellt.
Ursprünglich war geplant, eine statistische Reanalyse bestehender Tests anhand von Originaldaten (z. B. Korrelationsmatrix der Aufgaben) vorzunehmen, was jedoch aus verschiedenen Gründen scheiterte, die bei den folgenden (nicht rohdatenbasierten) Analysen erwähnt werden. Während die in Abschnitt 2.2 zu besprechenden, internationalen Vergleichsstudien große Aufmerksamkeit (Kraus, 2005; Payk, 2009) nach sich gezogen haben, kann dies für eher traditionelle psychologische Tests nicht behauptet werden.

2.1.1 Analyse des Rechentest 8+

Der Rechentest RT 8+ (Fisch, Hylla & Süllwold, 1965) stammt aus einer Zeit in der die Datenverarbeitung mittels Computern noch in den Kinderschuhen steckte. Da im Testmanual keine Korrelationsmatrizen enthalten sind, erübrigt sich die weitere Nachforschung. Ohnehin ist das Alter des Tests bereits an dessen sprachlichen

Formulierungen erkennbar, was auch bei der Konstruktion des Nachfolgers herausgestellt wurde (Bremm & Kühn, 1992). Beispielsweise heißt es an einer Stelle im Test *Verwandle 1/6 in einen Dezimalbruch* oder ein Testteil ist durch die Überschrift *Von Dezimalbrüchen* gekennzeichnet (Fisch et al., 1965). Der Test ist aufgeteilt in 6 Skalen und zwar von ganzen Zahlen (1), von den Maßen (2), von gemeinen Brüchen (3), von Dezimalbrüchen (4), vom Schlußrechnen (5) und vom Prozentrechnen (6). Die Korrelationen zwischen den einzelnen Subtests schwanken von $r = 0{,}42$ bis $r = 0{,}62$. Daraus und aus den noch höheren attenuationskorrigierten Korrelationen schließen die Autoren, dass die einzelnen Subtests auf denselben fundamentalen Fähigkeiten beruhen (Fisch et al., 1965, S. 11).

Im RT 8+ findet sich keine einzige Geometrieaufgabe oder eine Aufgabe, die eine Zeichnung enthält. Keine Aufgabe des Tests enthält übermäßig viel Text, gleichzeitig weisen nur zwei der sechs Subtests keine Aufgaben mit inhaltlicher Einkleidung auf. Diese oberflächliche Betrachtung der RT 8+ zeigt somit nicht unbedingt, dass es sich bei Mathematikfähigkeit um ein mehrdimensionales Konstrukt handelt, sondern eher, dass es möglich ist einen Mathematiktest zu konstruieren, der glaubhaft nur eine Dimension erfasst.

2.1.2 Analyse des MTAS

Der Mathematiktest für Abiturienten und Studienanfänger (MTAS) (Lienert & Hofer, 1972) wurde 7 Jahre nach dem RT 8+ (Fisch et al. , 1965) veröffentlicht, also auch zu einer Zeit, in der die Arbeit mit Computern in diesem Kontext kaum möglich war. Der Test weist einen Gültigkeitsbereich auf, der streng genommen ein anderer ist, als der in dieser Arbeit vorgegebene (Ende der Sekundarstufe I). So ist der Zweck des MTAS Abiturienten die Studiumsauswahl zu erleichtern (Lienert & Hofer, 1972, S. 5), es wird also auf dem zu erwartendem Niveau am Ende der Sekundarstufe II getestet. Da jedoch generell wenige Mathematiktests für den deutschsprachigen Raum (ab ca. 15 Jahren) existieren, erscheint ein Blick auf diesen Test lohnenswert. Lienert und Hofer (1972) unterscheiden zwischen Geometrie, Algebra und Funktionen, wobei die Interkorrelationen zwischen $r = 0{,}27$ (Algebra und Funktionen) und $r = 0{,}55$ (Algebra und Geometrie) variieren. Der Test ist komplett in einem Multiple-Choice Format zu beantworten (stets Alternativen: A, B oder C) und enthält nur dort Text, wo es unbedingt nötig ist. Im Gegensatz zu anderen in dieser Arbeit vorgestellten erhältlichen Tests umfasst der MTAS auch das Rechnen mit Logarithmen und Differentialrechnung, was das insgesamt höhere Fähigkeitsniveau der Zielpopulation unterstreicht. Interessant am MTAS ist vor allem, dass durch Konzentration

auf nur drei Inhaltsbereiche lediglich moderate Korrelationen zwischen den Skalen bestehen.

2.1.3 Analyse des Berufsbezogenen Rechentests

Die Korrelationsmatrix des Berufsbezogenen Rechentests (BRT) (Balser, Ringsdorf & Traxler, 1986) ließ sich leider anhand des Manuals nicht rekonstruieren. Zwar existiert eine Grafik, in der Trennschärfen und Schwierigkeiten abgetragen wurden (Balser, Ringsdorf & Traxler, 1986, S. 12), doch ist selbst hier eine Analyse nur eingeschränkt möglich. So existiert z.B. die Aufgabe 3 in dieser Grafik mehrfach (Testteil I) und für einige Aufgaben (z.B. Aufg. 33 (Testteil II), Aufg. 7 (Testeil II)) sind die Beschriftungen nicht zweifelsfrei dem Punkt im Diagramm zuzuordnen.

Hervorzuheben ist, dass es sich wohlgemerkt um keinen Fehler der Autoren handelt, sondern eher das verwendete Statistikprogramm (vermutlich SPSS 9 von 1982) die Ursache darstellt. Die Langform des BRT (Zeitbegrenzung 84 Minuten), unterteilt sich in 8 Skalen:

Dezimalbrüche (1), Maße (2), Algebra (3), Geometrie (4), Grundrechenarten (5), Gewöhnliche Brüche (6), Prozentrechnen (7) und Schlussrechnen (8) (Balser et al., 1986, S. 8). Die Reliabilität der Skalen variiert zwischen 0,65 (Skala 2) und 0,90 (Skala 6). Um welche Form der Reliabilitätsschätzung es sich handelt, wird nicht berichtet. Da die Rangkorrelation zwischen Anzahl der Items und Reliabilitätsschätzung r = 0,81 beträgt, handelt es sich vermutlich um Cronbach‘s α (Cronbach, 1951) oder eine Split-Half Korrelation. Informationen zu den Interkorrelationen der 8 Skalen fehlen. Gegen 8 eigenständige Skalen spricht, dass die Autoren selbst berichten (Balser et al., 1986, S. 17) eine 7-faktorielle Struktur gefunden zu haben, bestehend aus Aufgaben ohne Text (1), Textaufgaben (2), Schlussrechnen – Geometrie – Algebra (3), Algebra (4), Gewöhnliche Brüche (5), Schlussrechnen (6) und Maße (7). Leider sind weder die Interkorrelationen der 7 Skalen noch die Rotationsart der Faktorenanalyse benannt (Varimax, Oblimin etc.). Auch existiert keine Faktorladungsmatrix, was es sehr schwer, macht die Befunde zu bewerten.

2.1.4 Analyse des Rechentests 9+

Der Rechentest 9+ (RT9+) (Bremm & Kühn, 1992) unterscheidet zwischen Bruchrechnen (1), Prozentrechnen (2), Zinsrechnen (3), Gleichungen (4), Potenzen und Wurzeln (5), sowie Rechnen mit Größen (6). Die Autoren (Bremm & Kühn, 1992, S. 10) schlagen auf

Basis ihrer Daten eine 8-Faktorenlösung vor. Die sehr hohe Faktorenzahl ist fraglich und aufgrund der eher geringen Varianzaufklärung der letzten 6 Faktoren (alle Varianzaufklärung ≤ 5%), sowie nicht vorhandenen theoretischen Begründung kaum gerechtfertigt. Vor allem überrascht, dass die Anzahl der Faktoren (8) höher ist als jene der postulierten Skalen (6). Das Ladungsmuster (S. 11) ist unvollständig dargestellt und auch die vorhandenen Ladungen lassen Zweifel an der Trennbarkeit der 8 Faktoren aufkommen. Dies verwundert von daher nicht, da unbekannt bleibt, wieso sich z. B. Prozentrechnen (Faktor 4) und Zinsrechnen (Faktor 6) trennen lassen sollten. Zu den Interkorrelationen der Skalenwerte finden sich im Manual keine Informationen.

Als Basis für diesen Test wurden in erster Linie die Lehrpläne der seinerzeit alten Bundesländer herangezogen. Auf jegliche Geometrieaufgaben wurde verzichtet, auch auf den in allen Lehrplänen enthaltenen *Satz des Pythagoras*, aus „testökonomischen Gründen“ (Bremm & Kühn, 1992, S. 4), was jedoch nicht einleuchtet. Der RT 9+ enthält ausschließlich Aufgaben mit offenem Antwortformat, der Textanteil ist auch bei Aufgaben mit inhaltlicher Einkleidung recht gering (Bremm & Kühn, 1992). Eine über Lehrpläne hinausgehende theoretische Grundlage für den Test fehlt.

2.1.5 Analyse des Mathematiktest – Grundkenntnisse für Lehre und Beruf

Die theoretischen Ausführungen zum Mathematiktest für Lehre und Beruf (MATLUB) (Ibrahimovic & Bulheller, 2005) sind recht kurz (eine halbe Seite). Dort heißt es, eine Analyse der curricularen Anforderungen, Schulbücher, Lehrerurteile und die Analyse bestehender Verfahren seien die Grundlage für die Testkonstruktion gewesen. Das Itemformat besteht zum einen aus Aufgaben mit offenem Antwortformat und zum anderen aus Aufgaben, bei denen die Antwort durch Ankreuzen von einer oder mehreren Zahlen aus einer Reihe von 0 bis 9 besteht. Der Test setzt sich aus vier Subskalen zusammen (Textaufgaben, textfreie Aufgaben, Geometrie, Tabellen- und Grafikverständnis), von denen textfreie und Textaufgaben mit $r = 0{,}70$ (für Form B an zweiter Stelle mit $r = 0{,}67$) am höchsten miteinander korrelieren. Am niedrigsten korrelieren Geometrieaufgaben mit Aufgaben zum Tabellen- und Grafikverständnis (Form A: $r = 0{,}46$, Form B: $r = 0{,}43$).

Diese Struktur spricht nicht gegen einen praktischen Einsatz des Tests, doch stellt sich aus psychologischer Sicht die Frage, wieso gerade diese Subskalen gewählt wurden - eine Faktorenanalyse auf Itemebene wird nicht berichtet - und weshalb gerade textfreie und Textaufgaben am höchsten miteinander korrelieren. Da weder Iteminterkorrelationen noch eine Faktorenanalyse auf Itemebene durchgeführt wurden, gibt es praktisch keine Befunde

zur Faktorenstruktur des MATLUB. Bei einer rotierten Faktorenanalyse auf Skalenebene zusammen mit den Skalen anderer Tests (u.a. des Intelligenz-Struktur-Analyse Tests und des Frankfurter Aufmerksamkeitsinventars) laden alle MATLUB-Skalen am höchsten auf dem ersten Faktor. Die Autoren waren auch nach einer offiziellen schriftlichen und telefonischen Anfrage des Lehrstuhls Psychologie II der Universität Mannheim nicht bereit eine Korrelationsmatrix für Analysen zur Verfügung zu stellen.

2.1.6 Schlussfolgerung aus Sichtung aktueller Mathetests

Die Schlussfolgerung fällt eindeutig aus: Es gibt zu wenige Verfahren die für die hier betrachtete Altersgruppe anwendbar sind (nur ein aktuelles, den MATLUB). Leider lassen sich kaum Ideen aus den vorliegenden Tests - zwecks wissenschaftlicher Prüfung oder Aufstellung einer neuen Theorie - ableiten.

Die Tests wurden nicht entwickelt, um die Forschung in diesem Bereich voranzutreiben. Die vorliegenden Verfahren sind alle mehr oder weniger curriculumsbasiert. Wichtig ist, dass eine solche Orientierung die Frage nach der Konstruktbeschaffenheit (z.B. faktorielle Struktur) nicht ausschließt. Wenn überhaupt Versuche unternommen wurden, die einzelnen Aufgabentypen – z.B. anhand einer Faktorenanalyse - zu trennen, so wurden schlicht auf Basis von Aufgabenarten Faktoren gebildet, was zu schlechter Passung führte und schwer theoretisch begründbar ist (und auch nicht weiter begründet wird). Da sämtliche Normierungsdaten entweder nicht mehr auffindbar waren, oder sich die Autoren weigerten sie zu wissenschaftlichen Zwecken zur Verfügung zu stellen, ist die Konstruktion eines neuen Tests notwendig. Eine Neuentwicklung ist sinnvoll, um theoretische Annahmen zur Struktur der Mathematik empirisch zu prüfen. Darüber hinaus ergibt sich dadurch die Möglichkeit potentieller Anwender zwischen Verfahren zu wählen. Es sei herausgestellt, dass die theoretische Fundierung aller hier dargestellten Verfahren zwar dürftig ist, dies jedoch in erster Linie ein wissenschaftliches und nicht unbedingt ein praktisches Problem darstellt.

Ferner ist wichtig deutlich zu betonen, dass die obigen Analysen nicht als Kritik an der Sinnhaftigkeit der Tests zu verstehen sind. Es war nicht das Ziel der Tests, wissenschaftliche Hypothesen zu prüfen.

2.2 Mathematik in internationalen Vergleichsstudien

Innerhalb der letzen 15 Jahre hat die Diagnostik von Schulleistungen, maßgeblich verursacht durch die großen TIMSS und PISA Untersuchungen, deutlich an Bedeutung gewonnen. Nachdem eine detaillierte Betrachtung kommerzieller Tests vorgenommen wurde, ist die Frage, wie die beiden bereits erwähnten Großuntersuchungen - mit hohem politischem und wissenschaftlichem Einflussfaktor - Mathematikfähigkeiten definieren und erfassen. Die Ergebnisse der TIMMS Studien (Leibniz-Institut für die Pädagogik der Naturwissenschaften [IPN], 2000, S. 5) lassen sich aufgliedern in TIMMS I (Grundschule, ohne deutsche Beteiligung), TIMMS II (Mittelstufe) und TIMMS III (Oberstufe) wohingegen sich die PISA Studien (2000, 2003, 2006, 2009) auf 15 Jahre alte Schüler beziehen (OECD, 2007). Demnach befassen sich beide Studien mit einer Klientel, die altersmäßig nicht ganz dem typischen Auszubildenden am Ende der Sekundarstufe I entspricht (d.h. eher zu junge Teilnehmer). Da die Nähe zum Ende der Sekundarstufe I jedoch deutlich ist, scheint es geboten, mögliche Ordnungs- und Definitionsversuche der Vergleichsstudien zu betrachten.

2.2.1 Third International Mathematics and Science Study

Die TIMSS-Aufgaben basieren auf einer Inhalt mal kognitiver Anspruch-Matrix, die ursprünglich auf Blooms Lernzieltaxonomie (Bloom, Englehart, Furst, Hill & Krathwohl, 1956) und einen Ordnungsversuch von Wilson (1970) zurückgeht (IPN, 1998; IPN, 2000). Die Studien werden regelmäßig alle 4 Jahre an Viert- und Achtklässlern durchgeführt, jedoch seit 2003 mit einem kleineren Versuchsdesign, das im Wesentlichen die eigentlichen Leistungs-Fragebögen enthält. (Olsen, 2005, S. 23). Deutschland nimmt seit 1999 nicht mehr an den internationalen TIMSS-Studien mit Achtklässlern teil (International Association for the Evaluation of Educational Achievement [IEA], 2004a).
In der 2003er Studie (und sehr ähnlich in der 99er Studie) werden als Inhaltsdimensionen Nummern, Algebra, Messung und Geometrie sowie Daten und als kognitive Dimensionen Wissen von Fakten und Prozeduren, Konzepte anwenden, Routineprobleme lösen und Verarbeitungskapazität (reasoning) genannt (IEA, 2004b, S. 9). Bei der derzeit neuesten Studie aus dem Jahre 2007 wurden *Messung* und *Daten* durch *Daten und Zufall* ersetzt (IEA, 2008, S. 372). Statt der ursprünglichen Aufteilung der kognitiven Dimension (IEA, 2000) ist 2007 nur noch knapp von *Wissen*, *Anwenden* und *Verarbeitungskapazität* die Rede (IEA, 2008, S. 372). Das Ausmaß, in dem die einzelnen Inhaltsdimensionen der

neuesten Studie für Achtklässler zum Gesamttest beitragen sollen, zeigt die folgende Abbildung 1.

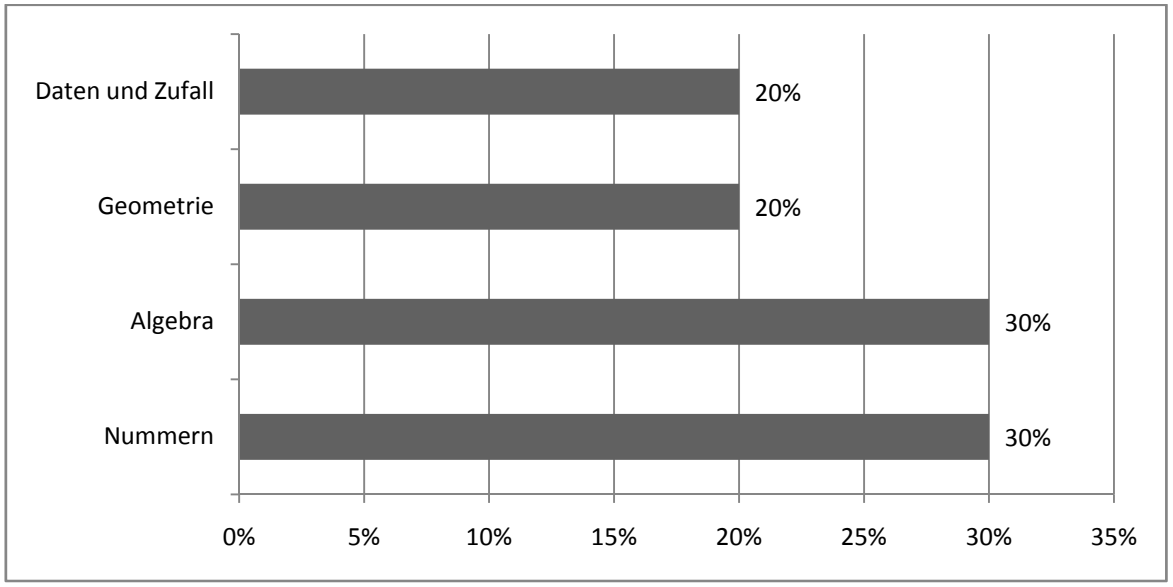


Abbildung 1 Ausmaß in dem die 4 Inhaltsdimensionen der TIMSS 2007-Untersuchung im Test enthalten sind.

Bei TIMSS wurde großer Wert auf die curriculare Validität der Aufgaben gelegt (IEA, 2005, S. 4; IEA, 2008, S. 198; IPN 1998), die Frage nach dem psychologischen Konstrukt der Mathematik wurde nicht untersucht. Die TIMSS-Studien zeigten und zeigen sicherlich eindrucksvoll einen Vergleich der Länder und erfassen, inwiefern curriculares Wissen beherrscht wird, doch die Frage, wie Mathematikfähigkeit psychometrisch geordnet und analysiert werden könnte, wurde nicht beantwortet, da sie nicht im Fokus der Untersuchung stand. In keinem der offiziellen Berichte zu den Studien von 1999 (IEA, 2000), 2003 (IEA, 2004b) und 2007 (IEA, 2008) finden sich Informationen zu den Korrelationen innerhalb (bzw. zwischen) den inhalts- oder kognitiven Dimensionen. In einen Dokument, das sich speziell den kognitiven Dimensionen (in Bezug auf Outcome-Vergleiche) in TIMSS 2003 widmet, wurden jedoch Korrelationen zwischen den kognitiven Dimensionen berichtet, wobei Verwirrung vorprogrammiert ist, da aufgrund von Schwierigkeiten die vier kognitiven Dimensionen gemäß TIMSS 2003 (siehe oben) inhaltlich zu trennen, sie zu den drei kognitiven Dimensionen gemäß TIMSS 2007 (siehe oben) zusammengelegt wurden (IEA, 2005, S. 9). Der Median der Korrelationen über alle Länder hinweg, lag bei $r = 0{,}95$ zwischen *wissen* und *anwenden*, und $r = 0{,}81$ zwischen *anwenden* und *Verarbeitungskapazität* sowie *wissen* und *Verarbeitungskapazität* (IEA, 2005).

Recht unklar bleibt neben der Konstruktbeschaffenheit der Mathematik (aus psychometrischer Sicht), die mögliche Eignung der Aufgaben zur Verwendung in Leistungstests zu Auswahlzwecken. Da nur der Vergleich der Leistungen im Fokus stand, wurden Korrelationen zu Außenkriterien, wie z. B. der Mathenote in den meisten bisher genannten Veröffentlichungen nicht einmal erwähnt. Die äußerst realitätsnahe Operationalisierung von Mathematikfähigkeit zeigt sich bei TIMSS auch in der Formulierung einiger Aufgaben, bei denen es z. b. heißt „Wie rechnest Du?“ oder „Schreibe deine Lösungsschritte auf“ (IPN, 1998, S. 50). Diese Sichtweise von Mathematik als sehr breitem Konstrukt spiegelt sich auch in der Verwendung eines sehr allgemeinen multidimensionalen Rasch-Testmodells (Adams, Wilson & Wang, 1997) wieder, das gegen Ende dieser Arbeit detailliert beschrieben wird.

2.2.2 Programme for International Student Assessment

In Bezug auf die PISA-Studien ist hier vor allem PISA 2003 von Interesse, da dort der Schwerpunkt auf dem Bereich Mathematik lag, wohingegen bei der Untersuchung aus dem Jahre 2006 Mathematik einen kleineren Bereich einnahm (Pisa-Konsortium Deutschland [PK], 2007). Unabhängig davon soll in jeder der PISA-Studien (auch) die so genannte *mathematical literacy* erfasst werden, deren offizielle Definition der OECD (2003, S. 15) wie folgt lautet:

„An individuals capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individuals life as a constructive concerned and reflective citizen “. Diese zweifelsohne normativ wirkende (Weinert, 2001, S. 288) Rahmensetzung hat durchaus auch zu teils harscher Kritik an der Studie geführt (Kraus, 2005; Wuttke, 2007), wobei einer der Kernpunkte die Frage zu sein scheint, was PISA genau misst. Zu dem Begriff der mathematical literacy wie oben beschreiben passt, dass die tatsächlichen Testaufgaben eine starke Einbettung in alltägliche schülernahe Kontexte aufweisen, wie z.B. den Kauf von Skateboardteilen bei begrenzten Ressourcen (Pisa-Konsortium Austria, 2009). Weiter weisen sie häufig einen hohen Textanteil auf, der sich in der starken Korrelation von mathematical literacy und *reading literacy* von $r = 0{,}77$ zeigt (Bodin, 2007).

Letztlich kann das, was im mathematischen Teil von PISA erfasst werden soll, auch als Problemlösekompetenz bezeichnet werden (OECD, 2003, S. 34), die dort durch einen so

genannten *mathematisation cycle* abgebildet wird (OECD, 2003, S. 38). Diese Problemlösekompetenz soll in Bezug auf vier Inhaltsdomänen (genannt overarching ideas) mit 85 Items erfasst werden (OECD, 2003, S. 35) und zwar Quantität (quantity), Raum und Form (space and shape), Veränderung und Zusammenhänge (change and relationships) sowie Unsicherheit (uncertainity). Die Interkorrelationen zwischen diesen latenten Dimensionen sind in der folgenden Tabelle 2 abgetragen.

Tabelle 2 Interkorrelationen zwischen den PISA-Skalen (overarching Ideas) der Studie von 2003 (OECD, 2005, S. 190)

	Veränderung und Zusammenhänge	Unsicherheit	Quantität
Raum und Form	0,89	0,88	0,89
Veränderung und Zusammenhänge		0,92	0,92
Unsicherheit			0,90

Wie ersichtlich fallen die Korrelationen zwischen den Dimensionen extrem hoch aus; bei der folgenden PISA-Erhebung stellte Mathematik keinen Schwerpunkt mehr dar, weshalb weniger Testzeit zur Verfügung stand und auf eine Aufteilung in mehrere Dimensionen im Bereich Mathematik gänzlich verzichtet wurde (OECD, 2009).

Etwas nachdenklich stimmt, dass die einzelnen Dimensionen von ihrer Benennung her recht vage erscheinen. Dazu passt der Abschnitt des PISA 2003 *Assessment Frameworks* (OECD, 2003, S. 26) der die theoretische Basis für die Mathematikerfassung darstellt und demzufolge sich die Mathematikdomäne aus den in Abbildung 2 dargestellten Komponenten zusammensetzt.

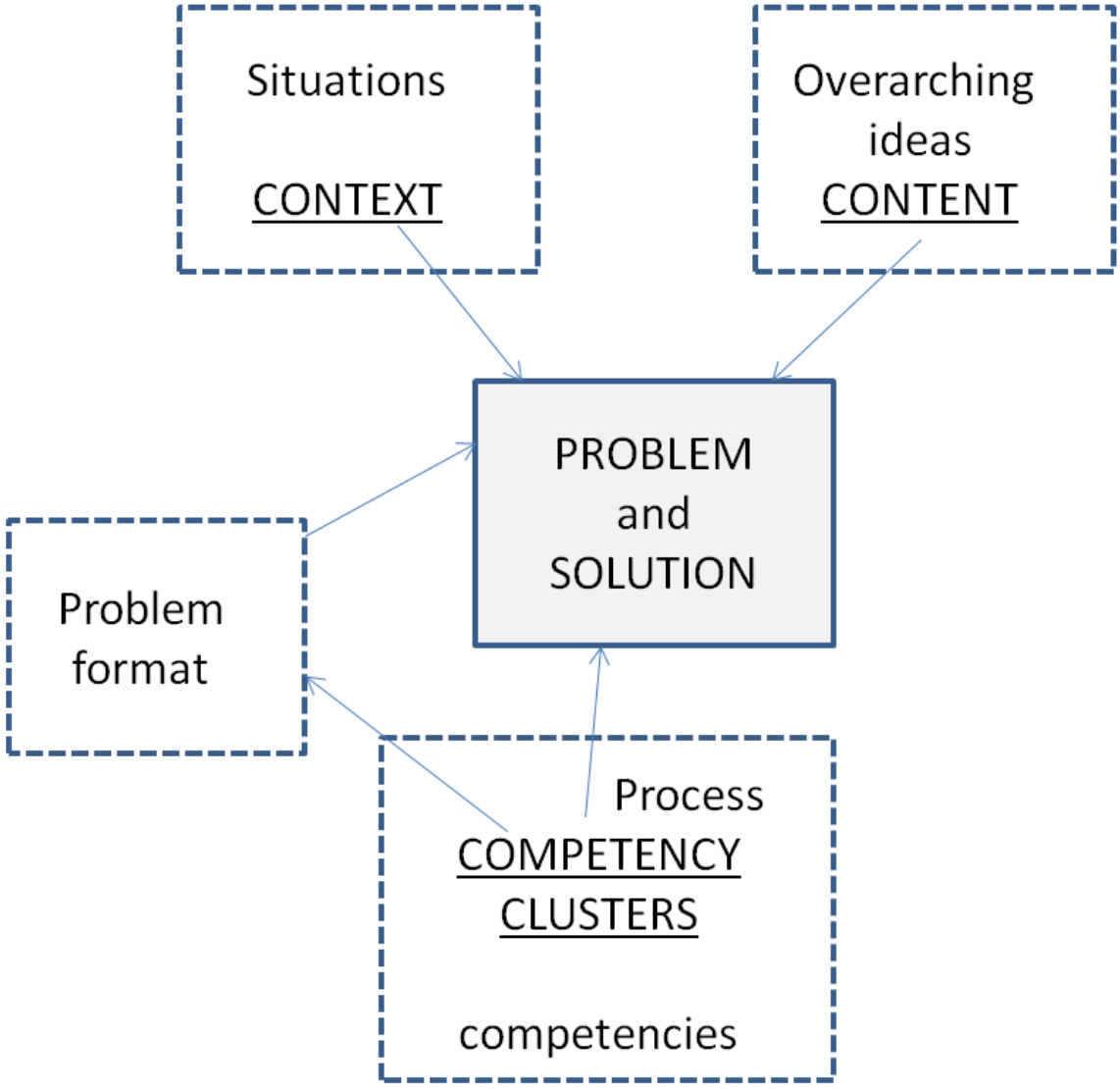


Abbildung 2 Organisation der Mathematikdomäne in PISA 2003, nach OECD (2003, S. 28)

Demnach sind es Situationen und Kontexte (z.B. persönlich, Ausbildung, öffentlich), der mathematische Inhalt (die vier Dimensionen gemäß Tabelle 2) und Kompetenzen (Reproduktion, Relation (connection), Reflektion (u.a. Problemlösen)) die zur Lösung eines Mathematik-Problem eingesetzt werden müssen (OECD, 2003). Die Beschreibung der Inhaltsbereiche, die schließlich auch die Dimensionen bildeten, erscheint teilweise recht unscharf (vgl. Jablonka, 2006, S. 160). So heißt es z.B. bei Raum und Form: „The study of shapes is closely connected to the concept of grasping space. This means learning to know, explore and conquer, in order to live, breathe, and move with more understanding in the space in which we live (Freudenthal, 1973)“ (OECD, 2003, S. 34). Wenige Zeilen später wird es konkreter mit “… also includes understanding how three-dimensional

objects can be represented in two dimension, how shadows are formed and must be interpreted…." (S. 34).

Eine der Beispielaufgaben besteht daraus, dass ein Umriss der Antarktis als Landkarte einschließlich Maßstab angegeben wird und vom Schüler geschätzt werden soll, wie groß die Fläche des Gebietes ist: „Schätze die Fläche der Antarktis, indem du den Maßstab der Karte benutzt. Schreibe deine Rechnung auf und erkläre, wie du zu deiner Schätzung gekommen bist. (Du kannst in der Karte zeichnen, wenn Dir das bei deiner Schätzung hilft.)" (OECD, 2001, S. 6).

Letztlich ist es schwierig zu beurteilen, warum einzelne Dimensionen in PISA 2003 nicht oder eben doch korrelieren sollten. Die Tatsache, dass viele Items eine große Menge an Text enthalten und die Aufgaben größtenteils in realitätsnahe Situationen eingebettet sind legt nahe, dass praktisch alle Items auf Grund des übergeordneten Konstrukts Problemlösen korrelieren (siehe auch Beginn dieses Abschnitts und Bodin, 2007, S. 31).

2.2.3 Schlussfolgerung aus der Betrachtung der TIMSS und PISA-Studien für eine psychometrische Ordnung

In Abschnitt 2.2 wurde bereits dargelegt, welche Bedeutung die Mathematik in den TIMSS- und PISA-Studien einnahm. Die Frage an dieser Stelle ist, welcher Schluss aus den Konzepten der internationalen Studien für die Psychometrie am Ende der Sekundarstufe I gezogen werden kann. In Bezug auf die Erstellung und Testung eines Strukturmodells der Mathematikfähigkeit scheint der Nutzen der Vergleichsstudien eher gering zu sein. Dies liegt wohl vor allem daran, dass es nicht das Ziel dieser Untersuchungen ist, empirisch haltbare Annahmen zur Struktur der Mathematik aufzustellen und zu prüfen.

Die Tatsache, dass gerade von Mathematikdidaktikern, wie z.B. Herrn Prof. Bender (Uni Paderborn) oder Frau Prof. Jablonka (Uni Luleå, Schweden) sehr deutliche Kritik an den TIMSS- und PISA-Konzepten geübt wird und wurde, ist sicher auch durch die unscharfe Definition dessen was PISA und TIMSS erfassen sollen begründet (Bender, 2005; Jablonka, 2005). Dies wird noch deutlicher, wenn man bedenkt, dass Heinz Rindermann 2006 in der *Psychologischen Rundschau* (provokativ) fragte „Was messen internationale Schulleistungsstudien?" (Rindermann, 2006, S. 69). Er kam zu dem Schluss, dass die Aufgaben unterschiedlicher Skalen (z.B. mathematic- versus reading literacy) einander sehr ähnlich sind, d.h. Leseaufgaben Grafiken enthalten und Mathematikaufgaben viel Text. Eine Schlussfolgerung zu der bereits Wittmann (2004), im Rahmen einer Reanalyse

von PISA-Daten, kam. Für Faktorenanalysen mit den Skalen Lesen, Mathematik und Naturwissenschaft aus Pisa 2000 und Pisa 2003 berichtet er eindeutig einfaktorielle Lösungen und kommt zu dem Schluss, dass die Ergebnisse der Schulleistungsstudien (TIMSS wie PISA) auf einen gemeinsamen G-Faktor zurückzuführen seien (Rindermann, 2006, S. 83). Dem wurde vehement von Manfred Prenzel und Kollegen – Prenzel war nationaler Projektmanager für PISA 2003 und 2006 – widersprochen. In einer Replik (Prenzel, Walter & Frey, 2007, S. 133) stellt er dazu die *Deviance* Werte einer ein- und fünffaktoriellen Lösung gegenüber. Diese Werte werden von dem IRT-Programm *Conquest* (vgl. Abschnitt 9.6) ausgegeben und können mittels χ^2-Differenztest auf Signifikanz geprüft werden (Wu, Adams, Wilson & Haldane, 2007, S. 40). Wieso er nicht berichtet, dass dieser Unterschied keine Signifikanz erreicht, d.h. das von ihm postulierte 5 Faktor-Modell keinen signifikant besseren Fit aufweist, bleibt unklar ($\chi^2_{\text{Diff}} = 12{,}28$, df = 18, d.h. p = 0,83) und spricht gegen die Annahme mehrerer Dimensionen. Hauptschwierigkeit scheint vor allem die zunehmend starke politische Färbung der PISA-Studien zu sein (ein Intelligenzvergleich, im Sinne eines G-Faktors, zwischen Ländern ist brisant) über die beispielsweise Kraus (2005, S. 119) berichtet.

Wie dem auch sei, letztlich soll (zumindest bei PISA) mathematical literacy erfasst werden (gemäß obiger Definition), um Länder in Rangreihen anzuordnen. Die darüber hinausgehende Forschung ist eher ein Begleiteffekt der Studien. Auch wenn keineswegs Einigkeit darin besteht, was mathematical literacy genau ist – der Begriff ist schon vor den PISA-Studien aufgetaucht (Julie, 2006, S. 62) – scheint die Definition der OECD durchaus wertvoll. Besonders der Aspekt einer Orientierung an der Realität, weg von inhaltlich losgelöster Mathematik, wirkt sinnvoll.

Eine weitere interessante Frage besteht darin, wo überhaupt der konzeptuelle Unterscheid zwischen TIMSS und PISA (2003) liegt. Dieser Unterschied zeigt sich auch in den Inhaltsdimensionen, die in Form von overarching ideas (siehe Abschnitt 2.2.2) in PISA sehr vage wirken, in TIMSS jedoch mit Geometrie, Algebra, Nummern sowie Daten und Zufall *(*vgl. Abschnitt 2.2.1*)* deutlich fassbarer erscheinen. Wu (2009) verglich PISA 2003 und TIMSS 2003 und kam zu dem Ergebnis, dass TIMSS deutlich mehr formale Aufgaben enthält. Sie schlussfolgerte in Bezug auf PISA: „An almost exclusive emphasis on real-life mathematics, particularly at the 15-year-old level, will likely restrict mathematics assessment to a set of items with lower mathematical content, and thus lead to an assessment that does not reflect all the mathematics topics taught in schools“ (Wu, 2009, S.

21). Diese Feststellung führt zu der Forderung, dass in dieser Arbeit beides notwendig ist, erstens die Realitätsorientierung von PISA und zweitens ein gutes Ausmaß an Überprüfung formaler Kenntnisse. Ferner wirkt der ursprüngliche TIMSS-Ansatz einer Aufteilung in Inhalte und kognitiven Anspruch, wie er auch bei Wilson (1970) vorgenommen wurde, per se viel versprechend, da er konkret und nachvollziehbar ist. Darüber hinaus erscheint – auch für das Vorhaben einer Testerstellung zu Zwecken der Leistungsmessung – die Unterscheidung von intendiertem, implementiertem und erreichtem Curriculum wie bei TIMSS (IEA, 2005; IEA, 2008), besonders in der Phase der Testkonstruktion, von Interesse.

Ein großes Problem für eine objektive Auswertung eines Leistungstests stellt die Möglichkeit von teilweise richtigen Antworten in den Vergleichsstudien dar (OECD, 2005). Dies lässt zuviel Interpretationsspielraum und sollte in Hinblick auf die Testobjektivität (Lienert & Raatz, 1994; Moosbrugger & Kelava, 2008) keinesfalls übernommen werden. Bei PISA und bei TIMSS wurde darüber hinaus ein so genanntes *Multidimensional Random Coefficients Multinomial Logit Model* (MRCML), eine Verallgemeinerung des Rasch-Modells, eingesetzt (IEA, 2004a; OECD, 2003). Das Modell ermöglicht es unter anderem, mehrdimensionale Konstrukte (im Falle von Pisa Literalitäten) zu skalieren bzw. zu modellieren. Ein Vorteil des Modells entsteht vor allem dann, wenn die verschiedenen Dimensionen stark untereinander korrelieren (Adams, Wilson & Wang, 1997). Es wäre interessant, einen neuen Leistungstest für Schüler am Ende der Sekundarstufe I auch anhand dieses Modells zu begutachten. Die latenten Korrelationen zwischen verschiedenen Inhaltsbereichen ließen sich dadurch gut mit den bereits berichteten Ergebnissen der internationalen Studien vergleichen.

3 Theoretische Strukturierung von Mathematikfähigkeit

In diesem Abschnitt wird versucht eine theoretische Grundlage für die Konstruktion eines Mathematiktests zu erstellen. Hierbei sollen Überlegungen aus dem Bereich der pädagogischen- und diagnostischen bzw. differentiellen Psychologie einfließen.

3.1 Intelligenzdiagnostische Überlegungen zur Ordnung von Mathematik

Nach wie vor gilt Intelligenz einerseits als Schlüsselmerkmal für Berufserfolg und andererseits ist trotz etwa 100 Jahren Forschung noch kein wirklicher Konsens über die Definition von Intelligenz zustande gekommen (Jensen, 1998; Süß, 2003). Dies geht so

weit, dass teilweise in Lehrbüchern innerhalb eines Abschnittes zur Intelligenz verschiedene Definitionen aufgestellt werden (siehe z.B. Schweizer, 2006, S. 2). Auch ein Versuch des *Board of Scientific Affairs* der *American Psychological Association* (Neisser et al., 1996) mit dem Titel *Intelligence: Knowns and Unknowns* ging im Wesentlichen nicht über eine Darstellung der aktuellen Konzepte verschiedener Forscherkreise hinaus. Eine präzise Definition sucht man auch dort vergeblich. Hierzu passt Dearys Feststellung in der es heißt (Deary, 2000, S. 2): „Incidentially, luminaries in the area of Intelligence have felt the need to slay the definition dragon at the start of their accounts… All refused to be halted by demands for an exact meaning-style definition, deciding that there was a sufficient corpus of research findings to be described …“

Dies ändert nichts daran das, wie Hülsheger, Maier, Stumpp und Muck (2006) für den deutschen Sprachraum anhand einer Meta-Analyse zeigten, Intelligenztests eine gute Vorhersage von Ausbildungserfolgen, operationalisiert durch Noten und Beurteilungen, liefern (korrigiertes r von 0,48 und 0,54). Dies schließt nahtlos an die Ergebnisse von Schmidt-Atzert, Deter und Jaeckel (2004) an, die mit einem G-Maß (Faktorwerte des ersten unrotierten Faktors der eingesetzten Tests) den Ausbildungserfolg vorhersagten und für die theoretischen Kenntnisse Validitäten zwischen $r = 0{,}31$ und $r = 0{,}44$ vorfanden (Ausnahme: Kaufleute für Bürokommunikation mit $r = 0{,}09$), die sich jedoch in einigen Fällen noch durch spezifische Tests erhöhen ließen.

Darüber hinaus gibt es einen deutlichen Zusammenhang zwischen der Leistung im Fach Mathematik, sei es operationalisiert durch Mathetests, Schulnoten oder Lehrerurteile und diverseren Intelligenztests (Holling, Preckel & Vock, 2004). Für die bereits in Abschnitt 2.1 vorgestellten Mathetests liegen lediglich für den MATLUB (Ibrahimovic & Bullheller, 2005) mir $r = 0{,}69$ (Intelligenz-Struktur-Analyse) und den Rechentest RT8+ mit $r = 0{,}65$ (nach Frey, 1973, LPS) tatsächliche Daten vor.

In Bezug auf die Mathematiknote zeigten z.B. Wittmann und Süß (1997), dass der deutlichste Zusammenhang zum Berliner Intelligenzstruktur-Test (siehe Abschnitt 3.1.5) auf Ebene der 12 Einzelzellen mir $R_{\text{adjustiert}} = 0{,}61$ vorliegt, da dort die Prädiktor-Kriteriensymmetrie am höchsten ausfällt. In Bezug auf die Zellenebene ergab sich der größte Zusammenhang zwischen Mathematiknote und Verarbeitungskapazität ($r = 0{,}52$, $N = 137$) gefolgt von numerischer Intelligenz ($r = 0{,}42$, $N = 137$). Süß (2001) fasst schließlich zahlreiche Befunde, einschließlich Meta-Analysen, zusammen und schreibt: „Schulnoten gehören zu den am häufigsten verwendeten Kriterien für

Intelligenzleistungen, und es ist unmittelbar plausibel, dass sich Intelligenztests bei diesen Vorhersagen bewähren müssen.“ (S. 129).

Aufgrund dieser Nähe von Mathematikleistung und Intelligenztestleistung erscheint es in dieser Arbeit sinnvoll, einen Blick auf die bekanntesten Konzepte der Intelligenzdiagnostik zu werfen und theoretische Modelle der Intelligenzdiagnostik zu identifizieren, die eine Hilfe bei der Strukturierung von Mathematikfähigkeit liefern können. Im folgenden werden daher zunächst einige bekannte Konzepte dargestellt und anschließend auf gebräuchliche Intelligenztests und ihren Bezug dazu eingegangen. Aufgrund der Vielzahl von Ansätzen ist eine erschöpfende Darstellung nicht möglich und auch nicht angestrebt. Einige Modelle sind aus Sicht des Autors dieser Arbeit schlicht ungeeignet, um im Kontext der Psychometrie der Mathematik von größerem Nutzen zu sein oder weisen sonstige gravierende Schwächen auf. Ein Beispiel für ersteren Fall stellt Guilfords *Structure-of-Intellect* Konzeption (Guilford, 1967) dar die aufgrund ihrer Umfänglichkeit (Guilfords Fähigkeitswürfel enthält 120 Teilfähigkeiten) den wesentlichen Aspekt eines Modells – nämlich ein vereinfachtes Abbild der Realität darzustellen – nicht ausreichend erfüllt. Ein weiteres, problematisches Beispiel findet sich in Form von Cyrill Burts Intelligenztheorie (Burt & Howard, 1956), da sich im Nachhinein herausstellte, dass Burt systematisch Daten gefälscht hat (Hearnshaw, 1979) was, wie es Vernon (1979, Vorwort) ausdrückt, seine Befunde „worthless“ werden lässt.

3.1.1 Thurstones primary abilities

Thurstone (1938) war der Meinung, dass sich Intelligenz nicht ausreichend durch ein G-Faktormodell erklären lässt und stellte mit seiner Methode der Faktorenanalyse eine Theorie der *primary mental abilities* auf, die allgemeiner als Spearmans (1904) spezifische Faktoren aber weniger allgemein als sein G-Faktor sein sollten. Bei den sieben primary abilities handelt es sich um Wortflüssigkeit, verbales Verständnis, schlussfolgerndes Denken, räumliches Vorstellungsvermögen, Merkfähigkeit, Rechenfähigkeit und Wahrnehmungsgeschwindigkeit (Thurstone, 1938, Übersetzung durch den Autor). Thurstone wählte entgegen seiner eigenen Überzeugung in der Ursprungsarbeit eine orthogonale Rotationsmethode wie er in seiner Autobiographie klarstellt: „Although my first text on multiple-factory analysis, The Vectors of Mind, had previously been published (1935), with a development of the concepts of communality, the rotation of axes, and the use of oblique axes, I hesitated to introduce all of these things in the first experimental

study. ... Instead of proceeding according to my convictions, that first factor study was published with the best fitting orthogonal frame, although we knew about more complete methods.“ (Thurstone, 1952, S. 316). Letztlich sieht er selbst seine sieben Primärfähigkeiten als korreliert an, was überhaupt erst die – von ihm selbst betriebene (vgl. Thurstone, 1944) - Extraktion Faktoren zweiter oder gar dritter Ordnung ermöglicht.

3.1.2 Cattells Theorie fluider und kristaliner Intelligenz

Horn und Cattell (1966) unterscheiden in dieser Theorie zwischen der überwiegend genetisch bedingten fluiden Intelligenz und der kristallinen Intelligenz, die kulturabhängig ist und bis zum Lebensende stabil bleibt oder gar ansteigt. Als eine Erweiterung postulierte Cattell (1987, S. 138) seine Investmenttheorie, die davon ausgeht, dass die bereits genetisch determinierte fluide Intelligenz, sowie Motivation und Qualität der Lerngelegenheiten entscheidend für den sukzessiven Aufbau kristalliner Intelligenz sind. In einem seiner letzten Artikel beschreibt Cattell (1998) seine Betroffenheit darüber, dass sowohl er, als auch Eysenck, Jensen und Herrnstein bereits physisch angegriffen wurden, als sie auch nur über die mögliche Erblichkeit von fluider Intelligenz referierten. In den letzten Jahren scheint es dennoch eine gewisse Evidenz dafür zu geben, dass auch fluide Intelligenz zu einem bedeutsamen Maß trainierbar ist, indem das Training anhand von Arbeitsgedächtnisaufgaben vorgenommen wird (Jaeggi, Buschkuehl, Jonides & Perrig, 2008; Sternberg, 2008). Eine Theorie, die sich explizit an Cattell orientiert und eine Weiterentwicklung darstellen möchte, ist die PPIK-Theorie (process, personality, interest, knowledge) nach Ackerman (1996).

3.1.3 Jägers Facettenmodell

Der Vorläufer von Jägers Facettentheorie der Intelligenz findet sich bereits in dessen Habilitationsschrift aus dem Jahre 1967, wo er zwischen anschauungsgebundenem, zahlengebundenem, und sprachgebundenem Denken sowie Einfallsreichtum, Merkfähigkeit, Konzentrationskraft/Tempo-Motivation und Verarbeitungskapazität unterschied (Jäger, 1967, S. 179). Aufgrund von Schwierigkeiten in mehreren Datensätzen, die drei Bereiche verbale, figurale und numerische Intelligenz als Faktoren zu extrahieren, entwickelte A. O. Jäger seine Facettentheorie (1982), die in Abbildung 3 dargestellt ist.

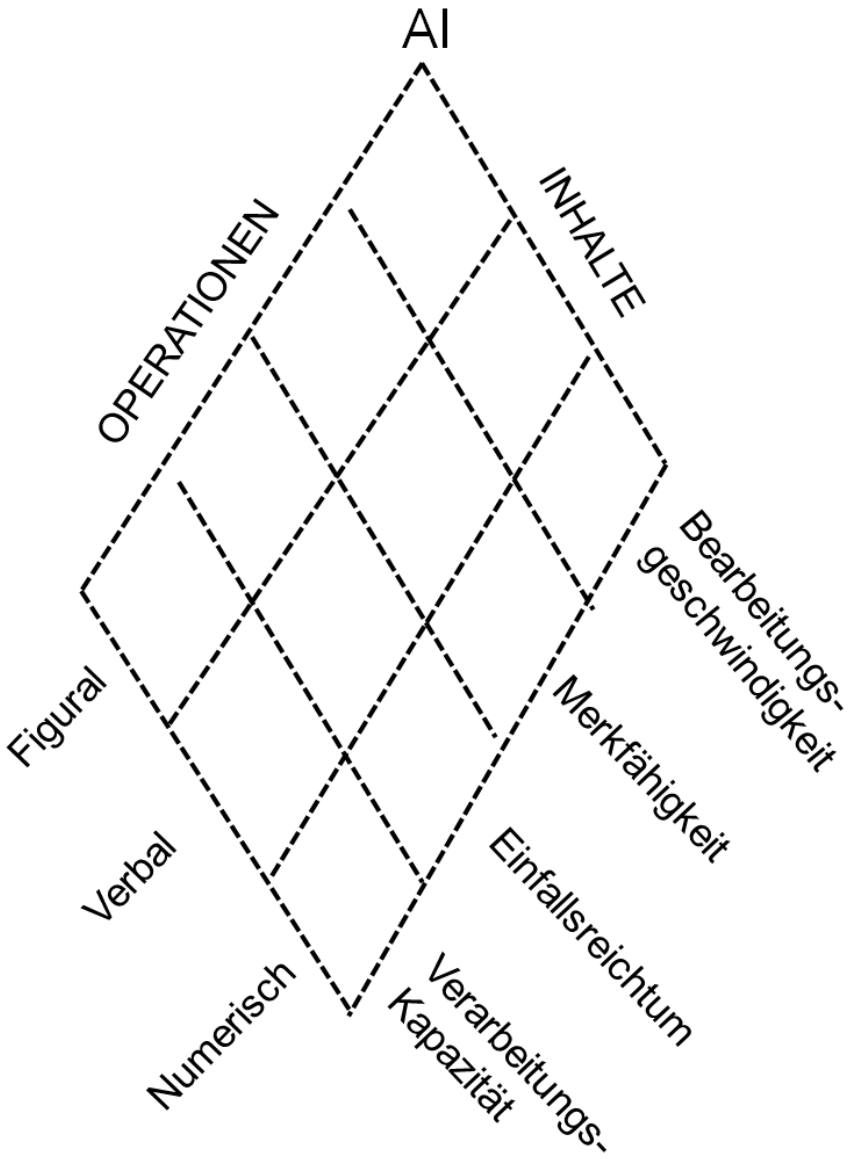


Abbildung 3 Facettenmodell der Intelligenz nach Jäger (1982)

Der Grundgedanke besteht darin, dass durch Parceling – eine theoriegeleitete Aggregation der Einzelitems zu Miniskalen (genaueres siehe Abschnitt 4.3.1) - entlang der Inhaltsfacetten (figural, verbal, numerisch) und eine anschließende Faktorenanalyse dieser Miniskalen, die drei Inhaltsfaktoren deutlich zutage treten und bei einer Bündelung entlang der Operationsfacetten (Bearbeitungsgeschwindigkeit, Merkfähigkeit, Einfallsreichtum, Verarbeitungskapazität) die vier operativen Faktoren (Jäger, 1982). Neben diesen insgesamt sieben Skalen (oder Faktoren) sieht das Modell eine Art G-Faktor (AI, Allgemeine Intelligenz) als Aggregat aller Aufgaben vor. Hervorzuheben ist die besonders für Forschungszwecke interessante Möglichkeit, Vorhersagen unter dem Blickwinkel von Symmetrie von Prädiktor und Kriterienseite (Wittmann, 1985, 1988) auf der jeweils sinnvollsten Aggregationsebene durchzuführen, d.h. auf Zellen-, Facetten- oder AI-Ebene.

Das BIS-Modell ist nach Brocke und Beauducel (2001, S. 28) vor allem durch Aufgaben-Integration (Herstellung eines Aufgabenpools, der für alle in der Intelligenzforschung verwendeten Aufgaben maximal repräsentativ ist) entstanden. Ob es sich letztlich um ein integratives Modell (Integration verschiedener Strukturmodelle durch meta-analytische, gemeinsame Auswertung) handelt, ist demnach streitbar (Brocke & Beauducel, 2001, S. 29).

3.1.4 Die Zwei-Faktoren-Theorie und integrative Modelle

Die Zwei-Faktoren-Theorie geht auf den britischen Psychologen Charles Spearman zurück, der 1904 durch die Analyse von Korrelationen zwischen unterschiedlichen Testleistungen in den verschiedensten Inhaltsbereichen zu folgendem Schluss kam: „All branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the others.“ (Spearman, 1904, S. 283).

Er unterschied also letztlich zwischen einer Intelligenzkomponente, die allen Aufgaben gemeinsam ist (G-Faktor) und weiteren Komponenten, die spezifisch für jede Aufgabe sind. Abbildung 4 verdeutlicht dieses Prinzip mit Hilfe von Ballantines.

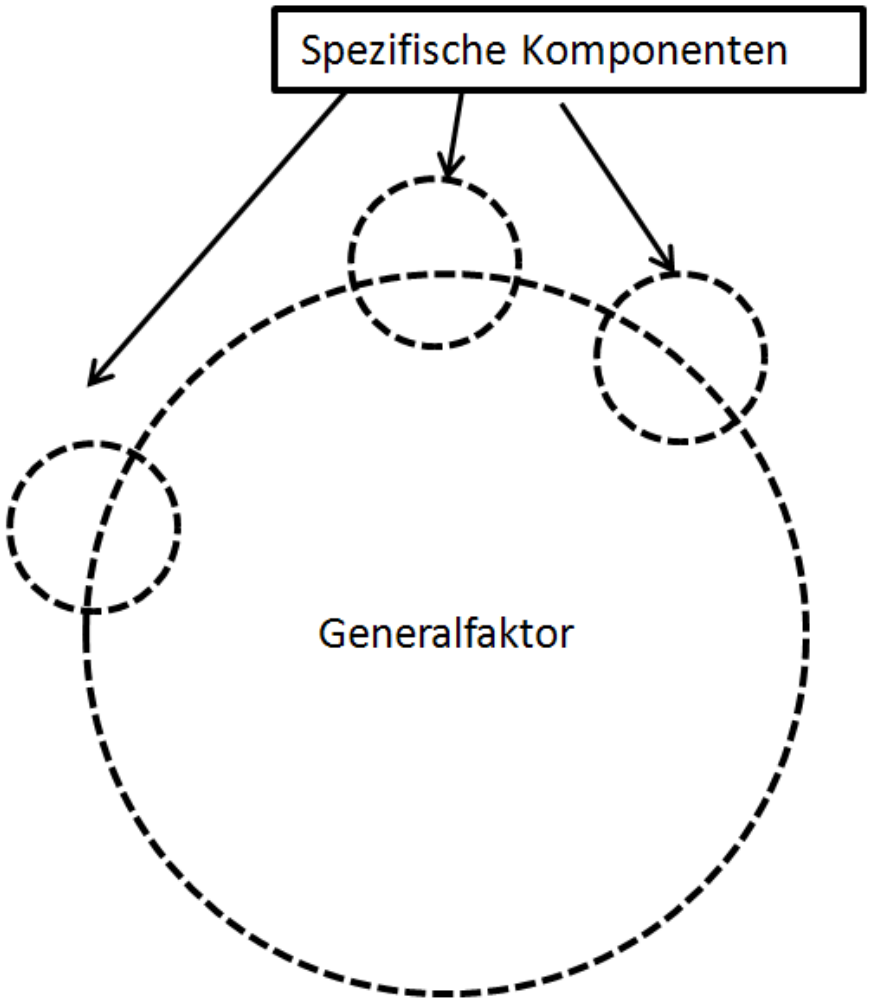


Abbildung 4 Intelligenzmodell nach Spearman (1904).

Nach Jensen (1998, S. 89) haben alle IQ-Tests als charakteristische Eigenschaft gemeinsam, starke Ladungen auf dem so genannten G-Faktor zu produzieren. Geht man von den weiter oben beschriebenen Eigenschaften von Tests nach Spearman aus, müsste mit steigender Anzahl von Aufgaben(gruppen) der Gesamtscore sukzessive mehr G-Komponenten enthalten, als spezifische Aufgabenanteile. Dies passt zu dem verbreiteten Vorgehen, bei Intelligenztests neben Subscores (für wie auch immer geartete Aufgabengruppen) einen Gesamtscore anzugeben, der meist die allgemeine Leistungsfähigkeit beschreiben soll.

Es ist von großer Bedeutung, dass die G-Faktor-Theorie keineswegs mit den bisher erläuterten Modellen unvereinbar sein muss, wie die Überschrift dieses Abschnitts bereits zum Ausdruck bringen soll. Es war Carroll, der 1993 eine (faktorenanalytische) Analyse von 461 Datensätzen aus 19 Ländern vornahm, die letztlich zu seiner *Three-Stratum-Theorie* kognitiver Fähigkeiten führte. Diese Theorie ordnet kognitive Fähigkeiten in drei

Ebenen unterschiedlicher Generalität (Stratum I, II und II), mit einem G-Faktor als höchsten Maß. Wittmann und Süß (1997) ziehen aus seinen Befunden und eigenen Arbeiten die logische Konsequenz, dass diese hierarchische Ordnung die auf den ersten Blick widersprüchlichen Theorien vereint: „Although debates will go on what the best hierarchical model of intelligence is, the very fact that g is there but also different group factors no longer needs to be challenged.“ (S. 5). Auch aus Sicht von Vernon (1979, S. 61), der ein hierarchisches Intelligenzmodell vorschlägt, ist eine Kombination von (korrelierten) Gruppenfaktoren mit *g* an ihrer Spitze sinnvoll. Letztlich ist es das Aggregationsniveau von der Einzelaufgabenebene über Gruppenfaktoren bis hin zu Gesamtscores, das hilft die Zweifaktorentheorie und die vorangegangenen Konzepte zu vereinen. Sternberg und Powell (1982) haben diese Erkenntnis in ihrem Evolutionsmodell der Intelligenztheorien zusammengefasst. Dieses Modell geht von drei Evolutionsstufen aus, beginnend mit Stufe I, auf der monoistische Theorien (Ia; G-Faktormodelle) und pluralistische Theorien mit vielen gänzlich unabhängigen Einheiten (Ib; z.B. nach Thorndike die Anzahl der S-R Verknüpfungen) dominieren. Der nahe liegende Konflikt dieser Theorien wird auf Stufe II durch ein hierarchisches Modell (IIa, z. B. Vernon), das einen Generalfaktor an der Spitze vorsieht, jedoch die Eigenständigkeit der darunter liegenden Faktoren betont, und ein non-hierarchisches Modell (IIb, z. B. Thurstone, unter der Annahme die primary abilities seien korreliert), das eine Überlappung der einzelnen Gruppenfaktoren zulässt, aufgehoben. Entscheidend für die (bisher) höchste Evolutionsstufe (Stufe III) ist nach Sternberg und Powell (1982, S. 988) die Kombination des hierarchischen Ansatzes von IIa und der Überlappung zwischen den Gruppenfaktoren gemäß IIb.

Die Zuordnung zu einzelnen Stufen kann leider nicht immer völlig eindeutig vorgenommen werden (wie z.B. bei Thurstone). Aus Sicht des Autors dieser Arbeit könnte das bereits erwähnte BIS-Modell von Jäger (1982) am ehesten die höchste Evolutionsstufe (Stufe III nach Sternberg & Powell, 1982) darstellen. In dieser Arbeit soll für Mathematik auf einer Generalitätsebene ähnlich aller in Abschnitt 2.1 vorgestellten Tests eine Struktur theoretisch begründet und empirisch geprüft werden. Dass sich auf einer noch allgemeineren als der hier anvisierten Ebene der Mathematikfähigkeiten ein G-Faktor befinden kann, wird demnach nicht ausgeschlossen.

3.1.5 Verbreitete Intelligenztests

Eine Testreihe mit dem Ziel den Spearmanschen G-Faktor zu erfassen wurde von John Raven entwickelt, einem Schüler Spearmans (Casé, Neer & Lopetegui, 2003; Jensen, 1998, S. 38). Ähnliche Tests, speziell für den deutschsprachigen Raum, existieren zum Beispiel in Form der Wiener Matrizen-Tests von Formann und Piswanger (1979), der einige identische Aufgaben enthält. Eine detaillierte, prototypische, Analyse eines Tests der Bongard-Figurenmuster zur Erfassung eines G-Faktors verwendet findet sich bei Jasper (2007). Inwiefern oben genannte Tests fluide Intelligenz nach Cattell (1998) oder den spearmanschen G-Faktor (1904) erfassen, ist wohl eher eine Frage der Interpretation der Theorien die - zumindest im Falle Spearmans - eher vage formuliert und hier nicht Gegenstand der Betrachtung sind. Das zeigt sich auch am Manual der deutschen Variante des Culture Fair Tests (Cattell & Weiß, 1971), in dem einerseits postuliert wird der CFT-3 erfasse hauptsächlich fluide Intelligenz (S. 18) und gleichzeitig explizit betont wird, Ziel sei es mit dem Test den spearmanschen G-Faktor zu erfassen (S. 6).

Die bekannten Hamburg-Wechsler-Intelligenztests (z.B. HAWIK-III, Tewes, Rossmann & Urs, 1999) haben das Ziel, ein breites Konstrukt zu erfassen, sind jedoch nicht ausnahmslos einer der bisher aufgezählten Theorien zuzuordnen. Für Wechsler war Intelligenz „…die zusammengesetzte oder globale Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen“ (Wechsler, 1961, S. 13). Die von ihm nach diesem Konzept entwickelten Intelligenztests werden demnach auch von Horn und Noll (1994, S. 163) als auf Mischtheorien der Intelligenz basierend angesehen. Die neueste deutsche Variante für Erwachsene aus dem Jahre 2006 (Aster, Neubauer & Horn, 2006) entspricht vom Intelligenzkonzept dem HAWIK-3 (Tewes et al. , 1999) mit seiner Einteilung in Verbal- und Handlungsteil, die von manchen Autoren jedoch als völlig überholt bezeichnet wird (Jacobs & Petermann, 2007; Schweizer Verband für Berufsberatung, 2006). Für den neuesten Test der Wechsler-Reihe, den HAWIK-4, gibt es keine solche Einteilung mehr und nur noch die vier Bereiche Sprachverständnis, wahrnehmungsgebundenes Denken, Arbeitsgedächtnis und Verarbeitungsgeschwindigkeit sowie einen Gesamt-IQ (Petermann & Petermann, 2008).

Das Leistungsprüfsystem ist auf der Untertestebene eindeutig von Thurstones Primärfaktorentheorie beeinflusst, so ordnet sein Autor im Manual den Großteil der 15 Untertests explizit den entsprechenden Fähigkeiten nach Thurstone zu (Horn, 1983).

Gleichzeitig stellt Horn (1983) nur wenige Seiten später selbst ein Pyramidenmodell zur Struktur der Begabung auf, das die unter Abschnitt 3.1.1 bis 3.1.4 beschriebenen Ansätze integrieren soll.

Auch der Wilde-Intelligenztest (in zweiter, kaum veränderter Auflage) bezieht sich ausdrücklich auf Thurstones sieben Primärfaktoren, die mit neun Subtests erfasst werden sollen (Jäger & Althoff, 1983). Jäger (1997) sieht die Anwendung des WIT an sich jedoch, unter anderem wegen der aus seiner Sicht nur vagen Anlehnung an diese Intelligenztheorie (keine rationale Testkonstruktion) und völlig veralteten Normen (von 1963) als nicht gerechtfertigt. Dem gegenüber basiert der WIT-2 von 2008 (Kersting, Althoff & Jäger, 2008) auf einem modifizierten Modell der primary mental abilities (MMPMA). Modifikationen betreffen die Annahme hierarchischer Stufen (ähnlich Carrolls Strata) und die Nutzung von Jägers Facettenansatz (vgl. Abschnitt 3.1.3), um einzelne Varianzkomponenten zu akzentuieren und andere abzuschwächen. Von den ursprünglichen 15 Subtests des WIT wurden 7 beibehalten bzw. überarbeitet und vier komplett neue Skalen hinzugefügt (z.B. E-Mails bearbeiten und Wissen-Informationstechnologie) (Kersting et al., 2008). Auch der WIT-2 orientiert sich an einigen von Thurstones Primärfaktoren. So weist das Manual explizit darauf hin, dass die Module sprachliches Denken, rechnerisches Denken und räumliches Denken direkt den Faktoren verbal comprehension, numerical ability und spatial ability gemäß Thurstone (1938) zuzuordnen sind (Kersting et al., 2008, S. 29).

Der Berliner Intelligenzstrukturtest ist seit 1997 (Jäger, Süß & Beauducel, 1997) erhältlich und stützt sich auf das unter Abschnitt 3.1.3 vorgestellte Modell Jägers.

Der I-S-T 70 von Amthauer (1973) ist ein Test, der neun Einzelskalen (sowie einen Gesamtscore) erfasst, die von ihrer Benennung teilweise an Thurstones primary abilities (z.B. Zahlenreihen ≈ Rechenfähigkeit oder Würfelaufgaben ≈ räumliches Vorstellungsvermögen) erinnern. Doch liegt dem Test kein explizites Intelligenzmodell zugrunde; so heißt es im Manual unter dem Abschnitt „Was wird mit den Aufgabengruppen des I-S-T untersucht?“ (Amthauer, 1973, S. 39), dass Korrelationen mit Außenkriterien und Faktorenanalysen entscheidend waren, die jedoch nicht berichtet werden. Der sehr ähnliche Vorgänger I-S-T basiert laut Manual, in dem jegliche konkreten Hinweise zu Entwicklung fehlen, anscheinend auf gar keiner Theorie (Amthauer, 1953). Wie Brocke, Beauducel und Tasche (1998) zeigten, sind die konkreten Prinzipien der Testkonstruktion, zum Beispiel Maximierung der Korrelation von Aufgabengruppen zum Gesamtscore und gleichzeitige Minimierung der Korrelation zwischen den

Aufgabengruppen nach heutigem Maßstab indiskutabel (bzw. widersprüchlich) und es lassen sich eigentlich nur figurale und verbale Fähigkeiten als Einheiten extrahieren. Überraschend ist in diesem Zusammenhang, dass der IST-70 bei Hogrefe (Stand: 18.05.2009) nach wie vor bestellt werden kann. Aufgrund diverser methodischer Mängel wurden der I-S-T 2000 und schließlich der I-S-T 2000 R entwickelt, der auf einem so genannten hierarchischem Rahmen – bzw. Protomodell der Intelligenzstrukturforschung (HPI) basiert (Amthauer, Brocke, Liepmann & Beauducel, 1999; Liepmann, Beauducel, Brocke & Amthauer, 2007). Sowohl das bereits erwähnte Modell von Carroll (1993), als auch das Radex-Modell von Guttmann (1957), Jägers Modell (Jäger, 1982), das Cattell-Horn Modell (Horn, 1983) und natürlich Thurstones Ansatz (1938) sollen nach Liepmann, et al. (2007) spezielle Formen des HPI darstellen. Dies wird z.B. von Schmidt-Atzert in einer Testrezension kritisiert, der meint „das hierarchische Rahmen bzw. Protomodell der Intelligenzstrukturforschung (HPI) ist mindestens so schwer zu verstehen wie sein Name vermuten lässt“ (Schmidt-Atzert, 2002, S. 54) und den unklaren Zusammenhang zwischen den Skalenbildungen und theoretischen Überlegungen anprangert. Der I-S-T 2000 R besteht aus einem Grund- und einem Erweiterungsmodul; das Grundmodul bietet als Maße figurale, verbale sowie numerische Intelligenz und das Erweiterungsmodul verbal, numerisch und figural kodierte Wissensmaße, als auch Kennwerte für fluide/kristalline Intelligenz und einen Gesamtwert für Wissen (Liepmann et al., 2007). Darüber hinaus besteht die Möglichkeit, Merkfähigkeit und schlussfolgerndes Denken mit Wissensanteilen zu erfassen.

3.1.6 Schlussfolgerung aus Betrachtung von Intelligenztests und Konzepten: Skalenkonzeption

Nach einer Sichtung unterschiedlicher Intelligenzmodelle- und Tests in den vorherigen Abschnitten werden folgende Schlüsse gezogen:

1. Die Einteilung in figurale, verbale und numerische Intelligenz zieht sich durch diverse Intelligenztests (z.B. IST-2000R und BIS sowie WIT-2) und wurde somit vielfach repliziert.
2. Viele Intelligenztests enthalten Mathematikaufgaben, wie z.B. Zahlenreihen und Rechenzeichen (z.B. BIS und IST-2000R), oder die Skala *Rechnerisches Denken* im WIT-2, die zwar sehr basale Fähigkeiten erfassen, jedoch eindeutig Mathematikaufgaben darstellen (Kersting et al., 2008).

3. Es scheint einen deutlichen Zusammenhang zwischen der Leistung in Mathematik als Schulfach und in Form von spezifischen Tests mit diverseren Intelligenztests zu geben (vgl. Abschnitt 3.1).
4. Es ergab sich aus Sichtung aller verfügbaren Mathetests (vgl. Abschnitt 2.1), dass Dimensionen der Mathematik (wie auch immer sie aussehen mögen) deutlich korreliert sind.

Diese vier Erkenntnisse legen es nahe, erstens ein Modell deutlich korrelierter Dimensionen der Mathematikfähigkeit anzunehmen und zweitens bezüglich Skalenkonstruktion eine Orientierung an figuraler, verbaler und numerischer Intelligenzdiagnostik vorzunehmen. Zusammen mit den bisherigen Schlussfolgerungen aus internationalen Vergleichsstudien – insbesondere zur mathematischen Literalität (vgl. Abschnitt 2.2.3) – werden im folgenden nun vier Skalenbeschreibungen aufgestellt.

3.1.6.1 Verbale Mathematikfähigkeit: Mathematische Literalität

Mathematische Literalität soll sich an den im Rahmen der PISA-Studien verwendeten Begriff der mathematical-literacy anlehnen (OECD, 2003), wobei in deutschen Veröffentlichungen häufig von mathematischer Grundbildung die Rede ist. Die Definition des Begriffs nach OECD in deutschen Veröffentlichungen lautet: „…Fähigkeit definiert, die Rolle, die Mathematik in der Welt spielt, zu erkennen und zu verstehen, begründete mathematische Urteile abzugeben und sich auf eine Weise mit der Mathematik zu befassen, die den Anforderungen des gegenwärtigen und künftigen Lebens einer Person als konstruktiven, engagierten und reflektierenden Bürger entspricht.“ (OECD, 2001, S. 19). Es sollen hier Aufgaben verwendet werden, die allesamt eine Einkleidung in möglichst alltagsnahe Sachverhalte enthalten. So soll es nötig sein, aus einer vorgegebenen Grafik Werte zu entnehmen, die zur Lösung der Aufgabe zwingend erforderlich sind oder aus einer Tabelle Zahlen abzulesen und zu verwenden.

Dennoch gibt es deutliche Unterschiede zwischen dem Begriff gemäß OECD und der hier angewendeten Konzeption. Zum Beispiel soll hier stets nur eine einzige korrekte Lösung existieren und keine halb- oder teilweise richtigen Lösungsmöglichkeiten. Auch muss die Breite der abgedeckten Themenbereiche - allein wegen der nötigen Beschränkung der Testzeit – wesentlich geringer sein. Neben der schon angesprochenen Alltagsnähe ist ein großer Textanteil typisch und erwünscht, weshalb solche Aufgaben weitestgehend klassischen Textaufgaben (wie sie auch in der Schule Verwendung finden) entsprechen.

Vom lateinischen littera (=Buchstabe) abgeleitet ist hier mit Mathematischer Literalität gemeint, dass sowohl die Sprache an sich als auch die Mathematik (gewissermaßen als formale Sprache betrachtet) beide unverzichtbare Kommunikationsmittel in der heutigen Berufswelt darstellen (vgl. Kaiser & Schwarz, 2003). Der Zusammenhang mit Intelligenzdiagnostik ergibt sich daraus, dass mathematische Sachverhalte in Form von verbalen Beschreibungen (= Oberflächenmerkmal) präsentiert werden.

3.1.6.2 Figurale Mathematikfähigkeit: Geometrie und grafische Funktionen

Am ehesten leuchtet wohl die Analogie von Geometriefaufgaben und der figuralen Intelligenzkomponente ein. Neben klassischen Geometrieaufgaben, wie dem Berechnen des Volumens von Körpern, eines fehlenden Winkels in rechtwinkligen Dreiecken und Grundlagen der Trigonometrie (Sin, Cos, Tan, etc.) enthält dieser Bereich ausdrücklich auch grafische Darstellungen von Funktionen. Das entscheidende Merkmal wenn Funktionen Inhalt dieser Skala sind, ist dass ihre grafische Darstellung wichtig für die Lösung der Aufgabe ist, also z.B. Punkte aus Grafiken abgelesen werden müssen und somit eine Orientierung im kartesianischen Koordinatensystem notwendig ist. Verwendete Grafiken sind zur Lösung der Aufgabe zwingend erforderlich oder ermöglichen zumindest eine deutliche Vereinfachung der Lösung. Die visuell / räumliche Komponente ist bei all diesen Aufgaben von Bedeutung. Die inhaltliche Einkleidung ist eher schlicht gehalten und der Instruktionstext auf das Nötigste beschränkt.

3.1.6.3 Numerische Mathematikfähigkeit I: Prozedurales Rechnen

Prozedurales Rechnen deckt einfache Rechenaufgaben ab, deren Durchführung weitestgehend automatisiert ablaufen sollte. Die Voraussetzung zur Lösung seitens einer Person ist das Grundverständnis des jeweiligen Rechenprinzips (Wissen). Weiß eine Testperson prinzipiell wie multipliziert und dividiert wird, so stellt eine reine Divisionsaufgabe keine große Herausforderung mehr dar. Vielmehr geht es dann nur noch darum, bekannte Lösungsschritte abzuarbeiten. Soll beispielsweise umgerechnet werden, wie viel Meter 7 Kilometer darstellen, ist eine Mischung aus Wissen (1km gleich 1000m) und Berechnung (7 mal 1000 gleich 7000) notwendig. Darüber hinaus enthält dieser Bereich die Abfrage reinen Wissens. Es geht also um (wichtige) handwerkliche Fähigkeiten, die in vielen Bereichen des Berufslebens gefordert werden. Ebenso wie für Geometrie und grafische Funktionen sollten die Aufgaben möglichst wenig Text enthalten.

Rein oberflächlich betrachtet weisen die Aufgaben deutliche Ähnlichkeit mit der folgenden Dimension auf. Die Betonung von Zahlen, Rechenzeichen, Standardoperationen und Formeln stellt hier den Zusammenhang mit der numerischen Intelligenzkomponente dar.

3.1.6.4 Numerische Mathematikfähigkeit II: Komplexes Rechnen

Häufig ist für die Lösung von Aufgaben dieser Skala das Beherrschen von reinen Rechenaufgaben eine Vorraussetzung (prozedurales Rechnen, vorheriger Abschnitt). Es geht am ehesten um fortgeschrittene Algebra einschließlich Funktionen, bei der meist mehrere Variablen gleichzeitig beachtet werden müssen (z.B. x und y). Es müssen zum Beispiel einfache Gleichungssysteme gelöst oder Terme umgeformt werden. Die durchzuführenden Operationen bestehen aus mehreren Teilschritten. Dieser Aspekt hat in dem Sinne Ähnlichkeit mit Problemlöseaufgaben als dass nicht sofort klar ist, auf welche Art und Weise man zu der Lösung gelangt (Eysenck & Keane, 2005, S. 442; Hussy, 1998). Die Aufgabe muss analysiert werden und es ist nötig sich einen Lösungsweg zu überlegen. Hier ist ein deutlicher Unterschied zwischen verschiedenen Klassenstufen und Schultypen zu erwarten. Neben der Betonung von Zahlen und Formeln lässt sich hier auch ein hoher Zusammenhang zur Verarbeitungskapazität – analog der Forschung zum komplexen Problemlösen – (Wittmann & Hattrup, 2004) erwarten. Generell sollten Aufgaben einer solchen Skala möglichst sprachfrei sein.

3.2 Taxonomien zur Ordnung von Mathematikfähigkeit

Der Begriff Taxonomie lässt sich ableiten vom griechischen táxis = Ordnung und nomos = Gesetz. Es handelt sich also um eine gesetzmäßige Ordnung, die durch eine Taxonomie beschrieben wird.

Die Frage ob es sinnvoll ist Mathematikfähigkeiten anhand von Lern(ziel)taxonomien für den kognitiven Bereich zu ordnen, ist an sich nicht unumstritten. Blumberg, Alschuler und Rezmovic (1982) beantworteten diese Frage in Bezug auf die Entwicklung von Testaufgaben zur Kontrolle von Lernzielerreichungen negativ. Sie fanden heraus, dass bei Testaufgaben gleichen Inhalts, konstruiert auf den drei Stufen, (1) *recall or recognition*, (2) *simple interpretation* und (3) *application to problem solution* keine Leistungsunterschiede zwischen den Personen auftauchten und es somit keine Schwierigkeitsunterschiede zwischen den einzelnen Stufen gab (kein differentieller

Effekt). Die weiterführende Frage, ob kognitive Taxonomien in der Testentwicklung genutzt werden sollten, griffen mehr als 10 Jahre später Cizek, Webb und Kalohn (1995) auf. Sie entwickelten Items zur Erfassung von *comprehension*, *application* sowie *analysis* und unterzogen die Items einer Faktorenanalyse, die jedoch eher eine Einfaktorlösung nahe legte. Auch die Analyse der Korrelationen der drei Subtests mit dem Gesamtscore sprach gegen einen differentiellen Effekt der drei Stufen, da alle Korrelationen sehr hoch ausfielen und die Unterschiede zwischen Ihnen nicht signifikant waren (r = 0,87 bis r = 0,98). Cizek et al. (1995) ziehen daraus den Schluss, dass Taxonomiestufen nur berichtet werden sollten, wenn sie empirisch validiert wurden und ihre Anwendung als Leistungsindiz in Pass/Fail-Entscheidungen als generell kritisch zu sehen ist. Letztlich fordern sie mehr Forschung zu diesem Bereich, was sich der Autor dieser Arbeit zu Herzen nimmt.

Allein um einen Leistungstest zu erstellen, ist demnach eine taxonomische Ordnung nicht notwendig, doch stellen auch beispielsweise Anderson und Krathwohl (2001) heraus, dass der Hauptnutzen einer Taxonomie wahrscheinlich in der Erleichterung der Kommunikation über den Sachverhalt liegt. Dies entbindet nicht von der empirischen Prüfung der theoretischen Annahmen einer Taxonomie. Doch es setzt den Fokus auf den tatsächlichen Nutzen einer solchen Ordnung und das ist nicht eine Verbesserung von Auswahlentscheidungen, sondern eine Verbesserung der Kommunikation dessen, was erfasst wird.

Von den im folgenden dargestellten Taxonomien stellen jene von Bloom et al. (1956) und deren Revision durch Anderson und Krathwohl (2001) die bekanntesten Ordnungsschemata dar, weshalb sie sich auch am Beginn dieses Abschnitts befinden. In ihrem Appendix vergleichen Anderson und Krathwohl (2001) elf alternative Taxonomien mit der ursprünglichen Taxonomie (Bloom et al., 1956) und acht Taxonomien mit der zweidimensionalen überarbeiteten Fassung, wobei alle Alternativen versuchen, das ursprüngliche Werk von Bloom entweder zu verbessern oder leichter nutzbar zu machen. Aufgrund dieser Fülle kann hier nur eine Auswahl zwecks Beschreibung getroffen werden, die vor allem darauf basiert, ob eine Anwendung auf den Mathematikbereich möglich erscheint.

3.2.1 Bloom et al. (1956)

Die erste Idee zu dieser wohl berühmtesten Lernzieltaxonomie entsprang einer Sitzung der APA im Jahre 1948 in Boston, wo über die Notwendigkeit einer Lernzieltaxonomie diskutiert wurde, was nach einem Erstentwurf 1951 zu einer ersten Auflage im Jahre 1956

führte, die im Laufe der Jahre in mindestens 18 Sprachen übersetzt wurde (Bloom, 1994). Bereits auf den ersten Seiten geben die Autoren klar ihr Ziel zur erkennen. Dort heißt es: „In our original consideration of the project we conceived of it as a method of improving the exchange of ideas and materials among test workers, as well as other persons concerned with educational research and curriculum development." (Bloom et al., 1956, S. 10). Den Aufbau der Taxonomie mit ihren sechs Stufen skizziert die folgende Abbildung 5.

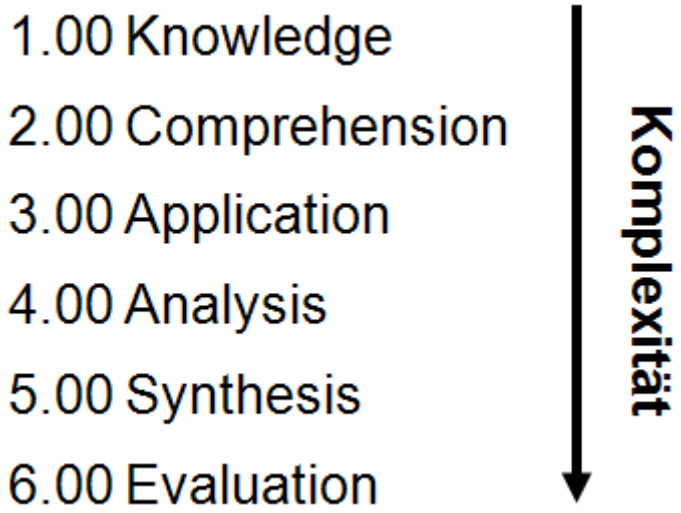


Abbildung 5 Lernzieltaxonomie nach Bloom et al. (1956).

Alle Stufen sollen sich demnach durch eine Ordnung steigender Komplexität auszeichnen. Bei Wissen (knowledge) geht es darum, dass ein Schüler zeigt, Ideen und Phänomene verstanden zu haben, die er im Lernprozess erworben hat und erinnert (und / oder wieder erkennt) (Bloom et al., 1956, S. 28). Auf der nächst höheren Stufe, Verständnis (comprehension) ist es entscheidend, dass der Schüler mit einer Kommunikation (verbal, schriftlich, symbolisch) konfrontiert ist und erkennt, was kommuniziert wird sowie in der Lage ist, die übermittelten Materialien und/oder Ideen zu benutzen. Hier wird eine weitere Unterteilung in Translation (übersetzen der Kommunikation in eine andere Form, z.B. Zahlen in Inhalte), Interpretation (die Bedeutung der einzelnen übermittelten Ideen/Konzepte wird erkannt, ebenso wie ihre Beziehungen untereinander) und Extrapolation (von den übermittelten Ideen/Konzepten ausgehend können Vorhersagen und Schätzungen getroffen werden) vorgenommen (Bloom et al., 1956, S. 89). Auf Stufe drei, Anwendung (application), geht es darum - gegeben ein neues zu lösendes Problem – die richtige Vorgehensweise darauf anzuwenden, ohne explizit darauf hingewiesen zu werden. Unter Analyse (analysis*)* verstehen Bloom et al. (1956, S. 144) ein „Breakdown of

the material into its constituent parts and detection of the relationships of the parts“ , wobei dies die Verbindung zwischen Elementen sowie ihre Relation untereinander einschließt. Die vorletzte Stufe ist am ehesten typisch für kreatives Verhalten, da hier die einzelnen Elemente zu einem neuen Ganzen zusammengesetzt werden müssen (Bloom et al., 1956, S. 162). Auf der letzten Stufe, Evaluation (evaluation*),* schließlich geht es um die Bewertung von Ideen, Arbeiten und Lösungen anhand interner (logische Schlüssigkeit) und externer (Zweckdienlichkeit zur Zielerreichung) Kriterien.

3.2.1.1 Empirische Bewährung der Taxonomie

Nach Kreitzer und Madaus (1994) unterteilen sich Studien zur empirischen Prüfung der Taxonomiestruktur in jene, die Items den Stufen der Taxonomie zuordnen lassen um die Interrater-Reliabilität zu bestimmen und solche, die versuchen, die kumulative Struktur der Taxonomie (eine höhere Stufe baut auf der nächst tieferen auf) statistisch zu prüfen. Für ersteren Typ von Untersuchung stellte Fairbrother (1975) fest, dass die Übereinstimmung von 22 Lehrern bei Zuordnung von Testaufgaben zu den ersten vier Stufen der Taxonomie unzureichend war. In zwei Datensätzen waren es einmal 14 und einmal 18 von 40 Aufgaben ($\alpha = 1\%$) für die Cohen‘s Kappa eine Übereinstimmung anzeigte, die über der zufällig zu erwartenden lag. Was die explizite Einordnung von Testaufgaben in das Bloomsche Schema angeht, kommt Lipscomb (1985) zu dem Ergebnis, dass eine Einschätzung von 18 Aufgaben durch Studenten anhand eines semantischen Differentials (mit den Endpunkten *simple* - *complex*) mit einer Einordnung anhand der Taxonomie sehr hoch korrelierte. Lipscomb (1985) wertet das Ergebnis in Bezug auf die Taxonomie kritisch, wobei die Tatsache, dass die beiden Einordnungsverfahren zu äquivalenten Ergebnissen führten, nicht unbedingt gegen die Taxonomie sprechen muss.

Für den zweiten Typ Untersuchung erstellte Seddon (1978) eine Überblicksarbeit, die zu dem Schluss kam, dass eine gewisse Evidenz für einen kumulativen Aufbau der ersten vier Stufen besteht (Seddon, 1978, S. 320). Eine präzise Testmethodik stellt jene von Hill und McGraw (1981) dar, die mit einem SEM-Ansatz einen Datensatz von Stoker und Kropp (1966) im Hinblick auf die kumulative Taxonomiestruktur überprüften. Sie erhielten erst einen akzeptablen Fit als die Wissenskategorie entfernt wurde. Einen gänzlich anderen Weg gingen Solman und Rosen (1986), indem sie Aufgaben ähnlichen Inhalts auf den sechs unterschiedlichen Taxonomiestufen erstellten und die Schüler einen kognitiven Orientierungstest (Figural Intersection-Test) durchführen ließen. Es ergab sich lediglich ein

deutlicher Unterschied zwischen Synthesis / Evaluation und den restlichen Stufen der Taxonomie, d.h. nur Schüler, die Aufgaben der letzten beiden Stufen richtig lösten, wiesen bessere Ergebnisse im kognitiven Test auf.

3.2.2 A revised taxonomy: Anderson und Krathwohl (2001)

Im November 1995, also fast 40 Jahre nach erscheinen der ersten Taxonomie, trafen sich kognitive Psychologen sowie Curriculums- und Testspezialisten in New York, um die Notwendigkeit einer Revision der Taxonomie zu besprechen (Anderson, 1999). Aufgrund des Umfangs der revidierten Fassung der Taxonomie (Anderson & Krathwohl, 2001) kann sie hier nicht in allen Details dargestellt werden und es wird versucht die für diese Arbeit wichtigsten Aspekte und Befunde herauszuarbeiten und darzustellen.

Ein Grund für die Überarbeitung bestand darin, dass die ursprüngliche Taxonomie zunehmend eher als historisches Dokument betrachtet wurde und weniger als handlungsrelevantes Schema, obwohl sie aus Sicht der Autoren ihrer Zeit weit voraus war (Anderson & Krathwohl, 2001, S. 16). Durch die Weiterentwicklung des Wissens über Lernprozesse, Schülerverhalten und allgemeine Fortschritte der Psychologie sollte die Taxonomie in vielerlei Hinsicht verbessert werden. Zum Beispiel sollte dafür gesorgt werden, dass die Taxonomie Ziele und Lernaufgaben näher zusammenbringt: Was führt zu welchem Ziel? (Krathwohl, 1994, S. 197).

Einer der wesentlichen Unterschiede zum Original (Bloom et al., 1956) besteht darin, dass nun zwei Dimensionen existieren, eine Wissensdimension und eine Dimension kognitiver Prozesse auf welche in den folgenden Abschnitten noch eingegangen wird. Eine weitere, wesentliche Neuerung besteht sicher auch in der expliziten Aufteilung von objectives, also dem was ein Schüler durch Unterricht erreichen soll, in global- (z.B. eine lernbereite Klasse), educational- (ein Schüler lernt Notenlesen) und instructional objectives (tägliche, abgeschlossene Lerneinheiten) (Anderson & Krathwohl, 2001). Darüber hinaus wird nun auch nicht mehr von einer strikt kumulativen Struktur ausgegangen (Anderson, 1999, S. 8), die sich - wie unter Abschnitt 3.2.1.1 berichtet - ohnehin als fraglich erwies. Als Anwendungsbereiche werden einmal *lernen* (es wird eingeordnet was ein Schüler lernen soll), *Instruktion* (wodurch soll er lernen) und *Anordnung* (Passung von Lerngegenstand und Lernmöglichkeiten) benannt (Anderson & Krathwohl, 2001, S. 16).

3.2.2.1 Zur kognitiven Dimension

Ein sofort auffälliger Unterschied zur ursprünglichen Taxonomie besteht darin, dass die Benennungen der einzelnen Kategorien nun aus Verben anstelle von Subjektiven bestehen und die Kategorien *evaluation* und *synthesis* vertauscht wurden (Anderson & Krathwohl, 2001). Die folgende Abbildung 6 fasst die Veränderungen von alter zu neuer Taxonomie in Bezug auf die kognitive Dimension zusammen.

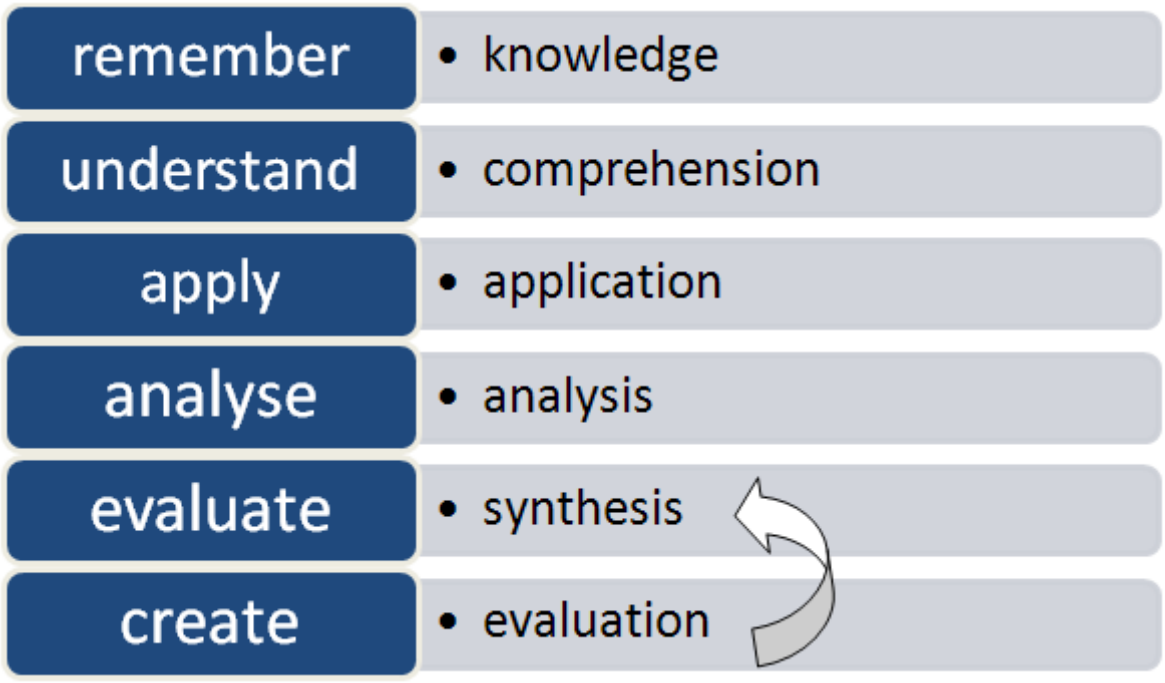


Abbildung 6 Veränderung von der alten (Bloom et al., 1956, rechts) zur neuen (Anderson & Krathwohl, 2001, links) Taxonomie.

Um einen ausreichend detaillierten aber nicht zu ausschweifenden Überblick bezüglich der kognitiven Dimension zu gewährleisten, wurden die kognitiven Prozesse mit einer kurzen Erläuterung und Beispielen in der folgenden Tabelle 3 zusammengefasst.

Tabelle 3 Auszug zur kognitiven Dimension nach Anderson und Krathwohl (2001) .

Kognitiver Prozess	Definition/Beispiel
Erinnern	Hier geht es um tatsächliches „Erinnern“, wie z.B. das Datum des Ausbruchs des 2. Weltkriegs UND/ODER Wiedererkennen. Wiedererkennen findet sich z.B. in vielen, aber nicht allen MC-Aufgaben.
Verstehen	Man muss die Aufgabe an sich verstehen, um sie zu lösen. Ein einfacher Abruf oder ein Wiedererkennen ist nicht mehr ausreichend. Zu verstehen gehören u.a. folgende Unterkategorien: - Übersetzen von einer Form in eine Andere. Z.b. von einer verbalen Beschreibung in eine mathematische - Beispiele für ein Konzept finden. Z.b. ein Beispiel für eine Primzahl finden. - Zusammenfassen
Anwenden	Hier geht es um das Ausführen einer vertrauten Aufgabe, wie eine Zahl durch eine andere zu dividieren, UND/ODER Das Ausführen einer unvertrauten Aufgabe, wie z.B. Prozentrechnen auf einen neuen Sachverhalt anwenden (Textaufgabe)
Analysieren	Hier muss das Material in seiner Einzelteile zerlegt werden. Z.b: Zuschreiben: Man liest ein politisches Programm und kann es der FDP, SPD oder CDU zuordnen.
Evaluieren	Bewertungen anhand von internen oder externen Standards vornehmen. - Z.B. Intern: Eine mathematische Herleitung ist in sich logisch.
Kreieren	Einzelteile werden zu einem Ganzen zusammengefügt. - Hypothesen aufstellen: Z.B.: Was sind mögliche Anwendungen des World-Wide-Web ? - Produzieren: Man erfindet ein neues Produkt, oder z.B. ein mathematisches Lösungsverfahren, oder man programmiert ein neues, kleines Statistik-Programm.

Anmerkung. Keine erschöpfende Darstellung.

Darüber hinaus erwähnenswert erscheint, dass zwar (wie unter 3.2.2) angesprochen die strikte kumulative Ordnung aufgegeben wurde, jedoch nach wie vor davon ausgegangen wird, dass sich die sechs Stufen prinzipiell nach Komplexität ordnen lassen (Krathwohl, 2002, S. 214).

3.2.2.2 Zur Wissensdimension

Die Wissensdimension (knowledge) teilt sich explizit in vier Unterkategorien auf und zwar Faktenwissen, konzeptuelles Wissen, prozedurales Wissen und metakognitives Wissen (Krathwohl, 2002). Diese vier Wissensarten lassen sich mit jeder der 6 kognitiven Stufen kombinieren, wodurch ein zweidimensionales Ordnungsschema entsteht. Die Autoren der überarbeiteten Taxonomie sehen Wissen als domänenspezifisch und kontextualisiert an (Anderson & Krathwohl, 2001). Ein Beispiel für Faktenwissen bestünde z.B. darin, dass Datum der letzten drei Kriege auf deutschem Boden zu benennen. Für konzeptuelles Wissen wäre ein Beispiel zu wissen, wie sich das deutsche Regierungssystem zusammensetzt (Parlament, Abgeordnete, Bundeskanzler etc.). Es ist also nötig die Struktur, den Modellcharakter, von etwas zu erfassen. Während für prozedurales Wissen entscheidend ist zu verstehen, wie man etwas macht (z.B. welches Vorgehen benötigt man zur Berechnung eines Klassendurchschnitts), ist bei metakognitivem Wissen eher strategisches Wissen und Wissen über sich selbst erforderlich (Anderson & Krathwohl, 2001). In diese Kategorie wäre sicherlich auch das Wissen über selbstregulative Lernstrategien (Schweizer, 2006) einzuordnen. Als Analogon zur Unterscheidung von Fakten- und prozeduralem Wissen kann die in der kognitiven Psychologie schon lange gebräuchliche Unterscheidung von prozeduralem und deklarativem Gedächtnis angesehen werden (Solso, MacLin & MaClin, 2005). Die Beschreibung von konzeptuellem Wissen hingegen erinnert deutlich an mentale Modelle, wie sie von Johnson-Laird (1980) beschrieben wurden. Solche Modelle bilden eine gedanklich manipulierbare, modellhafte Abbildung des bearbeiteten Realitätsausschnittes (Schnotz & Bannert, 2003).

3.2.2.3 Empirische Bewährung der Taxonomie

Obwohl die revidierte Fassung der Taxonomie bereits vor acht Jahren veröffentlicht wurde bleibt die verfügbare empirische Evidenz, welche über den wissenschaftlichen Appendix der Originalarbeit herausgeht, eher gering. Was die Dimension kognitiver Prozesse angeht (siehe Abschnitt 3.2.2.1) sind aufgrund der Ähnlichkeit zur Originalarbeit von 1956 (Bloom et al.) ähnliche Schwierigkeiten zu erwarten. Die in Abbildung 6 herausgestellte Vertauschung der letzten beiden Kategorien und der Übergang von Subjekten zu Verben ist bereits eine Reaktion auf empirische Befunde (Anderson & Krathwohl, 2001).

Die Frage ob sich die verschiedenen Facetten der Wissensdimension überhaupt trennen lassen ist bisher nicht direkt geprüft worden. Einer der wenigen Ansätze zur empirischen

Prüfung der revidierten Taxonomie stammt von Näsström und Henriksson (2008), die testeten ob eine Einordnung von schwedischen Bildungsstandards (die den gewünschten Endzustand anzeigen) im Fach Chemie und entsprechenden Assessment-Fragen anhand zweier Beurteiler reliabel gelingt. Sie kamen zu dem Schluss, dass die Interrater-Übereinstimmung zwischen den Beurteilern für die Bildungsstandards 53% betrug und für die Assessment-Fragen 60%, was deutlich besser war als die entsprechenden Werte einer anderen Taxonomie (37% versus 48%). Darüber hinaus waren in der revidierten Taxonomie nur wenige Teile der Bildungsstandards nicht klassifizierbar und es traten fast überhaupt keine doppelten Zuordnungen (d.h. zu mehreren Zellen gleichzeitig) auf, woraus die Autoren letztlich die Überlegenheit der Taxonomie ableiteten (Näsström & Henriksson, 2008).

3.2.3 Wilson (1970)

Wilson (1970) stellt ein Klassifikationssystem bestehend aus zwei Dimensionen dar, Inhalte (content) auf der einen Seite und Verhalten (behaviors) auf der anderen. Die Inhalte umfassen *number systems, algebra* sowie *geometry* und die Verhaltensweisen setzen sich zusammen aus *computation, comprehension, application, analysis, interest and attitudes* als auch *appreciation* (Wilson, 1970). Ein wesentlicher Unterschied zu den anderen bisher und im folgenden vorgestellten Taxonomien besteht darin, dass dieses Modell explizit für den Bereich Mathematik entwickelt wurde. Interessanterweise lehnt sich dieses Konzept deutlich an Blooms (Bloom et al., 1956) Taxonomie an, wenngleich die kognitiven Prozesse hier als Verhaltensweisen (behaviors) bezeichnet sind, was wiederum deutlich an die kognitive Dimension der revidierten Taxonomie (Anderson & Krathwohl, 2001) erinnert. Dort werden die Prozesse nicht mehr mit Subjekten, sondern Verben benannt. Zur Erfassung des Lernfortschritts unterscheidet Wilson (1970) schließlich zwischen knowledge (für computation und teils comprehension) und abilities (teils für comprehension und für application, analysis komplett). Man kann Wilsons Modell demnach als sehr fortschrittliche Mischform aus alter und (damals noch nicht existenter) neuer Taxonomie ansehen. Eine Zusammenfassung der wichtigsten Elemente findet sich in der folgenden Abbildung 7.

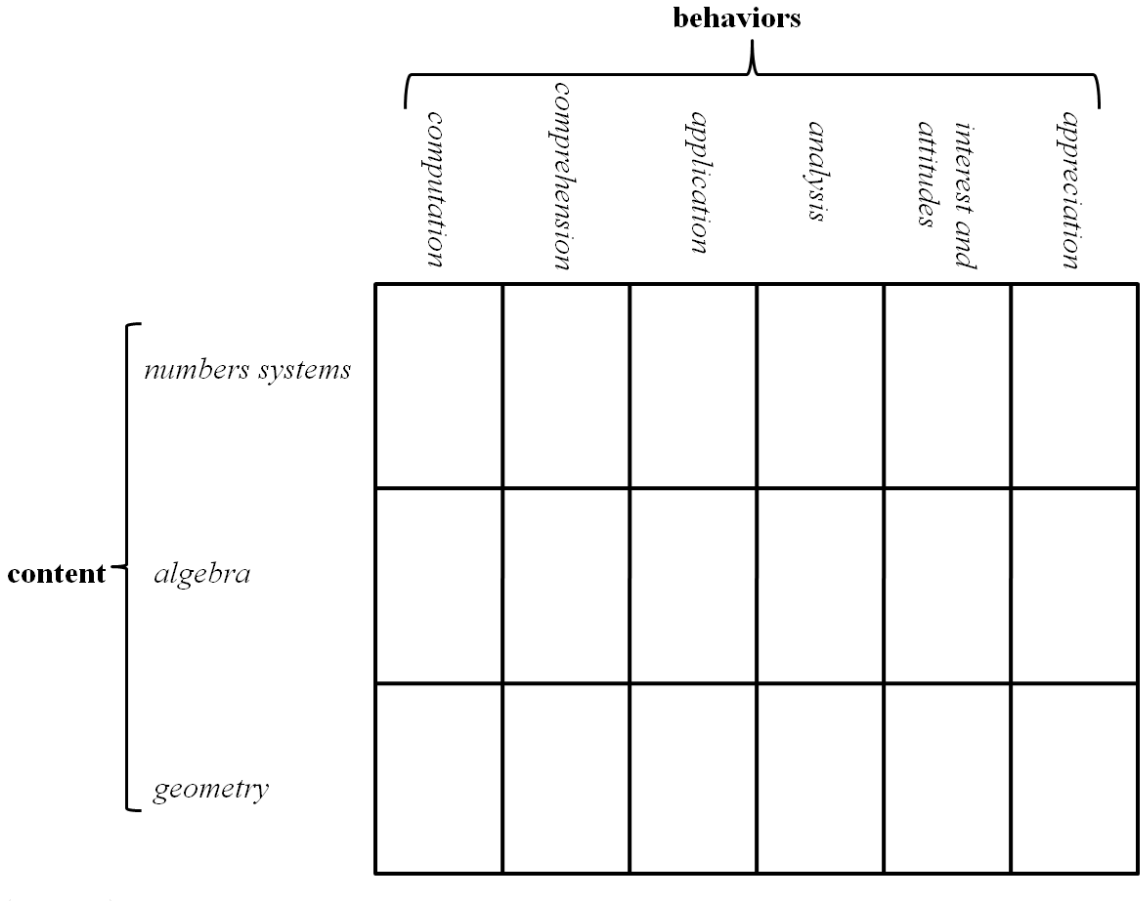


Abbildung 7 Zusammenfassung von Wilsons (1970) Modell

Zwei affektive Komponenten auf die hier nicht weiter eingegangen wird erinnern an eine bekannte Taxonomie für den affektiven Bereich (Krathwohl, Bloom & Masia, 1956) und finden sich in den Verhaltensweisen interest and attitudes sowie appreciation. Das besondere an Wilsons (1970) Herangehensweise ist, dass er explizit zwischen Inhalten und kognitiven Prozessen (er spricht von Verhaltensweisen, behaviors) unterscheidet.

3.2.4 Components Display Theory (CDT)

Ein zentraler Punkt von Merrills (1983) CDT ist die so genannte performance - content Matrix welche sich unterteilt in find, use sowie remember auf der einen Seite und fact, concept, procedure sowie principle auf der anderen Seite. Die Fakten, Konzepte und Prozeduren können als analog zu den gleichnamigen Facetten der Wissensdimension in der revidierten Taxonomie von Anderson und Krathwohl (2001) angesehen werden.

Der Fokus der CDT liegt letztlich klar auf der Beschreibung von Instruktionsstrategien (Choi, 1986), was auch dadurch deutlich wird, dass Merrill (1983) sehr intensiv auf primäre und sekundäre Präsentationsformen eingeht, die mit entscheidend für das

Erreichen eines Lernziels sind. Über die mögliche psychometrische Trennung einzelner Zellen seiner Taxonomie äußert er sich nicht. So heißt es an anderer Stelle von Merill (1999, S. 1): „CDT describes instructional strategy in terms of strategy components: primary presentation form (PPFs), secondary presentation forms (SFPs), and interdisplay relationships (IDRs).“, was den beschriebenen Fokus der Taxonomie unterstreicht. Dazu passt auch, dass Merrill (1983) Empfehlungen vorgibt, wie viel Prozent der Aufgaben jeder Zelle der performance – content-Tabelle korrekt gelöst werden sollten (z.B. bei remember/fact 0% Fehler), um von einem Lernerfolg auszugehen.

Hier in dieser Arbeit ist jedoch sehr wichtig, dass eine Taxonomie nicht nur den idealen Lehr-Lernprozess schildert, der nicht Gegenstand dieser Arbeit ist, sondern vielmehr auch eine detaillierte - und vor allem empirisch belastbare - Hilfe bei der Ordnung von erreichten Lernergebnissen liefert. Für ein mögliches Schema im Mathematik-Kontext ist festzuhalten, dass auch hier kognitive Prozesse auf der einen Seite (find, use, remember) und verschiedene Wissensarten auf der anderen Seite (concept, procedure, principle, fact) – analog zu Merrill (1983) - unterschieden werden können.

3.2.5 Ein integratives Modell

Wie bereits aus den vier vorgestellten Beispielen (Bloom, Anderson & Krathwohl, Wilson sowie Merrill) hervorgeht, scheinen die Taxonomien zur Klassifizierung von Lernzielen und dazugehörigen Assessment-Strategien im kognitiven Bereich deutliche Ähnlichkeiten aufzuweisen. Dementsprechend stellten Reigeluth und Moore 1999 ein Rahmenmodell auf, das als Synthese verschiedener anderer Modelle gelten soll und in der folgenden Tabelle 4 abgetragen wurde.

Tabelle 4 Auszug eines Lernzieltaxonomien-Vergleichs nach Reigeluth und Moore (1999, S. 54).

Bloom	Gagné	Ausubel	Merrill	Reigeluth
knowledge	verbal information	rote	remember verbatim	memorize information
comprehension	verbal information	meaningfull	remember paraphrased	understand relationships
application	intellectual skill		use	apply skills
analysis synthesis evaluation	cognitive strategy		find	apply generic skills

Demnach lassen sich Lernzieltaxonomien in die vier Stufen memorize, Information, understand relationships, apply skills und apply generic skills einordnen (Reigeluth & Moore, 1999). Dass solche Vergleiche durchaus problematisch sind, war den Autoren klar, so heißt es an einer Stelle: „In many ways, trying to compare the theories is like comparing apples and oranges“ (Reigeluth & Moore, 1999, S. 55). Deshalb werden auch die Schemata von Gagné und Ausubel nicht genauer dargestellt. Gagnés (1984) Taxonomie besteht eigentlich aus fünf Aspekten – sie enthält zusätzlich motor skills und attitudes – was sie als ganzes im Kontext dieser Arbeit schwer anwendbar macht. Bei Ausubels (1968) Ansatz gibt es tatsächlich, wie in Tabelle 4 dargelegt, nur eine Dichotomie zwischen diskreten, isolierten Wissenselementen (rote learning) und dem begrifflich etwas schwer zu fassendem meaningful learning, das an den Begriff der literacy aus den PISA und TIMSS-Studien (vgl. Abschnitt 2.2) erinnert. Im Vergleich zu Blooms (1956) Taxonomie scheint der wesentliche Unterschied von Reigeluths und Moores (1999) Ansatz darin zu bestehen, dass sich apply generic skills über die letzten drei Stufen nach Bloom erstreckt.

3.2.6 Schlussfolgerung

Es wurden mehrere Taxonomien gesichtet und zwei Ansätze stechen sicherlich hervor. Zum einem jener von Wilson (1970) wegen seiner expliziten Orientierung an Mathematikfähigkeiten und zum anderen der Ansatz von Bloom und seine Überarbeitung, letztere alleine schon wegen ihrer historischen Bedeutung. Dies beantwortet die Frage danach, welcher der bisher vorgestellten Ansätze den meisten Mehrwert erbringt, jedoch noch nicht erschöpfend. Einen für diese Arbeit entscheidenden Aspekt stellen Reigeluth und Carr-Chellman (2009, S. 65) heraus, indem sie schreiben: „While Bloom’s taxonomy is well known and thoroughly describes a number of naturally cohesive learning outcomes, we feel that Bloom’s Taxonomy was primarily designed to describe and assess learning outcomes rather that to select different sets of methods.“. Genau das, was die beiden Autoren gewissermaßen kritisieren, ist in dieser Arbeit – was die Taxonomien angeht – von Interesse: Einmal das Beschreiben von Lernzielen, d.h. was wird eigentlich von einem Berufsanfänger - am Ende der Sekundarstufe I im Bereich Mathematik - erwartet und wie Fragen zu diesem Assessment klassifiziert werden können. Demnach fällt hier die Wahl auf eine Orientierung an Blooms (revidiertem) Ansatz.

Auf Basis der (revidierten) Bloomschen Taxonomie (Anderson & Krathwohl, 2001; Bloom et al. 1956) sollen die Mathematikaufgaben geordnet werden. Die zu entwickelnden Aufgaben werden jedoch nicht auf Basis einer Taxonomie erstellt, da die Ergebnisse dazu

eher ernüchternd waren (vgl. Abschnitt 3.2). Die Vorhersagevalidität eines Tests lässt sich durch eine nachträgliche Einordnung in eine Lernzieltaxonomie natürlich nicht erhöhen, doch macht dieses Vorgehen den Test für den Anwender verständlicher, indem es die Kommunikation über das, was der fertige Test misst vereinfacht.

3.3 Erweiterte Integration: Ein kognitives Prozess x Inhalte –Modell

Nach Sichtung aktueller Intelligenzkonzepte, Lernzieltaxonomien und internationaler Vergleichsstudien scheint es wünschenswert, alle drei in ein mögliches Modell zu integrieren. Nachdem als Schlussfolgerung aus Intelligenzdiagnostik und internationalen Vergleichsstudien in Abschnitt 3.1.6 bereits vier Skalenkonzeptionen entwickelt wurden, stellt sich die Frage nach einer möglichen Erweiterung des Modells im Sinne einer kognitiv/taxonomischen Ordnung. Dafür wird auf den bereits geschilderten (Abschnitt 3.2.3) Ansatz von Wilson (1970) zurückgegriffen, der Inhalte und kognitive Prozesse (bei ihm behaviors genannt) kombinierte. Momentan liegen eine Konzeption für vier inhaltliche Skalen auf der einen Seite und die Entscheidung für Blooms (revidierte) Taxonomie auf der anderen Seite vor. Diese beiden Konzepte lassen sich analog zum Vorgehen bei Wilson (1970) – einer zweidimensionalen Ordnung in Inhalte und kognitive Prozesse (bei ihm behaviors) – vereinen. Bezieht man gedanklich zusätzlich noch das BIS-Modell (Abschnitt 3.1.3) mit seiner Rautenform und Ordnung in Inhalte und Operationen heran resultiert das vorgeschlagene Modell gemäß Abbildung 8.

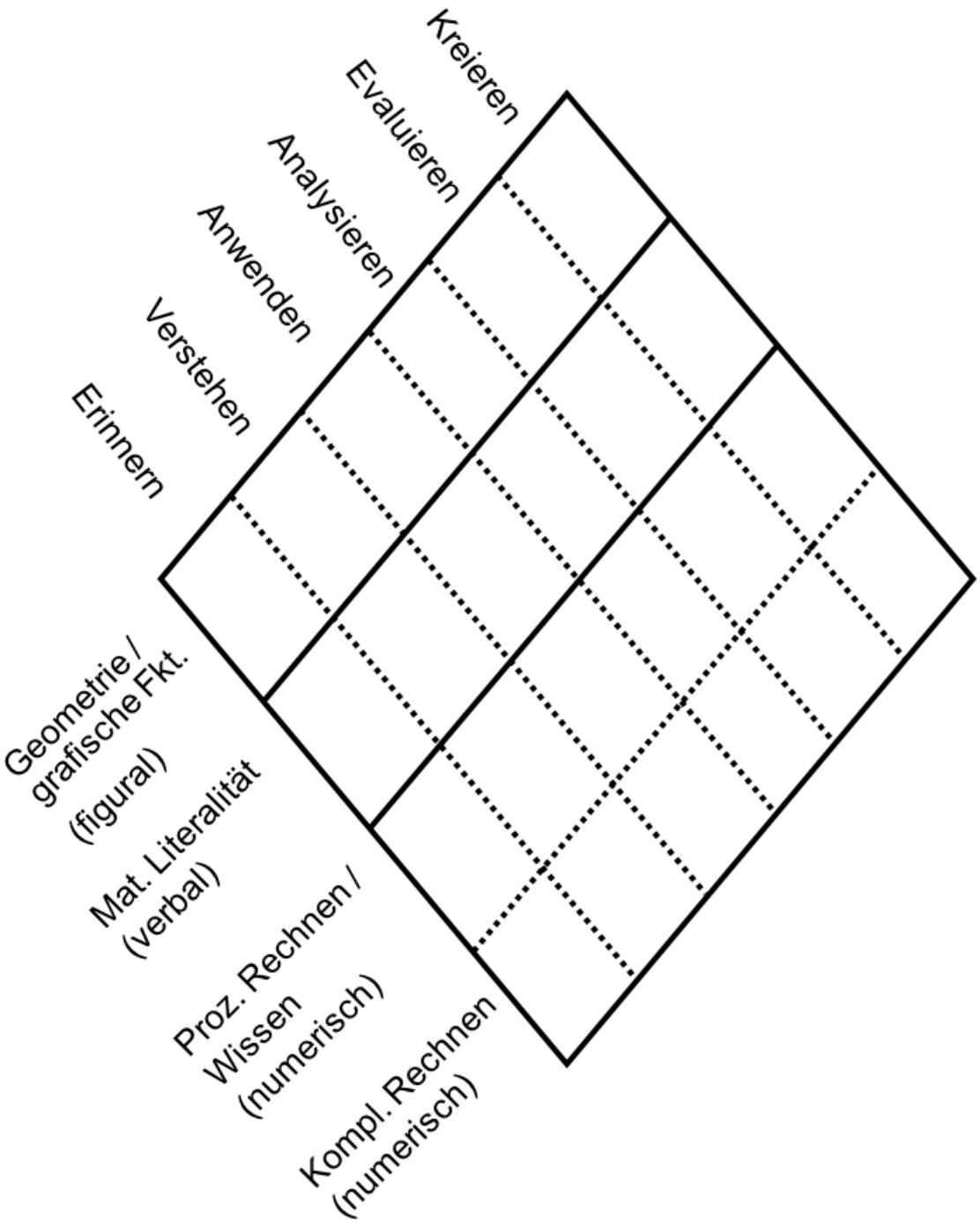


Abbildung 8 Modell zur Ordnung der Mathematik auf Basis einer Kognitive Prozesse x Inhalte-Matrix.

Demnach wird die Wissensdimension (vgl. Abschnitt 3.2.2.2) nach Anderson und Krathwohl (2001) entfernt und durch die bereits erarbeiteten Skalenkonzeptionen – angelehnt an die Intelligenzdiagnostik – ersetzt. Dieses Vorgehen kann durchaus kritisiert werden, doch sei vorab darauf hingewiesen, dass die Autoren der überarbeiteten bloomschen Taxonomie selbst schreiben:

“Like the original framework, our revision will be most beneficial to those who adapt it to their purposes.“ (Anderson & Krathwohl, 2001, S. 259).

Zwar wurden alle sechs kognitiven Prozesse übernommen, doch ist nicht gesagt, dass sie auch alle am Ende der Sekundarstufe I auftauchen müssen. Die gestrichelten Linien in Abbildung 8 signalisieren ferner, dass die empirische Unterscheidung einiger Bereiche schwierig ist. Für die Inhalte prozedurales Rechnen und komplexes Rechnen ist dies zu erwarten, da beide Facetten aus ähnlichen Inhalten (numerisch, Zahlen) bestehen (vgl. Abschnitt 3.1.6.). Diese beiden Facetten könnten insbesondere zur Trennung von sehr guten und sehr schlechten Personen nützlich sein. Die Rautenform des Modells gemäß Abbildung 8 signalisiert in Anlehnung an den BIS (Jäger, 1982; Abschnitt 3.1.3) eine Korrelation der Dimensionen. Dies ergibt sich jedoch nur dadurch, dass jede Testaufgabe gleichzeitig einer Skala und einer kognitiven Stufe zuzuordnen ist.

Während eine statistische Prüfung zur Trennbarkeit der Inhaltsfacetten durchaus Erfolg versprechend scheint, soll für die kognitiven Prozesse der Weg einer Experteneinschätzung durch Lehrer vorgenommen werden, die jedoch auch statistisch auf ihre Konsistenz geprüft werden kann. Eine ebenfalls sehr interessante Perspektive stellt die Möglichkeit dar, durch Experten zu erfassen, was von der Zielpopulation erwartet wird und dies damit in Bezug zu setzten, was der Test – ebenfalls aus Sicht von Experten – erfasst (Abschnitt 8.4.3).

II EMPIRISCHER TEIL

4 Vorprüfungen zu den bisherigen theoretischen Überlegungen

Im Abschnitt 3 wurde ein Modell der Mathematikfähigkeiten aufgestellt, das es zu testen gilt. Während für die kognitiven Prozesse (Taxonomiestufen) die Urteile von Experten eingeholt werden sollen (vgl. vorheriger Abschnitt) ist es für die Inhaltsdimension empirisch durchaus Erfolg versprechend, eine rein statistische Trennbarkeit der vier vorgeschlagenen Bereiche zu prüfen. Exploratorisch betrachtet entspricht dies der Frage: Finden sich die vier vorgeschlagenen Skalen, oder zumindest drei davon, in den Daten? Der Nachweis der kognitiven Prozesse (Taxonomiestufen) gestaltet sich vor allem deswegen als schwierig, weil Taxonomiestufen nicht sinnvoll gemittelt werden können (vgl. Abschnitt 8.4.3) und wird in späteren Abschnitten dieser Arbeit behandelt (vgl. Abschnitt 8.4).

Ein Problem stellt natürlich dar, dass bisher noch kein Mathematiktest auf Basis des vorliegenden Modells entwickelt wurde. Auch liegen für keinen der zu Beginn

vorgestellten Tests Daten vor. Was jedoch vorliegt, sind Daten aus einem Mathematiktest der in einem Experimentalpraktikum der Uni Mannheim über mehrere Semester hinweg entwickelt wurde (Jung, Kempf & Seggewiß, 2007; Orth, 2006). Auch dieser Test wurde nicht gemäß den 4 Skalenkonzeptionen des Abschnitts 3.1.6 entwickelt. Da jedoch sowohl Geometrieinhalte, als auch Textaufgaben und einige Rechenaufgaben in diesem Test enthalten sind, stellt er eine gewisse Näherung an kommerzielle Verfahren dar.

4.1 Hypothesen I

Unter der Annahme einer gewissen Allgemeingültigkeit der aufgestellten Skalenkonzeptionen müssten sich vier oder zumindest drei der Skalen gemäß Abschnitt 3.1.6.1 bis 3.1.6.4, wenigstens ansatzweise, in den Daten zu diesem Test wieder finden lassen. Daraus folgt, vorsichtig formuliert:
H1:Der Test ist mehrdimensional.

Eine Skalenkonzeption ist nur sinnvoll umsetzbar, wenn einigermaßen deutlich erkennbar ist, zu welcher Skala eine Aufgabe gehören soll. Daher sollten auch keine gravierenden Probleme bei dem Versuch auftauchen, die Items dieses Tests den eigenen Skalenkonzeptionen zuzuordnen. Daraus resultiert direkt Hypothese H2:
H2: Tendenziell lassen sich die Items des Tests den vorgeschlagenen Skalen zuordnen.

Unmittelbar aus der zweiten Hypothese ergibt sich eine weitere Hypothese:
H3: Die in das Schema eingeordneten Items lassen sich tendenziell statistisch trennen

Insbesondere für die Hypothesen H1 und H3 sind methodische Erörterungen notwendig und werden im folgenden Abschnitt vorgenommen.

4.2 Bestimmung der N-Dimensionalität eines Tests

Bevor die Dimensionalität eines Tests geprüft wird, sollte klargestellt werden wieso es überhaupt in der Psychologie von Bedeutung ist, dass Informationen über die Anzahl der Dimensionen eines Tests (bzw. seiner einzelnen Skalen) vorliegen. Ein sehr früher Kommentar hierzu stammt von McNemar (1946, S. 298), der beschreibt, dass insbesondere falls es intendiert ist, Personen in eine Rangreihe zu bringen, nur im Falle von Unidimensionalität gewährleistet ist, dass Teilnehmer mit gleichem Rang quantitativ und (in Grenzen) qualitativ ähnlich sind. Ohne Frage wäre es für eine differenzierte

Diagnostik von Fähigkeiten unerwünscht, wenn beispielsweise völlig mangelhaftes räumliches Vorstellungsvermögen durch überdurchschnittliche Leistung in Textaufgaben kompensiert werden könnte. Doch genau dies wäre der Fall, wenn z.B. ein Intelligenztest der solche Aufgaben ähnlicher Schwierigkeit enthält nur einen Gesamtscore bieten würde. Dies könnte auch ein Grund für die relativ geringen Geschlechterunterschiede internationaler Vergleichsstudien sein, auf die in Abschnitt 9.4 genauer eingegangen wird. Zwei völlig unterschiedliche Profilgestalten könnten genau demselben Level zugeordnet werden und es würde sich um ein kompensatorisches Modell handeln. Zweifelsohne hätte dies auch deutliche Auswirkungen auf die Validität eines solchen Tests. Würde man hiermit Studiumsanwärter für einen technischen Studiengang (z.B. Elektrotechnik) selektieren, hätten Personen mit sehr guten räumlichen Vorstellungsvermögen und eher schlechten verbalen Fähigkeiten dieselben Chancen auf einen Studienplatz wie Personen mit exakt umgekehrtem Fähigkeitsprofil.

Derzeit existiert eine Fülle von - teilweise theoretisch fraglichen - Indizes und Verfahren zur Feststellung der Dimensionalität eines Tests. Hattie (1985, S. 141) unterscheidet in einem Überblicksartikel zwischen fünf, im folgenden kurz dargestellten, unterschiedlichen Ansätzen zur Beurteilung der Unidimensionalität eines Tests oder einer Skala. Es werden nicht alle Ansätze ausführlich behandelt, sondern stets typische Vertreter kurz dargestellt und ihre Bedeutung für die vorliegende Arbeit herausgearbeitet. Im letzten Abschnitt werden aktuellste Verfahren besprochen, deren Entwicklung erst nach Hatties (1984, 1985) Arbeiten vonstatten ging. Zunächst gilt es jedoch eine Definition von Unidimensionalität (und damit auch Multidimensionalität) aufzustellen.

4.2.1 Begriffklärung: Unidimensionalität

Bevor die Hypothese 1 geprüft werden kann muss geklärt werden, was in dieser Arbeit überhaupt unter Unidimensionalität verstanden wird. Eine Definition von Hattie (1984) bezieht sich letztlich auf die Annahme eines einzigen latenten Traits. So soll für die Items eines eindimensionalen Tests gelten, dass die Wahrscheinlichkeit das Item richtig zu beantworteten nur von der Ausprägung einer Person auf dem latenten Trait θ und dem Ausmaß indem das Item diesen Trait zwecks Lösung benötigt abhängig ist und dies muss natürlich für alle Items des Tests gelten. Diese Antwortwahrscheinlichkeit wird in Item-Response-Modellen (Embretson & Reise, 2000; Kubinger, 1988) durch eine Normal-Ogive mit den Parametern θ und der Aufgabenschwierigkeit (häufig ξ benannt) dargestellt.

Hattie (1985) verwendet statt dieser Funktion den Platzhalter f,was zum Ausdruck bringen soll das es sich nicht um die logistische Funktion handeln muss, sondern f auch eine Stufenfunktion (siehe Abschnitt 4.2.2), oder eine lineare Funktion (siehe Abschnitt 4.2.3) darstellen kann.

Analog zu McDonald (1981) wird damit von einem abgeschwächten Prinzip der stochastischen Unabhängigkeit ausgegangen. Es reicht an dieser Stelle aus, wenn nach obiger Annahme keine Korrelationen mehr zwischen den Items bestehen. McDonald (1981) nennt dies abgeschwächte lokale stochastische Unabhängigkeit, da nach wie vor Zusammenhänge zwischen den Items bestehen könnten die nicht durch lineare Korrelationen erfasst werden sondern nonlinearer (quadratisch, kubisch usw.) Natur sind.

4.2.2 Antwortpattern

Einer der bekanntesten Vertreter dieser Gruppe ist Guttmans Reproduzierbarkeitskoeffizient bei dem davon ausgegangen wird, dass gegeben eine Personen löst eine Aufgabe korrekt, sie einen höheren Skalenwert aufweist als alle anderen Personen, die diese Aufgabe nicht richtig lösten (Guttman, 1944, S. 143). Auch Guttman war bereits klar, dass dieses idealtypische Muster bei realen Tests kaum zu erreichen war, weshalb er den Reproduzierbarkeitskoeffizienten, CR = 1 - (Inkonsistente Antworten/Alle Antworten), vorschlug (Guttman, 1950, S. 77). Doch selbst wenn eine ausreichende Reproduzierbarkeit erreicht wird bleiben formallogische Probleme des Verfahrens bestehen. Guttman (1944, S. 143) selbst wählte als Beispiel drei Mathematikaufgaben, die wie folgt lauten (übersetzt durch den Autor):

Item 1: Wenn r der Radius eines Kreises ist, wie lautet seine Fläche?

Item 2: Welche Werte für x sind für folgende Gleichung gültig: $ax^2 + bx + c = 0$

Item 3: Was ist $\frac{de^x}{dx}$

Abbildung 9 zeigt das hierbei zu erwartende Idealmuster, lediglich VP 5 fällt aus dem Rahmen, da sie Item 1 nicht gelöst hat.

	Item 1	Item 2	Item 3
VP 1	0	0	0
VP 2	1	0	0
VP 3	1	1	0
VP 4	1	1	1
VP 5	0	1	1

Abbildung 9 Guttman-Pattern mit einer Abweichung (VP5).

Guttman (1944, S. 149) erwähnt, dass der Grund ein solches Muster hier zu erwarten hauptsächlich kultureller Natur wäre, da die Lösung von Item 1 in einer niedrigeren Klassenstufe (in Deutschland vermutlich die 8. Klasse) als Item 2 und wiederum Item 3 gelernt würde. Er empfiehlt, den Versuch eine Skala zu bilden aufzugeben, falls zu viele Personen, wie hier VP 5, vom erwarteten Muster abweichen (Guttman, 1944, S. 139).

Führt man diesen Gedanken weiter, stellt sich die Frage, was eine Skala eigentlich inhaltlich darstellen soll, die lediglich zwischen verschiedenen Entwicklungsstufen differenziert, die für sich genommen völlig unterschiedliche Inhalte aufweisen könnten. Eine solche Skala könnte leicht in verschiedene Teilbereiche, wie z.B. Textaufgaben und Rechenaufgaben aufgeteilt werden, für die durchaus die Frage gestellt werden darf, ob sie dieselbe Fähigkeit erfassen. Das grundsätzliche Problem, dass Items mehr als einen Inhaltsbereich erfassen können und dennoch eine perfekte Guttman-Skala bilden, ist bereits lange bekannt und wurde von Campbell und Kerckhoff (1957, S. 298) beschrieben. Sie verwenden hierfür ein Item von Guttman, mit dem Wortlaut: „Wenn du einen Sohn hättest, würdest du wollen, dass er ein gewisses Maß an Armeetraining zu Friedenszeiten, nach dem Krieg, erhält oder nicht?“ (übersetzt durch den Autor). Es ist auf den ersten Blick ersichtlich, dass dieses Item wohl nicht nur die Einstellung gegenüber Wehrdienst zu Friedenszeiten (wie von Guttman angedacht), sondern eine Vielzahl von anderen Aspekten erfasst (Vater-Sohn Beziehung, Einstellung zur Armee etc.). Auch andere Autoren, wie z.B. Stookey und Baer (1976) konnten zeigen, dass Guttman-konforme Skalen häufig mehr als eine Dimension erfassen. Dies passt zu der Feststellung von Amelang & Zielinski (2001, S. 139), dass eine Gutmann-Skalierung bisher nur in sehr wenigen Fällen vorgenommen wurde und lediglich für reine, begründet eindimensionale, Niveau-Tests Erfolg versprechend sei, was auch für die verbesserte aber konzeptuell sehr ähnliche Formel von Loevinger gilt. Ein weiteres Problem stellt die Tatsache dar, dass der Gesamtscore eines Tests bestehend aus einer gleich gewichteten Summe von

Einzelfähigkeiten, z.B. verbale, figurale und numerische Intelligenzaufgaben, im Sinne Guttmans (1944) perfekt skalierbar wäre (Hattie, 1985, S. 143). Dazu passt die Feststellung von Guttman (1950, S. 85), bezogen auf die Basis der Skalogramm-Analyse, in der es heißt (übersetzt durch den Autor): "Die Skalenanalyse als solche enthält kein Urteil bezüglich des Inhalts; sie nimmt an, dass das Inhaltsuniversum bereits definiert ist".

Die einzige Möglichkeit diesen Konflikt zu umgehen, besteht darin von Experten beurteilen zu lassen, welche Items zu einem Item-Universum gehören (Guttman, 1950, S. 84), was jedoch aufgrund begründeter Skepsis gegenüber derart subjektivem Vorgehen (Campbell & Kerckhoff, 1957, S. 298) keine wirkliche Lösung zu sein scheint. Zur Feststellung der N-Dimensionalität eines Tests erscheint die Vorgehensweise nach Guttman daher ungeeignet.

4.2.3 Reliabilität

Reliabilität ist definiert als die Genauigkeit mit der ein Test das interessierende Merkmal erfasst und wird als Paralleltest-, Restestreliabilität oder interne Konsistenz in Form von Reliabilitätskoeffizienten erfasst (Engel-Schermelleh & Werner, 2008; Horst, 1971, S. 14). Wichtig zu betonen ist, dass die Reliabilität eines Tests keineswegs zwingend von dessen Homogenität abhängig ist. Nun ist es so, dass ein Maß der internen Konsistenz (z.B. Cronbach's α) bei völlig reliablen Aufgaben einen sehr niedrigen Wert erreichen kann, da die Aufgaben gleichzeitig äußerst heterogen sind. Als Gedankenexperiment seien hier vier Aufgaben gegeben, die Sprachkenntnis in Englisch, Französisch, Schwedisch oder Italienisch erfassen sollen. Alle diese Aufgaben könnten hoch reliabel sein und gleichzeitig völlig heterogen.

Cronbachs α wurde von seinem Erfinder als Maß zur Schätzung der Paralleltestreliabilität (Äquivalenz) entwickelt (1951, S. 297) und nicht um die Homogenität eines Tests zu erfassen. Eine der Originalformeln nach Cronbach (1951, S. 323), $\bar{r} = \frac{\alpha}{n+(1-n)\alpha}$, lässt sich leicht wie folgt umformen:

$$\bar{r}(n+(1-n)\alpha) = \alpha \quad (1)$$

$$n\bar{r} + \bar{r}\alpha - \bar{r}n\alpha = \alpha \quad (2)$$

$$n\bar{r} = \alpha(\bar{r}n - r + 1) \quad (3)$$

$$\frac{n\bar{r}}{\bar{r}n - r + 1} = \alpha \tag{4}$$

$$\alpha = \frac{\bar{r}n}{1 + (n-1)\bar{r}} \tag{5}$$

wodurch sich in (5) die in vielen Lehrbüchern abgedruckte Formel ergibt (Bei Cronbach (1951) ist sie nicht in Form von (5) zu finden). Bedingung für die Anwendung dieser Formel ist die Annahme gleicher Itemvarianzen und Kovarianzen der einzelnen Items, was jedoch praktisch nie der Fall sein dürfte und zu einer Überschätzung führt (Lienert & Raatz, 1994, S. 185). Weiterhin ist ersichtlich, dass mit zunehmendem *n* (Anzahl der Items), also formal $\lim_{n\to\infty} \frac{\bar{r}n}{1 + (n-1)\bar{r}} = 1$, d.h. Cronbach's α gegen 1 (für $\bar{r} > 0$) geht. Beide Einschränkungen wurden unter anderem von Green, Lissitz und Mulaik (1977, S. 833) behandelt, die herausstellen, dass α ein in vielen Fällen ungeeignetes Maß zur Prüfung der Homogenität darstellt. So impliziert Homogenität interne Konsistenz jedoch interne Konsistenz nicht Homogenität (Green et al., 1977, S. 831). Sinnvoll ist Cronbach's Alpha in erster Linie, wenn bereits von einer Homogenität des Tests ausgegangen wird (z.B. aus inhaltlichen Gründen) und die Reliabilität durch die interne Konsistenz, geschätzt werden soll. Sowohl Green et al. (1977) als auch Hattie (1984) konnten in Simulationsstudien zeigen, dass Cronbach's α aus den genannten Gründen schlecht zur Prüfung der Dimensionalität eines Tests geeignet ist. So ergibt sich für einen Test bestehend aus 36 Items bei dem jedes Item stets nur auf einem von 4 orthogonalen Faktoren deutlich lädt (Kommunalität: $h^2 = 0{,}90$) ein α von $\alpha = 0{,}90$. Cortina (1993) fasst die Problematik gut zusammen, indem er empfiehlt, α als nur konfirmatorisches Werkzeug zu verwenden, falls es bereits begründet erscheint, eine Skala zu bilden. In diesem Stadium scheint demnach α, ebenso wie seine in Simulationsstudien (Hattie, 1984, S. 71) schlecht abschneidenden Abänderungen, zur Prüfung der N-Dimensionalität eines Tests kaum geeignet zu sein.

4.2.4 Faktorenanalyse

Zwei grundlegende Verfahren aus diesem Bereich stellen die Hauptachsen und Hauptkomponentenanalyse dar, wobei sich erstere vor allem dadurch auszeichnet, dass die Kommunalitäten der Ausgangsvariablen geschätzt werden müssen und somit für jede Variable unique Anteile vorgesehen sind (McDonald, 1999; Überla, 1977). Dadurch wird

eine perfekte Reproduktion der ursprünglichen Korrelationsmatrix mit weniger Faktoren als Ausgangsvariablen möglich, im Gegensatz zur Hauptkomponentenanalyse. Die Frage, welche der Methoden wann angebracht ist, wird bereits seit einiger Zeit geführt, obwohl die Unterschiede in den Ergebnissen eher gering ausfallen und vernachlässigt werden können (Thompson & Brown, 2001, Velicer & Jackson, 1990, S. 21). Speziell in einem Kontext, bei dem für beide Verfahren die selbe Anzahl an Faktoren extrahiert wird, zeigen sich sehr ähnliche Lösungen (Velicer & Jackson, 1990, S. 5). Zur Bestimmung der Anzahl von Faktoren fasst Hattie (1985, S. 146) auf der einen Seite Verfahren auf Basis der Faktoreigenwerte (λ_p = quadrierte Summe der Ladungen auf einem Faktor p) zusammen, die jedoch alle nicht in der Lage sind das grundlegende Problem, nämlich wie hoch der Eigenwert sein sollte um eine ein-, zwei, n-Dimensionale Lösung zu wählen, lösen können. Auf der anderen Seite weist auch das alternative Heranziehen der prozentual aufgeklärten Varianz der n Ausgangsvariablen $\left(\lambda_p / n\right)$, prinzipiell ähnlich dem Scree-Plot, ein sehr subjektives Element auf. Die Differenz aufeinander folgender Eigenwerte also z.B. $Diff_1 = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_3}$ wurde als ein Kriterium für Homogenität vorgeschlagen, was jedoch nach Hattie (1985, S. 146) einen logischen Fehler darstellt, schließlich würde eine Variante mit lediglich ähnlich hohen Eigenwerten für Faktor 2 und 3 häufig einen hohen Index ergeben (da der Nenner des Bruchs dann sehr klein wird). Eine weitere Variante, die Verwendung von Indizes basierend auf den Kommunalitäten, weist das praktische Problem auf, dass diese wiederum geschätzt werden müssten. Ein zusätzliches Problem entsteht bei Anwendung der Faktorenanalyse auf dichotome Variablen (Stewart, 1981, S. 60) für das Lösungsansätze vorgeschlagen wurden, die das grundsätzliche Problem - der Nonlinearität durch Dichotomie - jedoch nur mindern (Lienert & Raatz, 1994, S. 113). So schlägt Green (1983) eine Faktorenanalyse binärer Items zur Prüfung der Unidimensionalität nur bei bereits sorgfältig entwickelten Tests vor. Der Gedanke das Problem durch Normierung der Korrelation auf ihr von den Itemschwierigkeiten abhängiges Maximum zu lösen, hat sich als praktisch untauglich erwiesen, da der Wert eines solchen Index stark von der Besetzung der Antwortkategorien eines Items abhängig ist (Davenport & El-Sanhurry, 1991). Da auch in Simulationsstudien (Hattie, 1984) die bisher erwähnten Indizes, ebenso wie Indizes die auf Residualmatrizen der Faktorenanalyse zurückgreifen (gewissermaßen ähnlich wie die Goodness-Of Fit Tests im SEM-Bereich), enttäuschende Ergebnisse zur Bestimmung der Dimensionalität aufwiesen, ist die lineare Faktorenanalyse nur mit Einschränkungen für

diesen Zweck zu verwenden. Praktisch bedeutet dies, dass den Ergebnissen von Collins, Norman, McCormick und Zatkin (1986) folgend mit Phi-Koeffizienten durchaus sinnvolle Ergebnisse bei binären Datensätzen erreicht werden können, dies jedoch sicherlich nicht den Königsweg darstellt. Insbesondere die Bestimmung der Anzahl von Faktoren bereitet größere Probleme, ebenso das Risiko der Entstehung von Schwierigkeitsfaktoren (auf die in Abschnitt 4.3 näher eingegangen wird); es ist wichtig diese Einschränkungen bei allen Analysen im Auge zu behalten.

4.2.5 Latent Trait Modell-Indizes

Die entscheidenden der Item Response Theorie (IRT) zugrunde liegenden Annahmen betreffen Monotonie, lokale stochastische Unabhängigkeit und Unidimensionalität (Moosbrugger, 2008; Nandakumar & Ackerman, 2004), wobei letztere in Abschnitt 4.2.1 bereits erläutert wurde. Bei Indizes basierend auf IRT-Modellen muss zunächst zwischen solchen für das Ein-, Zwei und Dreiparametermodellen unterschieden werden, wobei die meisten Indizes vom Einparameter-Modell ausgehen (Hattie, 1985, S. 151). Für alle drei Modelle ist eine wesentliche Vorraussetzung die Unidimensionalität des betreffenden Traits weshalb es zunächst plausibel erscheint Fit-Indizes der Modelle, wie z.B. Yen's-Q zur Prüfung dieser Annahme heranzuziehen (Hambelton, Swaminathan & Rogers, 1991). Hierbei handelt es sich um einen typischen, χ^2-verteilten, Item-Fit Index der sich ergibt als $\chi_B^{\,2} = \sum_{j=1}^{G} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}$, wobei G für die Anzahl der Fähigkeitsintervalle, O_{ij} für den Anteil korrekter Antworten in Fähigkeitsintervall j (für Item i), E_{ij} für den IRT-basierten, erwarteten Anteil korrekter Antworten und N_j für die Anzahl von Personen in Intervall j stehen (Dodeen, 2004, S. 264). Diese Familie χ^2 basierter Kennwerte wurde eigens von Wollenberg (1982, S. 83) entwickelt, um die Insensitivität traditioneller Testverfahren (vgl. Kubinger, 1988) wie dem Martin-Löf-Test, Anderson's Likelihood-Test oder dem Fischer-Scheiblechner-Test gegenüber Verletzungen der Unidimensionalität zu beheben. Leider weisen solche Indizes keinen eindeutigen Zusammenhang mit dem Merkmal der Unidimensionalität auf und auch alternative Kennwerte, die auf der Residualmatrix nach dem Fitten eines IRT-Modells beruhen, zeigten ausschließlich bei fast orthogonalen Dimensionen ausreichende Diskriminationsfähigkeiten zwischen ein- und mehrdimensionalen Modellen (Hattie, 1985, S. 155).

Eine weitere Methode, deren Logik jedoch - bedenkt man die Erkenntnisse aus Abschnitt 4.2.4 - unklar bleibt, ist die Anwendung einer Hauptkomponentenanalyse auf die Residualmatrix, erhalten durch Fitten eines Raschmodells. Diese Methode wird von Bond und Fox (2007, S. 255) vorgeschlagen, jedoch ist fraglich wieso, wenn ein Rasch-Modell gut fittet, relevante Korrelationen für eine Faktorenanalyse verbleiben. Wendet man das Verfahren jedoch an falls das Rasch-Modell nicht fittet, ist die Hypothese der Eindimensionalität in den meisten Fällen ohnehin schon verworfen.

Ein verglichen mit den bisher vorgestellten Methoden sehr neues Verfahren basiert auf einer Verallgemeinerung des Rasch-Modells auf mehrere (korrelierte) Dimensionen, das zudem noch polytome Antwortformate ermöglicht. Die Rede ist von dem MRCML-Modell nach Adams et al. (1997) welches bereits im Rahmen der Diskussion internationaler Vergleichsstudien erwähnt wurde. Mit dem Programm Conquest (Adams et al., 1997) ist es möglich beispielsweise ein eindimensionales gegen ein zweidimensionales Modell zu testen und die Verbesserung im Fit auf Signifikanz zu prüfen. Nachteil dieses Vorgehens ist, dass bisher keine Simulationsstudien vorliegen und dass selbst mit modernen Computern für mehr als zwei Dimensionen schnell extrem hohe Rechenzeiten entstehen (vgl. Abschnitt 9.6). An dieser Stelle erscheint es legitim zu fragen, wie es um den praktischen Nutzen der bisher angesprochenen Ansätze bestimmt ist. Embretson und Reise (2000) fragen diesbezüglich, ob viele der von Hattie (1984) beschriebenen Verfahren nun gänzlich unnütz seien, schließlich könnte die bisherige Analyse durchaus diesen Eindruck erwecken. Sie beantworten die Frage dahingehend, dass die exploratorischen Verfahren zwar isoliert betrachtet fragliche Ergebnisse liefern, jedoch durchaus helfen können bereits bestehende Annahmen zur Teststruktur zu bestätigen (im Sinne zusätzlicher Evidenz).

Letztlich existierten zwei noch nicht behandelte Verfahrensklassen, die als relativ modern angesehen werden und den eher klassischen Ansätzen vorgezogen werden sollten. Eines der Verfahren wurde bereits bei Hattie (1984, 1985) als viel versprechend gelobt, die nonlineare Faktorenanalyse. Das andere basiert auf einem Algorithmus von Stout (1987), und wird in Form der DIMTEST und DETECT Methode umgesetzt. Beide Verfahren werden in den beiden folgenden Abschnitten beschrieben und ihr potentieller Nutzen für diese Arbeit diskutiert.

4.2.6 Nonlineare Faktorenanalyse

Bereits in den 60er Jahren des letzten Jahrhunderts begann Roderick P. McDonald (1967) einen allgemeinen Ansatz zur nonlinearen Faktorenanalyse zu entwickeln. Schon einige

Jahre zuvor waren die Probleme um Schwierigkeitsfaktoren bei Faktorenanalysen dichotomer Items bekannt. Schwierigkeitsfaktoren entstehen vor allem bei dichotomen Itemformaten (so genannte Nonlinearität durch Dichotomie) weshalb sich andeutete, dass ein Bedarf für ein Verfahren zu Modellierung nichtlinearer Zusammenhänge zwischen latenten und mainfesten Variablen bestand (McDonald, 1962, S. 398). Auch in jüngster Zeit stellt die unangemessene Faktorenanalyse bei dichotomen Variablen unter Psychologen ein Problem dar (Kubinger, 2003). Eine seitens Kubinger (2003) vorgeschlagene Vorgehensweise, die Verwendung von tetrachorischen Korrelationen, ist auch zu hinterfragen. Die zugrunde liegende Annahme, dass die Variablen durch Dichotomisierung einer ursprünglich normalverteilten Variable entstanden sind, ist als problematisch anzusehen, da sie in der Praxis nur sehr selten zutrifft. Aus einer hitzigen Diskussion zwischen Karl Pearson und George Yule (ein Schüler Pearsons) darüber ob, von Yule polemisch formuliert, jemand der tot sei mehr oder weniger tot sein könne, leitete sich Yules frühe Kritik an Verfahren ab, die (wie bei tetrachorischen Korrelationen) eine Normalverteilung annehmen auch wenn es unrealistisch ist (Pearson & Herron, 1913, S. 161). Letztlich muss bei tetrachorischen Korrelationen davon ausgegangen werden, dass eine Überschätzung des Zusammenhanges umso stärker sein wird, umso eher die Annahme der zugrunde liegenden Normalverteilung verletzt wurde (Cohen, Cohen, West & Aiken, 2003; McDonald, 1999, S. 246). McDonald (1999, S. 270) sieht die IRT als eine Weiterentwicklung der Faktorenanalyse, speziell für dichotome Variablen. Sein Modell lässt sich skizzieren als $P\{U_j = 1 \mid \theta\} = N(a_j(\theta - b_j))$, wobei U_j die Antwort einer Person darstellt, θ einen Trait, a_j den Diskriminationsparameter und b_j den Itemparameter (McDonald, 1997). Die multivariate Verallgemeinerung lautet $P\{U_j = 1 \mid \underline{\theta}\} = N(\beta_{j0} + \underline{\beta}'_j \underline{\theta})$, wobei $\underline{\theta}$ für einen Traitvektor steht und $\underline{\beta}_j$ durch weitere Transformation einen Vektor von Faktorladungen darstellt (Gierl & Wang, 2005, S. 7). Eine inhaltlich sinnvolle Deutung von β_{j0} ist nur über Umwege möglich. So ist hier der Wendepunkt der ICC nicht bei β_{j0} (Analog zum Item-Parameter) sondern bei $\beta_{j0} + \underline{\beta}'_j \underline{\theta}$ =0 (McDonald, 1997). Das Programm bietet einen konfirmatorischen Modus, der es z.B. ermöglicht im Falle eines Mathetests festzulegen, dass die Items einer Skala Geometrie und grafische Funktionen auf nur auf einem Faktor laden und Textaufgaben auf einem anderen (McDonald, 2003).

Des weiteren kann bestimmt werden, ob die latenten Traits (Faktoren) korrelieren dürfen oder nicht. Als Indikatoren für den Fit des Modells sind zwei Ansätze zu beachten. Zum einen gibt NOHARM den Tanaka Index of Fit aus, $\tau = 1 - \left(\frac{Tr(Rs^2)}{Tr(S^2)} \right)$, wobei Rs^2 die Residual-Kovarianzmatrix darstellt und S die Stichprobenkovarianzmatrix (McDonald, 1997, S. 266). Da es sich um die Spur (trace) der Matrizen handelt, wird der Index umso größer, je geringer die Residualvarianz im Vergleich zur Ausgangsvarianz ausfällt (in der Matrizendiagonale befinden sich die Varianzen). Dieser Index ist auch unter dem Namen *GFI* bekannt (Ayala, 2008, S. 299; McDonald, 1999, S. 83).

Daneben soll nach McDonald (1997) unbedingt auch immer die (Wurzel der) Höhe der mittleren quadrierten Residuen betrachtet werden (der RMSR). Aus seinen Überlegungen geht hervor, dass jenes Modell zu bevorzugen ist, das einen hohen Tanaka-Index aufweist und gleichzeitig in einem möglichst niedrigen RMSR resultiert.

Bezüglich des Tanaka-Index existieren keine rationalen Entscheidungsregeln für die Einschätzung des Modell-Fit, abgesehen davon das größere Werte besseren Fit indizieren (Gierl & Wang, 2005, S. 12). Als Daumenregel kann nach McDonald (1999, S. 84) angenommen werden, dass ein Fit größer 0,90 als akzeptabel und größer 0,95 als gut angesehen werden kann (vgl. auch Abschnitt 8.3.2.1 zu Fit-Indizes).

Was den RMSR angeht, gibt es die Empfehlung einen $RMSR \leq 4 \cdot 1/\sqrt{N}$, wobei N die Stichprobengröße bezeichnet, als gut anzusehen (Ayala, 2008, S. 299; Fraser & McDonald, 1988). Bei Modellen mit sehr ähnlichen Kennwerten ist das einfachere Modell, im Sinne der Sparsamkeit (Occam‘s Razor), zu bevorzugen. Es existieren bereits viele Studien, die die Eignung von NOHARM die N-Dimensionalität eines Tests zu erfassen geprüft haben (z.B.: Champlain & Gessaroli, 1996; Hattie, 1985; Nandakumar, 1994), so dass es sinnvoll erscheint das Programm auch zu diesem Zweck (hier für den Inhaltsbereich Mathematik) einzusetzen.

4.2.7 Die DIMTEST-Prozedur

Bei der DIMTEST-Prozedur handelt es sich um ein in den 80ger Jahren von Stout (1987) entwickeltes, non-parametrisches Verfahren. Getestet wird stets die Hypothese, der Test sei eindimensional. Hierfür werden die Items in zwei Subtests aufgeteilt (Stout, 1987). Einen *assessment test* (AT), von dem angenommen wird, dass er Items enthält, die alle denselben Trait erfassen und einen *partitioning test* (PT) bei dem dies unklar ist.

Nun werden die Personen des PT basierend auf ihren Antwortwerten im AT in *k* Gruppen eingeteilt, wobei die theoretische Varianzschätzung der beiden Tests im Falle von Unidimensionalität einander entsprechen sollte. Interessant ist an dieser Stelle, dass es einen exploratorischen Modus zur Zusammenstellung der AT und PT-Subtests (basierend auf einer Faktorenanalyse mit tetrachorischen Korrelationen) und einen konfirmatorischen Modus gibt, bei dem die Zuordnung zu AT und PT vom Untersucher vorgenommen wird (Nandakumar & Ackerman, 2004, S. 97). Zum Verständnis der Logik von PT und AT sei nun auf Abbildung 10 verwiesen.

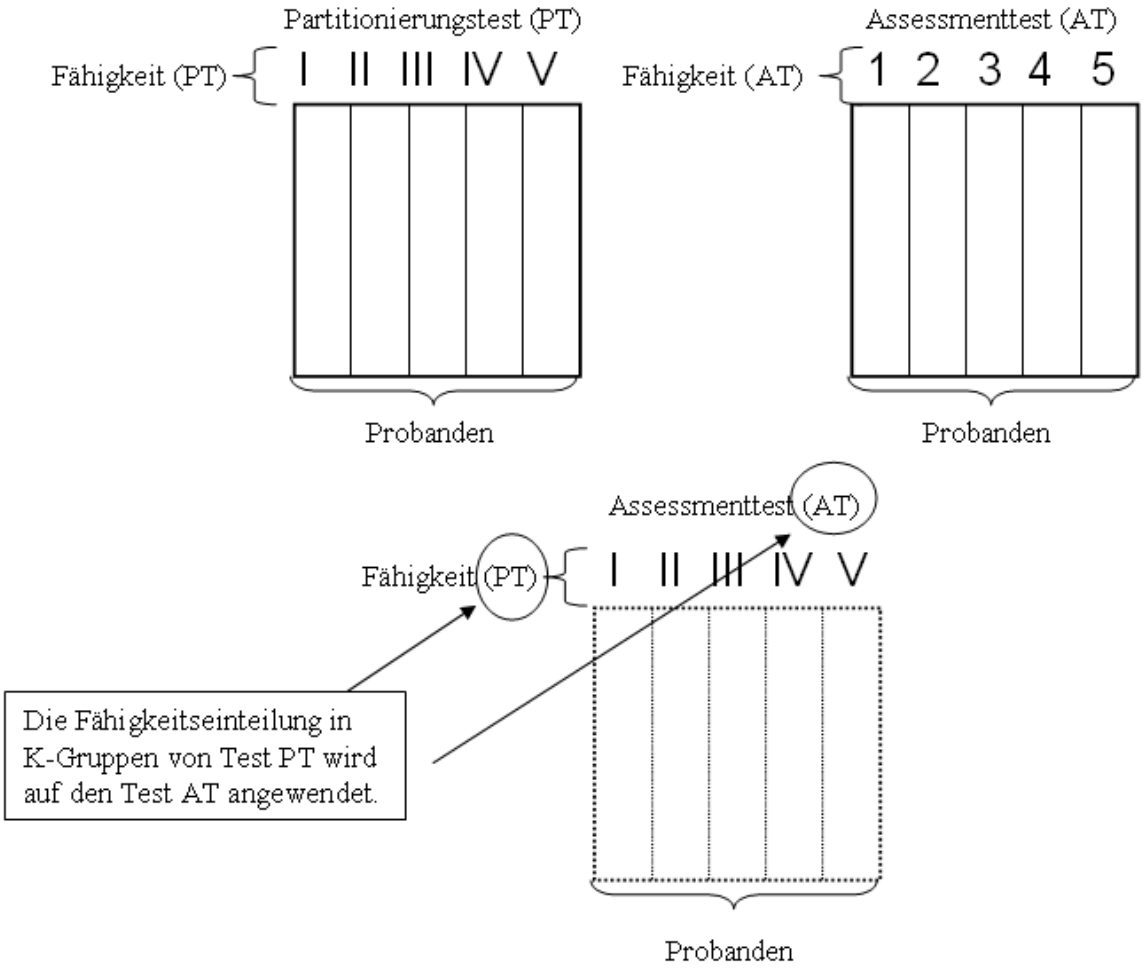


Abbildung 10 Logik der Aufteilung in AT und PT.

Eine etwas formalere Zusammenfassung von Stouts Verfahren findet sich bei Hattie, Krakowski, Roger und Swaminathan (1996, S. 2) in Form von

$$\frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} |Cov(U_i, U_j \mid \theta)| \simeq 0 \qquad (8)$$

, wobei U_i und U_j einfach Items eines Tests der Länge *N* darstellen. Im Endeffekt ist dieses Prinzip jenem der abgeschwächten lokalen stochastischen Unabhängigkeit nach McDonald

(siehe Abschnitt 4.2.6) sehr ähnlich. Die von Stout (1987, S. 594) für dieses Verfahren erarbeitete und im Jahre 1999 (Zhang & Stout) weiterentwickelte Teststatistik lautet:

$$T = \frac{1}{\sqrt{K}} \sum_{k=1}^{k} \left(\frac{\sigma^2{}_k - \sigma^2{}_{u,k}}{S_K} \right) \tag{9}$$

Hierbei stellt $\sigma^2{}_k$ eine Varianzschätzung für den *k*-ten Subtest dar, die sensitiv auf Verletzungen der Unidimensionalität reagiert, $\sigma^2{}_{u,k}$ hingegen eine Varianzschätzung die den selben Wert ergibt, egal ob der Test unidimensional ist, oder nicht. (siehe auch Abbildung 10). Für die Varianz $\sigma^2{}_k$ spielt also die vorgenommene Aufteilung in *k* Subtests eine bedeutende Rolle, im Gegensatz zur Varianz $\sigma^2{}_{u,k}$ welche, unabhängig von der Homogenität von AT zu PT, denselben Wert annimmt (Hattie, 1996, S. 3). Für die Herleitung der Standardisierung (S_k) und einer Bias-Korrektur muss aus Platzgründen - und weil hier nur die Logik des Verfahrens von Interesse ist - auf die Arbeit von Stout (1987) und Nandakumar und Stout (1993) verwiesen werden.

DIMTEST ist lediglich in der Lage zu prüfen, ob die Annahme der Unidimensionalität den tatsächlichen Daten gerecht wird, nicht jedoch wie viele Dimensionen dem Test zugrunde liegen. Zu diesem Zweck wurde von Zhang und Stout (1999) die DETECT-Methode entwickelt, die im Anschluss an die DIMTEST Ergebnisse durchgeführt werden kann. DETECT schätzt das Ausmaß an multidimensionaler Einfachstruktur, das sich in einem Datensatz findet (Tate, 2003, S. 171). Die Logik des DETECT Verfahrens besteht darin, dass die minimale Anzahl Dimensionen gesucht wird, die gleichzeitig die Bedingung $Cov(X_{i1}, X_{i2} \mid \Theta_{TT} = \theta) = 0$ bestmöglich erfüllt (Zhang & Stout, 1999, S. 217). Θ_{TT} ist ein (gewichteter) Test-Composite, ihm könnten in einem Mathematiktest z.B. $\Theta_{Algebra}$ und $\Theta_{Geometrie}$ zugrunde liegen. θ stellt eine spezielle Ausprägung dar, die auf eine, oder mehrere Personen zutreffen kann. Es handelt sich also praktisch um eine Realisation des Test-Composite und in diesem Beispiel würde er auf $\theta_{Geometrie}$ und $\theta_{Algebra}$ zurückgehen. X_{i1} und X_{i2} wären die Antworten von Personen auf zwei Items. Von Bedeutung ist ferner, dass die beiden Werte X_{i1} und X_{i2} nicht in den Test-Composite Θ_{TT} eingehen (Gierl & Wang, 2005, S. 4; Roussos & Ozbek, 2006, S. 219). Die Itemkovarianz für jedes Itempaar sollte minimal sein, wenn die Fähigkeit auf dem Test-Composite konstant gehalten wird. Schließlich sollten es ausschließlich die Fähigkeiten sein, die zu Kovarianzen zwischen

zwei Items führen. Die Bedingung entspricht im Endeffekt der Forderung nach paarweiser (da immer zwei Items betrachtet werden) lokaler stochastischer Unabhängigkeit, hier für den multidimensionalen Fall (Zhang & Stout, 1999). Abbildung 11 zeigt ein Beispiel für den Test-Composite bei einem Test dem zwei angenommene Fähigkeiten, Geometrie und Algebra, zugrunde liegen.

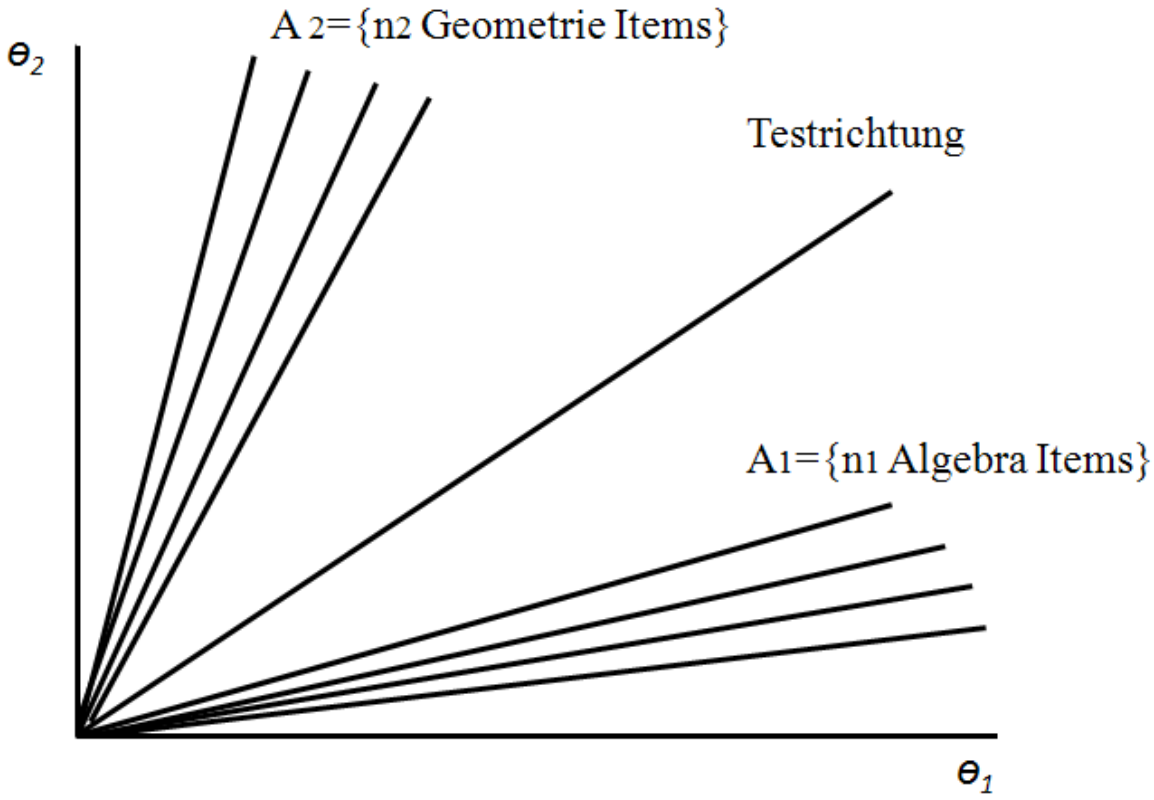


Abbildung 11 Veranschaulichung der Logik hinter DETECT, nach Zhang & Stout (1999, S. 218)

Eine Partitionierung P des Tests zu finden, für die die paarweisen, bedingten Itemkovarianzen (bzw. ihr Erwartungswert) den Wert 0 annehmen ist unrealistisch. Deutlich interessanter ist ein Maß, das Auskunft über den Grad der Multidimensionalität gibt. Zu diesem Zweck wurde die so genannte DETECT D-Statistik entwickelt. Gemäß dargelegter Logik sollten zwei Items aus demselben (homogenen) Inhaltsbereich eine positive hohe Kovarianz aufweisen, Items aus unterschiedlichen Bereichen hingegen eine niedrigere (negative) oder keine Kovarianz (Zhang & Stout, 1999, S. 219). Daraus ergibt sich die DETECT Statistik (Gierl & Wang, 2005, S. 6)

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \le i1 \le i2 \le N} \delta_{ij} E[Cov(X_i, X_i \mid \Theta_{TT} = \theta)] \qquad (10)$$

, wobei δ_{ij} so definiert ist, dass es für zwei Items aus der selben Dimension (z.B. Algebra) den Wert 1 und für zwei Items aus unterschiedlichen Dimensionen (z.B. Algebra und Geometrie) den Wert -1 annimmt. Der Wert D(P) soll einen möglichst hohen Wert annehmen. Die Logik sei abschließend an einem stark vereinfachten (eigentlich handelt es um geschätzte Erwartungswerte und es gibt mehrere Schätzvarianten) Gedankenexperiment erläutert.

Betrachtet man alle Items eines Test der z.B. Mathematikfähigkeit erfassen soll, so werden sich im Falle von Unidimensionalität teils positive, teils negative Kovarianzen ergeben und D(P) einen Wert nahe 0 annehmen. Im Falle von Multidimensionalität hingegen müssen sich jedoch nach Zhang und Stout (1999) bedingte Kovarianzen ergeben, die positiv für Items innerhalb einer Partition und negativ für Items zwischen Partitionen sind. Dadurch muss sich ein D(P) Index mit einem Wert größer Null ergeben. Je höher also der Wert des DETECT-Indexes D(P) ausfällt, desto mehr Multidimensionalität findet sich in den Daten. Das DETECT-Programm wird jene Anzahl und Aufteilung des Tests in Itemcluster finden, die den größtmöglichen DETECT-Wert darstellt (Zhang & Stout, 1999). Wichtig zu erwähnen ist noch, dass die Logik beinhaltet, dass dies nicht zwingend viele Dimensionen sein müssen. Versucht man z.B. die Big-Five mit 10 statt 5 Dimensionen zu beschreiben, werden häufig bedingte Kovarianzen zwischen Items verschiedener Dimensionen auftreten, denn die 10 Traits würden die wahre Struktur nicht richtig beschreiben.

Zusätzlich zum soeben beschriebenen Index wird häufig auch noch ein r_{max} Wert berichtet, der darüber informiert zu welchem Ausmaß die gefundene Anzahl (und Aufteilung) von Dimensionen einer Einfachstruktur entspricht (Gierl & Wang, 2005, S. 6). D.h. DETECT liefert zwei Werte: Einen D(P) Index, der das Ausmaß an Multidimensionalität in den Daten widerspiegelt und einen $r_{\max}$ Index der die Annäherung der gefunden Struktur an eine Einfachstruktur beschreibt.

Die einzelnen Herleitungen zum (genauen) DETECT-Vorgehen wurden in einigen statistischen Arbeiten (Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987), einschließlich Monte-Carlo Studien zur Überprüfung der Annahmen (Meara, Robin & Sireci, 2000; Seraphine, 2000), dargelegt und sind nicht Zentrum dieser Darstellung. Entscheidend ist, dass sich DETECT und DIMTEST z.B. bereits bei Tate (2003) oder Gierl und Wang (2005) als wirkungsvolle Verfahren zu Bestimmung der N-Dimensionalität von psychologischen Tests erwiesen haben, was ihre Anwendung rechtfertig.

4.2.8 Clusteranalyse

Die Logik des HCA/CCPROX-Verfahrens, PROX im Namen steht für Proximitäten und HCA für hierachical-cluster-analysis, ist leicht erklärt (Marden, Roussos & Stout 1998): So wird stets jenes Paar von Items (bzw. Clustern) vereinigt, das - gegeben die Scores der Personen auf allen anderen Items - die geringste Kovarianz aufweist. Abswoude, Ark und Sijtsma (2004, S. 9) notieren hierfür formal $E\{Cov[X_j, X_k \mid R_{(-j,-k)}]\}$ wobei X_j und X_k für die Scores der Personen auf den Variablen j und k stehen und $R_{(-j,-k)}$ für die Scores auf den restlichen Variablen. Jenes Item- oder Clusterpaar mit dem geringsten Wert für diesen Ausdruck wird in jedem Durchgang vereinigt. Eine grundsätzliche Frage bei allen Clusteranalytischen Verfahren ist die Bestimmung der Entfernung verschiedener Cluster. Marden et al. (1998, S. 21) kamen in einer vergleichenden Analyse zu dem Schluss, dass das UPGMA-Maß in Kombination mit dem HCA/CCPROX-Ansatz die besten Ergebnisse liefert. UPGMA steht für unweighted pair-group method of average, hierbei wird die Vereinigung mehrerer (Variablen oder) Cluster zu einem neuen Cluster basierend auf dem ungewichteten Mittel der Proximitäten (Kovarianzen) aller Einzelpaare bestimmt (Marden et al., 1998). Bei der HCA handelt es sich um ein agglomeratives Verfahren, d.h. es wird mit so vielen Clustern wie vorhandenen Variablen gestartet und in jedem Schritt zwei Variablen vereinigt. Nach jeder Vereinigung stellen die zwei Variablen nun eine neue Variable dar. Hier stellt sich die Frage, an welcher Stelle mit dem HCA-Verfahren abgebrochen werden sollte. Da von Marden et al. (1998) hierzu keine klaren Vorgaben existieren und das Programm keinerlei Fit-Indizes zu liefern vermag, ist es nur möglich den Versuch zu unternehmen, inhaltlich / theoretisch den agglomerativen Schritt mit der am besten interpretierbaren Lösung zu wählen. Dies ist auch nach Abswoude et al. (2004) die zu favorisierende Vorgehensweise, jedoch zugleich als extreme Einschränkung diese Methode zu betrachten, was ihre Verwendung in diesem Stadium der Arbeit jedoch nicht zu sehr einschränkt.

4.2.9 Schlussfolgerungen für diese Arbeit

Eine Frage die sich aus den bisher angerissenen mehr oder weniger verbreiteten Ansätzen ergibt, ist jene nach einem praktischen Vorgehen zur Dimensionalitätsbestimmung. Es wurde bereits erklärt, dass Guttman's (1950) Idealstruktur unwahrscheinlich ist, Cronbach's α (Cronbach, 1951) stark von der Anzahl der Items abhängig ist, die lineare

Faktorenanalyse bei binären Items mit Vorsicht zu genießen ist und auch moderne non-parametrische Verfahren (DIMTEST/DETECT, NOHARM) mit Bedacht angewendet werden sollten. Des weiteren sollte nicht vergessen werden, dass die N-Dimensionalität eines Tests nicht nur von den Items, sondern auch von der Personenstichprobe abhängen kann (Hattie, 1985, S. 159). Es ist durchaus denkbar, dass bei sehr niedriger Fähigkeit keine Ausdifferenzierung in verschiedene Bereiche (Geometrie, Algebra etc.) vorliegt, bei generell hoher Fähigkeit jedoch schon (oder umgekehrt, vgl. folgender Abschnitt 4.3).
Bei bereits vorliegenden Tests erübrigen sich klassische Itemanalysen, da diese bereits dem Manual zu entnehmen sind. Sind diese noch nicht vorhanden, oder unvollständig, sollte ein Vorgehen gemäß Abbildung 12 angewendet werden.

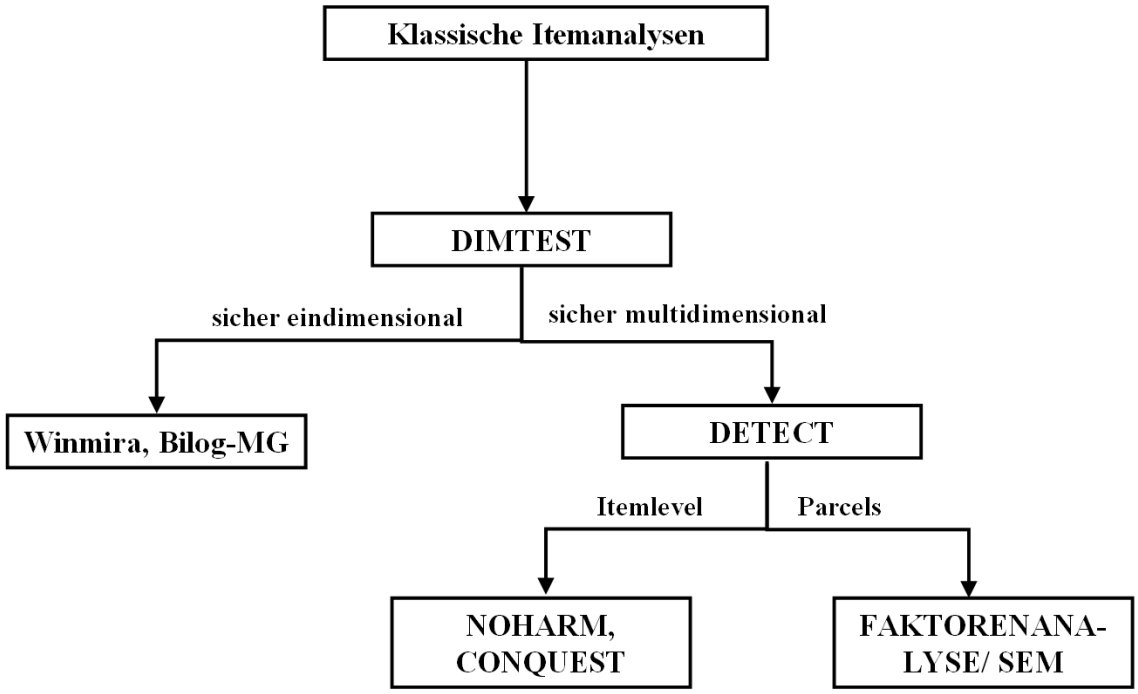


Abbildung 12 Ablaufschema zur Prüfung der N-Dimensionalität eines Tests.

Komplexen Verfahren wie DIMTEST keine klassischen Analysen vorzuschalten wäre sicherlich abzuraten, weshalb das Schema um diesen, eigentlich selbstverständlichen, Aspekt erweitert wurde. Klassische Kennwerte, wie akzeptable Trennschärfen und eine ausgewogene Schwierigkeitsverteilung sollten die Vorraussetzung für eine Anwendung komplexer Verfahren sein, um deren Ergebnisse überhaupt sinnvoll interpretieren zu können. Schließlich komprimiert DIMTEST sämtliche Information auf einen einzigen Kennwert, (T-Statistik, vgl. Abschnitt 4.2.7), DETECT auf die Anzahl der Dimensionen und zwei Kennwerte (r_{max} und D(P)).

Gemäß Abbildung 12 soll mit DETECT geprüft werden, wie viele Dimensionen eine optimale Aufteilung ergeben. Zwar liefert das Programm auch eine Aufteilung der Items auf die vorgeschlagenen Dimensionen, jedoch keinerlei Kennwerte auf Itembasis hierzu, auch ist nicht gesagt, dass eine DETECT Lösung sinnvoll interpretierbar ist; schließlich ist das Verfahren ausschließlich datengetrieben. Hauptzweck von DETECT in dem Schema gemäß Abbildung 12 ist es einen Anhaltspunkt zu liefern, wie viele Dimensionen überhaupt sinnvoll zu den Daten passen, um im Anschluss detaillierte Thesen zur Struktur zu prüfen. Auf Itemebene geschieht dies anhand von NOHARM (McDonald, 1999) und CONQUEST wobei die Anwendung von CONQUEST nur bei großen Stichproben und präzisen Hypothesen sinnvoll ist (Adams et al. 1997). Da NOHARM auch einen explorativen Modus bietet, wird dieser darüber hinaus genutzt, um die Schlussfolgerungen aus den DETECT Ergebnissen abzusichern, indem explorative Modelle mit ein, zwei, drei und 4 Faktoren in Bezug auf ihren Fit verglichen werden.

Ein Werkzeug auf Ebene von Parcels stellen Strukturgleichungsmodelle und Faktorenanalysen dar. Dieser letzte Aspekt, die Bündelung von Aufgaben zu Miniskalen, ist in einen größeren Kontext einzubetten und wird deshalb im folgenden Abschnitt dargestellt.

Das Programm HCA/CCPROX wurde nicht in das Schema gemäß Abbildung 12 aufgenommen, weil es keinerlei Fit-Indizes – weder deskriptive noch inferenzstatistische – bietet (Marden et al., 1998). Möglicherweise kann dieses Programm jedoch helfen Strukturen in Daten zu entdecken, was in Abschnitt 4.4.2.2 überprüft wird.

4.3 Die Bedeutung der Itemschwierigkeit für Strukturanalysen

In Abschnitt 4.2 wurde im Zusammenhang mit Schwierigkeitsfaktoren (Faktorenanalyse) bereits darauf hingewiesen, dass insbesondere bei dichotomen Items Probleme entstehen können. In den vorherigen Abschnitten lag der Fokus des Interesses vor allem auf der Bestimmung der Anzahl latenter Dimensionen eines Tests. Von den dort im Detail vorgestellten Verfahren scheint einzig NOHARM (McDonald, 1999) dazu geeignet, konfirmatorische Prüfungen der Struktur eines Mathematiktests durchzuführen. Leider handelt es sich bei NOHARM, im Gegensatz zur Faktorenanalyse und dem Ansatz der Strukturgleichungsmodelle (Loehlin, 2004), um ein wenig etabliertes Verfahren. Das schränkt seine Anwendbarkeit nicht ein, sehr wohl jedoch seine Kommunizierbarkeit (z.B. in einem Testmanual). Darüber hinaus bieten vor allem Strukturgleichungsmodelle, als

Kombination von Pfadmodellen und konfirmatorischer Faktorenanalyse, weitereichende Möglichkeiten präzise und umfangreiche Theorien zu testen (Schumacker & Lomax, 2004, S. 6). Deshalb muss für Strukturanalysen, das heißt wenn nicht mehr nur die Anzahl der Dimensionen entscheidend ist, sondern auch das spezifische Ladungsmuster und weitere Variablen (z.B. andere Tests) die Frage der Anwendbarkeit erneut bewertet werden. Das Problem der Schwierigkeitsfaktoren bei binären Items ist bereits lange bekannt und es wurden verschiedene Möglichkeiten damit umzugehen empfohlen (Gebbert, 1977; McDonald & Ahlawat, 1974; Witte & Caspar, 1977). Letztlich bleibt das Parceling (Gorsuch, 1983) durchaus eine sinnvolle Option falls klassische Faktorenanalyse und Strukturgleichungsmodelle eingesetzt werden sollen. Deshalb widmet sich der folgende Abschnitt dieser Technik, die durchaus auch Kritik hervorgerufen hat (siehe z.B. Bandalos & Finney, 2001).

Dass das Problem von Schwierigkeitsfaktoren keinesfalls realitätsfern sondern vielmehr ernst zu nehmen ist, wird in Abschnitt 8.2.1 an einer Endform des zu entwickelnden Tests demonstriert.

4.3.1 Parceling

Eine eng mit Strukturgleichungsmodellen und verschiedenen Formen der Faktorenanalyse verbundene Technik stellt das so genannte Parceling dar (Jäger, 1982; Kishton & Widaman, 1994). Diese Technik erlangte zuerst durch Anwendung im Rahmen der Konstruktion von Cattells (1956) 16PF-Fragebogen eine gewisse Bekanntheit und wurde von ihm selbst als (neben dem Instrument) wohl wichtigster Beitrag des Papers beschrieben (S. 208). Cattell (1956) führte ein so genanntes radial-parceling durch, bei dem die Items basierend auf einem so genannten Kongruenzmaß zu Parcels zusammengesetzt werden. Generell existieren verschiedenste Varianten des Parcelings und die Forschungsergebnisse zu den Folgen dieser unterschiedlichen Vorgehensweisen sind keineswegs eindeutig (Little, 2002; Nasser & Wisenbaker, 2006; Rogers & Schmitt, 2004). Da sich in einer Untersuchung von Bandalos und Finney (2001) herausstellte, dass bei Fähigkeitstests die am häufigsten eingesetzte Variante das schwierigkeitsbasierte Parceling von Items innerhalb möglichst homogener Subskalen darstellt, steht diese Variante hier im Fokus. Ziel ist vor allem der Bildung von Schwierigkeitsfaktoren vorzubeugen, welche eher die Schwierigkeit der Items als deren inhaltliche Eigenschaften widerspiegeln (Gorsuch, 1983). Ein Beispiel für die Anwendung dieser Parceling-Technik stellt die Entwicklung des Intelligenzstrukturmodells nach Jäger (1982) dar, welches dem Berliner

Intelligenz-Strukturtest (Jäger et al., 1997) zugrunde liegt (siehe auch Abschnitt 3.1.3). Der theoretische Hintergrund dieser Parceling-Variante besteht darin, dass die Kovarianz zwischen zwei Variablen gemäß folgender Gesetzmäßigkeit (Hays, 1994, Wittmann, 1985, S. 110ff.) zu Ungunsten der Einzelvarianzen und Fehlervarianzen gestärkt wird.

Letztlich kann man dieses Vorgehen als theoriegeleitete Akzentuierung von Zusammenhängen und Technik zur Konstanthaltung von unsystematischen Fehlern und unerwünschten Varianzanteilen betrachten. Ein weiterer Aspekt der für die Verwendung von Parcels spricht, ist die Reduktion der nötigen Stichprobengröße für sinnvolle Lösungen wobei dieser Aspekt aus der Forschung zu SEM entlehnt ist (Hall, Snell & Singer, 1999). Parceling ist alles andere als unumstritten, wobei sich die einzelnen Pro- und Kontra-Argumente auf die Frage zurückführen lassen ob eine stark empiristisch-konservative Wissenschaftsauffassung, d.h. kontra Parceling, oder eher eine pragmatisch-liberale Auffassung vertreten wird (Little et al., 2002, S. 152).

Ob Parceling hier in dieser Arbeit akzeptabel ist hängt in erster Linie davon ab, ob es eine theoretische Begründung gibt Parcels zu bilden. So schreiben Worthington & Whittaker (2006) ebenso wie Kline (2005, S. 197), dass im Rahmen der Entwicklung einer Skala eher von Parceling abzusehen ist, da es vorhandene Interitem-Zusammenhänge verschleiern könnte. Demnach soll Parceling in dieser Arbeit nur zur Testung von begründeten Strukturhypothesen verwendet werden, jedoch nicht zur exploratorischen Ergründung der Konstruktbeschaffenheit. Einfach davon auszugehen, dass einzelne Aufgabengruppen Faktoren begründen (z.B. ein Faktor Prozentrechnen, ein Faktor Multiplikationsaufgaben etc.) – wie es bei einigen der unter Abschnitt 2.1 beschriebenen Tests der Fall war - ist sicherlich keine ausreichende theoretische Begründung. Bei der in Abschnitt 4.4 folgenden Reanalyse eines nicht rein rational entwickelten Mathematiktests wären theoretische Begründungen für eine Bildung bestenfalls vage, weshalb diese Technik erst in Abschnitt 8.3, bei der Endform eines neuen Tests mit theoretischer Fundierung, eingesetzt wird. Ein mehr technischer Aspekt, der auch aus einer Bildung von Parcels resultiert, sind spezielle Schätzverfahren für den Modellfit von Strukturgleichungsmodellen, die im folgenden Abschnitt besprochen werden.

4.3.2 Alternative SEM-Schätzverfahren

Neben dem bereits angesprochenem Parceling, dessen Anwendung auch den Autoren aktueller SEM-Software bewusst ist und von ihnen keineswegs prinzipiell abgelehnt wird, (Bentler, 2003) stellt die im Falle einer SEM-Lösung gewählte Schätzmethode eine

wichtige Rolle. Zwar gilt der am weitesten verbreitete Maximum Likelihood (ML) Algorithmus auch bei Verletzung seiner Vorraussetzungen z.B. der multivariaten Normalverteilung, als eher robust (Benson & Fleishman, 1994, S. 117; Satorra, 1990, S. 383), doch sollten zur Absicherung von Befunden die Ergebnisse auch mit anderen Schätzverfahren geprüft werden (vgl. z.B. Wagener, 2008). Die Verletzung der multivariaten Normalverteilung wird beispielsweise bei dichotomen Items immer und bei Parcels zumindest häufig der Fall sein. Den Königsweg würde an dieser Stelle das ADF (asymtotic distribution free) Verfahren darstellen, welches keine Verteilungsannahmen benötigt, doch verlangt es nach geradezu gigantisch großen Stichproben und kann überhaupt erst, rein rechnerisch, angewendet werden, wenn mindestens so viele Fälle vorhanden sind wie nicht-redundante Parameter in der Stichprobenkovarianzmatrix (Bentler & Yuan, 1999, S. 182). Die Anzahl nichtredundanter Parameter ergibt sich aus der Anzahl von Elementen im unteren (oder oberen) Dreieck der Stichprobenkovarianzmatrix einschließlich ihrer Hauptdiagonale und beträgt bei m manifesten Variablen $df_{ges} = (m(m+1))/2$ (Kline, 2005). Um den Zusammenhang von minimaler Stichprobengröße und Anzahl von manifesten Variablen zu verdeutlichen wurde er in Abbildung 13 abgetragen.

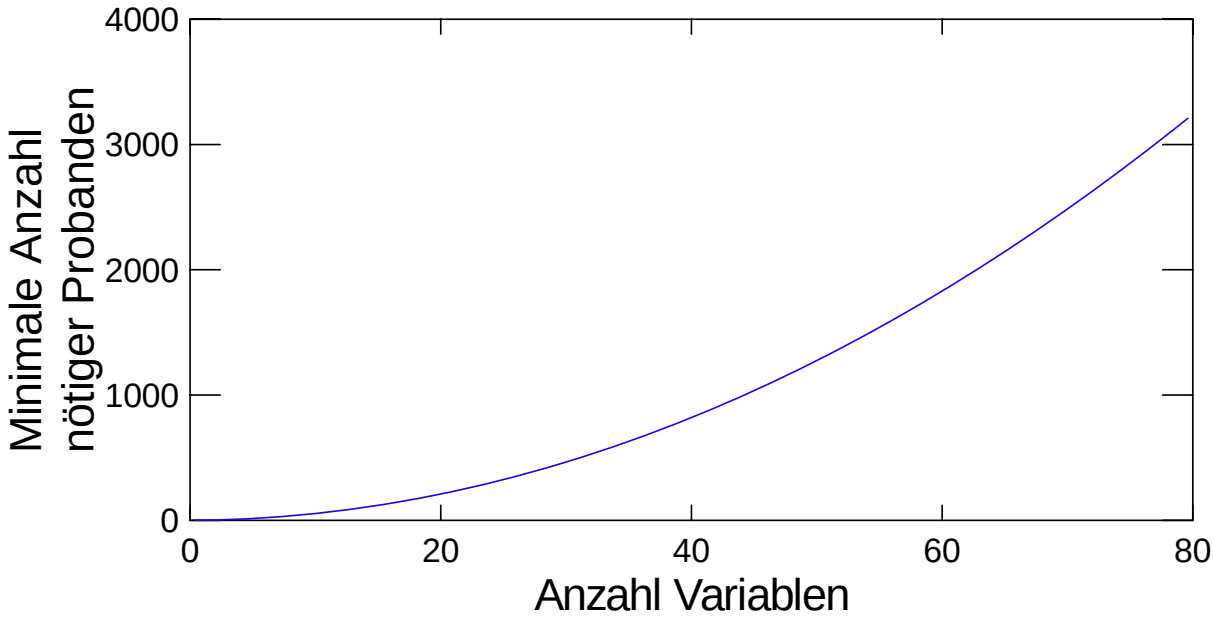


Abbildung 13 Notwendige Stichprobengröße im ADF-Verfahren.

Wie ersichtlich steigt das absolute Minimum für eine Anwendung des ADF-Verfahrens bereits ab etwa 40 Variablen astronomisch an. Dem entspricht auch die Schlussfolgerung

welche Peter Bentler (2003, S. 126) im Manual zu EQS zieht: “Unless sample size is really huge, the ADF test statistic based on an optimal weight matrix W yields distorted conclusions about the adequacy of a model”. Aus diesem Grund entwickelten Satorra und Bentler (1994) ein Korrekturverfahren für den ML-basierten χ^2-Fit Index und die Standardfehler, das umso stärkere Auswirkungen hat, je deutlicher die Daten schief verteilt sind. Zwar weist dieses Verfahren für Modelltests gute Eigenschaften auf, doch entspricht die Parameterschätzung nach wie vor dem ML-Verfahren, weshalb Finney und Distefano (2006) im Falle von kategorialen Variablen mit geordneten Kategorien ein weiteres Vorgehen namens WLSMV (weighted least squares with mean and variance adjusted) empfehlen. Hierbei handelt es sich um ein von Muthén (1993) entwickeltes Verfahren, bei dem von der weiter oben bereits erwähnten Gewichtungsmatrix W lediglich die Diagonale notwendig ist, was die benötigte Stichprobengröße immens verringert.

Eine empirische Prüfung des WLSMV-Verfahrens wurde von Muthén (1997) vorgenommen und führte selbst bei geringen Stichprobengrößen (N = 200) zu guten Ergebnissen. Um einen für konfirmatorische Faktorenanalysen im Bereich der psychologischen Forschung typischeren Datensatz zu simulieren führten Beauducel und Herzberg (2006) ebenfalls eine Monte-Carlo-Studie durch. Sie kamen zu dem Schluss, dass insbesondere für kategoriale Variablen mit wenigen Kategorien WLSMV der ML-Methode klar überlegen ist.

Für diese Arbeit wird geschlussfolgert, dass bei der Prüfung eines SEM neben der weit verbreiteten ML-Methode, im Falle von kategorialen Variablen und Parcels, stets auch die WLSMV-Methode appliziert werden sollte.

4.3.3 Law of diminishing returns

Nicht nur im Bereich der Mathematikfähigkeiten, sondern auch in anderen Inhaltsdomänen und sogar für andere Fähigkeitsbereiche gilt, dass eine Binnendifferenzierung umso eher sinnvoll ist, je höher das Leistungsniveau der Testanden ist (Abad, Colom, Juan-Espinosa & Garcia, 2003; Deary et al., 1996; Detterman & Daniel, 1989). Eng verbunden mit dieser Frage der Konstruktbeschaffenheit in Zusammenhang mit dem Leistungsniveau der Probanden ist die bereits von Spearman (1927) aufgestellte Differenzierungs-Hypothese, welche besagt, dass der Anteil von G an der Korrelation zwischen mentalen Tests mit steigendem Level von G zugunsten eigenständiger Varianzanteile der einzelnen Tests zurückgeht. Fogarty und Stankov (1995) fanden jedoch heraus, dass Spearmans Gesetz nur unter bestimmten Bedingungen gültig ist und zwar bei dem Vergleich von sehr schlechten

Probanden (IQ $\leq$ 78) und den nächst besseren. Sie kommen letztlich zu der Schlussfolgerung, dass bei sehr guten Probanden die Korrelationen der Subtests deswegen abfallen, weil die Teilnehmer zu gut werden, um in allen Skalen den Rang zugeordnet zu bekommen, der ihrer Fähigkeit entspricht. Demnach würde der differentielle Effekt sensu Spearman (1927) verschwinden, wenn die Tests nur komplex genug sind. Letztlich führen damit Fogarty und Stankov (1995) Spearmans Gesetz auf Deckeneffekte zurück, was bedeutet, dass das Gesetz eine Funktion von Aufgabenkomplexität und Probandenfähigkeit darstellt (vgl. Saklofske, Yan, Zhu & Austin, 2008). Da Aufgabenschwierigkeit – und erst recht Komplexität – nur schwerlich losgelöst von der Personenfähigkeit definiert werden können, scheint diese Begründung arbiträr. Die teils widersprüchlichen Ergebnisse gehen mittlerweile so weit, dass in zwei Studien mit demselben Intelligenztest (Wechsler Adult Intelligence Scale III), in Spanien (Abad et al., 2003) und den USA (einschl. australischen und kanadischen Daten) (Saklofske et al., 2008) einmal das Gesetz nachgewiesen werden konnte, ein anderes Mal hingegen nicht. Hartmann und Reuter (2006) prüften, ob die unterschiedlichen Ergebnisse in diesem Forschungsbereich vielleicht auf verschiedene Methoden der Subgruppenbildung zurückzuführen sind (anhand von Subtests, oder allgemeinen Fähigkeitsscores) kamen jedoch zu folgendem Schluss: „The study could not confirm Spearman's „Law of Diminishing returns" for any of the methods applied and did not find any relevant differences across methods applied" (S. 47).

Saklofske et al. (2008) empfehlen schließlich ob der nach wie vor ungeklärten Befundlage die Anfertigung einer Meta-Analyse oder die Durchführung einer Längsschnittstudie, um die Frage nach der Gültigkeit des Gesetzes im Rahmen einer kognitiven Entwicklungshypothese zu beantworten.

Bezüglich dieser Arbeit bedeutet dies, dass für einen fertigen Test – gegeben den Fall die Struktur entspricht nicht den Erwartungen – geprüft werden sollte, ob dies lediglich ein Artefakt von Spearmans Gesetz darstellt. Insbesondere für die Skalen prozedurales Rechnen und komplexes Rechnen, gemäß Skalenkonzeption aus Abschnitt 3.1.6, könnte es sich lohnen die Trennbarkeit mit dem Fähigkeitslevel der Probanden in Bezug zu setzen.

4.4 Reanalyse eines an der Uni Mannheim entwickelten Tests

Im Rahmen des alljährlichen Experimentalpraktikums an der Universität Mannheim wurden beginnend mit dem Wintersemester 03/04 die Entwicklungsschritte zur Erstellung eines Mathematiktests geübt. Aus diesem Versuchen entstand eine erste Vorform (Orth,

2006), deren Weiterentwicklung durch Jung et al. (2007) die Basis für die hier reanalysierte Testform darstellt. Während die ersten Formen des Tests eher experimentellen Status aufwiesen und hier nicht behandelt werden, kann die letzte Entwicklungsstufe als Basis zur in Abschnitt 5 vorgenommenen Weiterentwicklung betrachtet werden.

4.4.1 Testaufbau

Die letzten Testformen wurden in erster Linie durch Prüfung und Integration der Hauptschulcurricula Bayerns, Hessens, und Schleswig-Holsteins sowie von Informationen seitens der Mitarbeiter des Arbeitsamtes und von ausbildenden Unternehmen erstellt. Die Zusammenstellung der Skalen *kaufmännisches Rechnen*, *graphisch-geometrische Fähigkeiten* und *mathematisches Grundwissen* ergaben sich eher aus praktischen Gesichtspunkten, weniger aus einer psychologischen Theorie und ließen sich faktorenanalytisch nicht bestätigen (Jung et al., 2007). Zwei Items der Skalen mathematisches Grundwissen und kaufmännisches Rechnen sind in Tabelle 5 dargestellt.

Tabelle 5 Aufgaben 1d (mathematisches Grundwissen) und 7d (kaufmännisches Rechnen) des Studententests (Jung, Kempf & Seggewiß, 2007; Orth, 2006).

Item	Aufgabentext	Lösung
1d	H. bezahlte 23,08€ für den Hummer. Es waren 0,3 kg. Jetzt möchte er wissen, wie viel er für ein Kilo zahlen müsste.	76,93€
7d	Herr Eidam bekommt eine Erbschaft, die er zu einem jährlichen Zinssatz von 5% anlegt. Nach einem Jahr erhält er 550€ Zinsen. Wie hoch war die Erbschaft?	11000€

So lässt sich Aufgabe 1d leicht wie folgt lösen: *23,08€ = 0,3x*

Dementsprechend erhält man für Aufgabe 7d das Ergebnis: *550€ = 0,05x*

Von der Art der Berechnung her sind die Aufgaben fast identisch. Der Hauptunterschied besteht darin, dass für Aufgabe 7d bekannt sein muss, dass *5% = 5 : 100 = 0,05* beträgt. Aus theoretischer Sicht ist es schwierig zu begründen, wieso diese Aufgaben zu verschiedenen Skalen gehören müssen. In erster Linie inhaltliche Einkleidung und dadurch hervorgerufene Schwierigkeit wäre vielleicht als Grund denkbar. So ergibt sich z.B., wenn man die Items 1a-d, 6a-d, 7a-d und 8a-d betrachtet, die alle ähnlich den Aufgaben in Tabelle 5 sind, dass nur eines der Items aus Aufgabengruppe 1 die höchste Korrelation zu einem Item der eigenen Gruppe aufweist und zwar Item 1b mit $r = 0{,}33$ ($p = 0{,}00$, $N =$

182). Auch wieso Aufgaben vom Typ: *Löse nach x auf: 5x=15* (Aufgabengruppe 5) und *Berechnen Sie den Flächeninhalt eines Kreises mit dem Durchmesser d = 6m* (Aufgabengruppe 9) zu einer Skala gehören müssen, ist eher schwer zu begründen. Andere Aufteilungen wären also durchaus denkbar. Die bisher erwähnten Skalen weisen ein Cronbach's α von α = 0,71 (mathematisches Grundwissen, 18 Aufg.), α = 0,77 (kaufmännisches Rechnen, 17 Aufg.) und α = 0,64 (graphisch-geometrische Fähigkeiten, 16 Aufg.) auf, was Zweifel an der Homogenität der Skalen nahe legt (Horst, 1971, S. 282). Die Tatsache, dass Cronbach's α für die Skala 9 alleine einen Wert von α = 0,64 aufweist und sich durch Hinzufügen der Aufgabenreihe 5 nicht erhöhen lässt spricht statistisch gegen die Bildung einer gemeinsamen Skala, so müsste sich durch Hinzufügen von homogenen Items der Anteil der Kovarianz an der Gesamttestvarianz erhöhen und demnach α einen größeren Wert annehmen.

An dieser Stelle soll nicht versucht werden nach (ungefragten) Begründungen für diese oder jene Skalenstruktur zu suchen. Vielmehr gilt es die vorliegenden Daten zu verwenden, um mögliche andere Strukturen zu erkennen. Wie in Abschnitt 4.2.9 dargelegt, wird dafür zunächst ein Blick auf klassische Kennwerte geworfen, bevor komplexere Verfahren eingesetzt werden.

4.4.2 Klassische Itemkennwerte

Bei einer Analyse des Range der Daten (Hays, 1994), die lediglich in dichotomer Form vorliegen, zeigte sich, dass für einen Probanden bei einem Item (A9_F) mit dem Wert 10 ein Eingabefehler vorliegt. Der plausibelste Wert beträgt 1 und eine dementsprechende Korrektur wurde vorgenommen. Die klassischen Itemkennwerte sind Anhang 12.1.1 zu entnehmen da sie, bis auf zwei Items, relativ unauffällig sind.

Zwar ergibt sich für Cronbach‘s *α* des Gesamttests aufgrund der hohen Itemanzahl eine durchaus akzeptable Höhe von α = 0,87, doch wird an den Trennschärfekoeffizienten deutlich, dass es sich um ein nicht sonderlich homogenes Konstrukt zu handeln scheint. Neben der Tatsache, dass einige Items Trennschärfen nahe Null aufweisen (9J, 7D, 1A), was auch nicht immer durch extreme Schwierigkeiten erklärbar ist (z.B. Item 1A mit p = 0,69), bestehen zu wenig hohe Trennschärfen. Auffällig ist, dass für ein Item (1A) die interne Konsistenz steigt, nachdem das Item entfernt wird (von 0,87 auf 0,88) und darüber hinaus ein Item (9J) eine Trennschärfe von $r_{it} = 0$ aufweist. An dieser Stelle wird das Problem der nicht im Datensatz kodierten Missings sichtbar. So handelt es sich bei dem

Item (9J) um das vorletzte des Tests. Es wäre möglich, dass die Schwierigkeit des Items falsch eingeschätzt wird, weil viele schlechte Probanden aus Zeitgründen nicht bis zu diesem Item gekommen sind. Diese beiden Items (1A und 9J) werden für die folgenden Analysen entfernt, da sie die Ergebnisse verfälschen könnten.

4.4.2.1 DIMTEST und DETECT

Um zunächst zu prüfen, ob entsprechend der DIMTEST Logik (Stout, 1987) die Hypothese der Eindimensionalität zurückgewiesen werden kann und es möglich ist, Hypothese 1 aus Abschnitt 4.1 anzunehmen, wird das Verfahren im explorativen Modus durchgeführt. Hierbei ergibt sich eine T-Statistik von T = 4,27 (p = 0,00), womit von Multidimensionalität ausgegangen werden kann. Der Assessment Subtest, der von DIMTEST automatisch so ausgewählt wird, dass die Items maximal homogen sind (zur logischen Basis vgl. Abschnitt 4.2.7) besteht aus den in Tabelle 6 abgetragenen Aufgaben.

Tabelle 6 Nummerierung der AT-Test Items in DIMTEST, Benennung im Test und Trennschärfen.

Benennung im Test																			
2A	2C	2D	2E	2F	2G	2H	2I	5D	5E	6A	7B	7C	9A	9B	9C	9F	9G	9H	9K
Trennschärfe korrigiert																			
0,31	0,52	0,24	0,44	0,40	0,29	0,28	0,30	0,37	0,22	0,32	0,36	0,46	0,23	0,32	0,30	0,40	0,08	0,41	0,35
DIMTEST AT-Nummer																			
4	6	7	8	9	10	11	12	26	27	28	33	34	40	41	42	45	46	47	49

Betrachtet man die von dem Programm vorgeschlagene Aufteilung als Skala ergibt sich ein Cronbach's α von α = 0,77. Der eher niedrige Weg relativiert sich, da keine der ursprünglichen Skalen einen höheren Wert erreichte (höchster Wert kaufmännisches Rechnen mit α = 0,77).

Das Verfahren DETECT (Zhang & Stout, 1999) schlägt mit den Standardeinstellungen eine Lösung bestehend aus 4 Faktoren vor, die in Abbildung 14 dargestellt ist. Wie ersichtlich wurde versucht die Aufgaben den in Abschnitt 3.1.6 aufgestellten Skalen zuzuordnen. Die Items A5c und A5e wären eigentlich der Skala komplexes Rechnen zuzuordnen. Da sie jedoch die einzigen Items des Tests zu dieser Skala darstellen würden werden sie im folgenden der Skala prozedurales Rechnen zugeordnet.

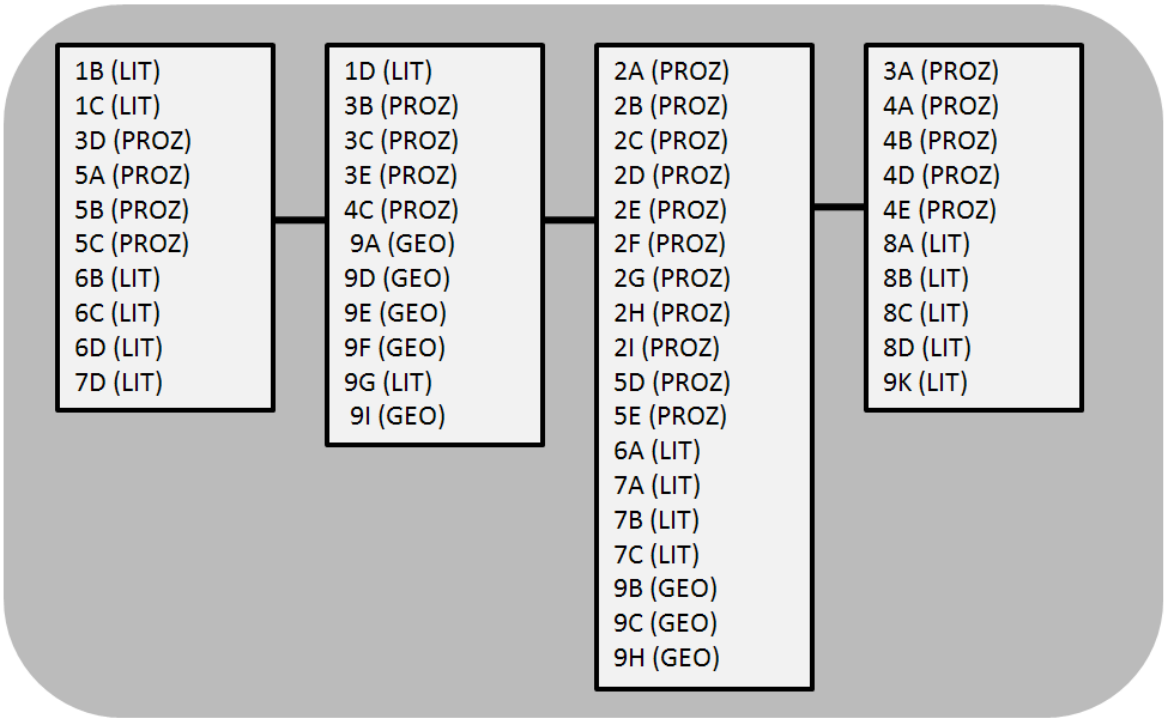


Abbildung 14 Von DETECT vorgeschlagene Cluster. Die Verbindungslinien zwischen Clustern verdeutlichen, dass sie nicht unabhängig sind LIT (mathematische Literalität), PROZ (prozedurales Rechnen), GEO (Geometrie und grafische Funktionen).

Es handelt sich um eine sehr grobe Zuordnung, die den aufgestellten Skalen nur im Sinne einer Analogie entspricht. Dennoch schienen die Skalenonzeptionen präzise genug, um eine Zuordnung zu treffen und somit Hypothese 2 zu bestätigen.

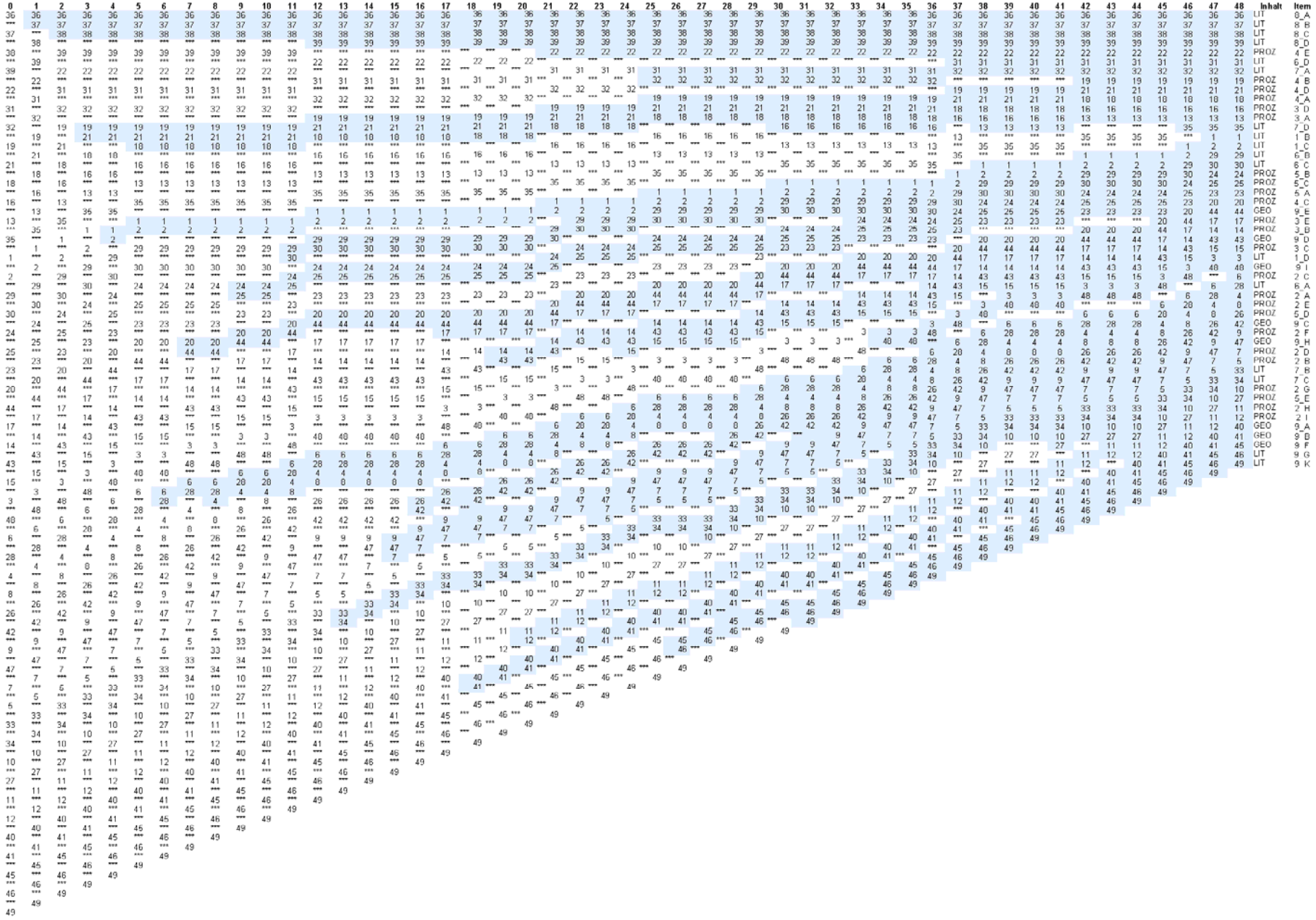
Der DETECT Index *D(P)* beträgt D = 0,47 was gemäß Gierl und Wang (2005, S. 6) einen schwachen Hinweis auf Multidimensionalität liefert (ab 0,51 moderat). Gleichzeitig liefert r_{max} mit einem Wert von r_{max} = 0,51 Evidenz für eine nicht vorhandene Einfachstruktur. Dies ist insofern interessant, da der *D(P)*-Index verglichen mit einer Untersuchung von Gierl und Wang (2005) zur Dimensionalität des amerikanischen Mathematik-SAT wesentlich höher ausfällt (dort nur D = 0,14) und zugleich r_{max} eine ähnliche Höhe wie in dieser Reanalyse aufweist. Die hier vorliegenden Daten scheinen also einen höheren Grad an Multidimensionalität aufzuweisen. Die inhaltliche Deutung der Cluster gemäß Abbildung 14 ist sehr schwer möglich. Dies liegt sicherlich auch an der starken Abweichung von der Einfachstruktur (gemäß r_{max}). Rein subjektiv erscheint es so, als wenn der erste Cluster (von l. nach r.) viele Items der Skala mathematische Literalität enthält (im Verhältnis zu deren Gesamtanzahl), der zweite Cluster die meisten Geometrie und grafische Fkt.-Aufgaben, der dritte Cluster vor allem Items des Bereichs prozedurales Rechnen und der vierte Cluster eine Mischung aus Items der Bereiche mathematische

Literalität und prozedurales Rechnen. Es wird deutlich, dass das DETECT-Verfahren an dieser Stelle an seine Grenzen stößt. Was fehlt sind die im faktoranalytischen Kontext verfügbaren Faktorladungen, die eine detailliertere Interpretation der Daten, sowohl über Eigenwerte, als auch über Betrachtung des Ausmaßes von Nebenladungen ermöglichen. Für diesen Abschnitt bleibt die Schlussfolgerung, dass der Mannheimer Test nicht eindimensional ist, also Hypothese 1 angenommen werden kann. Da keine Einfachstruktur zu erwarten ist, gestaltet sich eine, zumindest tendenzielle, Abgrenzung der verschiedenen Inhaltsbereiche als kompliziert.

4.4.2.2 HCA/CCPROX

Um das Vorgehen der HCA/CCPROX-Analyse (Marden et al., 1998) zu verdeutlichen wurde in Tabelle 7 ein so genanntes Dendrogramm erstellt, das die Bildung der Cluster nachvollziehbar macht. Hier existieren nach dem 17. Schritt immer noch 10 Cluster (ein Cluster mindestens 1 Item). Nach dem 34. Schritt sind es noch 11 Cluster. Während, wie in den obersten Zeilen sichtbar, vier der Aufgaben, die mathematischer Literalität zuzuordnen sind, bereits relativ früh einen Cluster bilden (12. Schritt) gilt dies nicht für die restlichen 13 textlastigen Aufgaben. Letztlich existiert zu keinem Zeitpunkt eine Lösung aus 3 oder 4 Clustern, die als inhaltlich sinnvoll interpretiert werden könnte. Tendenziell zeigt sich, dass zwar Items die aus inhaltlichen Gründen einen Cluster bilden sollten tendenziell auch dazu neigen (siehe z.B. die Aufgabengruppe 9_x, in den letzten Zeilen), dass dies jedoch erst sehr spät der Fall ist und dass diese Cluster kurz nach ihrer Bildung zu sehr globalen Clustern zusammengefasst werden. Dies ist ein Zeichen für deutliche Korrelationen zwischen den einzelnen Inhaltbereichen. Auf Basis der bisherigen Ergebnisse kann nur gefolgert werden, dass die Inhaltsgruppen mathematische Literalität, prozedurales Rechnen und Geometrie und grafische Fkt. in einigen Fällen sinnvolle Cluster bilden (zwei Beispiele wurden genannt) doch eine realistische 3 oder 4-Faktorlösung mit diesem Verfahren keine - inhaltlich begründbar - homogenen Cluster ergibt.

Tabelle 7 Dendrogramm. Clusterbildungen ab zwei Objekten wurden grau hinterlegt. Die oberste Zeile zeigt den Schritt an.



4.4.2.3 NOHARM

NOHARM (Fraser & McDonald, 1988; McDonald, 1999) bietet die Möglichkeit eines exploratorischen sowie eines konfirmatorischen Modus, wobei hier zunächst der exploratorische Modus angewendet wurde. Tabelle 8 zeigt die wichtigsten Fit-Indizes einer ein- bis 5-faktoriellen Lösung von NOHARM.

Tabelle 8 Fit-Indizes der exploratorischen NOHARM-Lösungen für ein bis 5-faktorielle Modelle.

Anzahl Faktoren	Tanaka (GFI)	RMSR
1	0,870	0,015
2	0,907	0,012
3	0,925	0,011
4	0,941	0,010
5	0,951	0,009

Alle RMSR-Werte weisen gemäß der Daumenregel (siehe Abschnitt 4.2.6) von McDonald und Fraser (1988), $RMSR_{GUTER\ FIT} \leq 0,2965$ (N=182), einen guten Fit auf. Vermutlich aufgrund der eher geringen Stichprobengröße, ist demnach aus dieser Empfehlung keine brauchbare Handlungsgrundlage ableitbar (stets überdurchschnittlich guter Fit). Für den GFI zeigt sich, wie zu erwarten war, eine Verbesserung des Index mit steigender Anzahl von Faktoren, wobei in diesen Index nicht die Anzahl der Modellparameter im Sinne der Sparsamkeit eingehen. Allgemein ist zu erkennen, dass der Abstand der GFI-Werte zwischen der 4 und 5-faktoriellen Lösung am geringsten ausfällt. Dies und die Tatsache das DETECT vier Dimensionen vorschlägt, werden zum Anlass genommen, nicht mehr als vier Dimensionen für diesen Test anzunehmen. Anhand McDonalds Daumenregel zum GFI (vgl. Abschnitt 4.2.6) kann die einfaktorielle Lösung ausgeschlossen werden (GFI $\leq$ 0,90). In Bezug auf eine zweifaktorielle Lösung handelt es sich um eine Ermessensfrage.

Für diesen Test soll nun geprüft werden, welchen Fit ein auf inhaltlichen Überlegungen aufgebautes Modell aufweist. Hypothese 2, für die die Items den Skalenkonzeptionen zugeordnet werden müssen, wurde bereits im vorherigen Abschnitt bestätigt; eine Zuordnung scheint möglich. Dabei stellte sich jedoch heraus, dass der Test fast keine Aufgaben enthält, die der Skala komplexes Rechnen zuzuordnen sind, weshalb auch im folgenden nur von prozeduralem Rechnen die Rede ist. Deshalb muss sich die konfirmatorische NOHARM-Lösung auf die verbleibenden drei Skalen stützen. Die Skalenzuordnung und Faktorenladungen einer NOHARM Lösung finden sich in der folgenden Tabelle 9.

Tabelle 9 Konfirmatorische, dreifaktorielle NOHARM-Lösung des Expra-Tests.

Item	Faktor			Skala
	F1	F2	F3	
1B	0,30			LIT
1C	0,37			LIT
1D	0,41			LIT
2A		0,45		PROZ
2B		0,37		PROZ
2C		0,52		PROZ
2D		0,26		PROZ
2E		0,66		PROZ
2F		0,46		PROZ
2G		0,43		PROZ
2H		0,53		PROZ
2I		0,46		PROZ
3A		0,47		PROZ
3B		0,34		PROZ
3C		0,66		PROZ
3D		0,54		PROZ
3E		0,71		PROZ
4A		0,52		PROZ
4B		0,80		PROZ
4C		0,56		PROZ
4D		0,66		PROZ
4E		0,60		PROZ
5A		0,60		PROZ
5B		0,56		PROZ
5C		0,51		PROZ
5D		0,55		PROZ
5E		0,58		PROZ
6A	0,42			LIT
6B	0,51			LIT
6C	0,30			LIT
6D	0,31			LIT
7A	0,55			LIT
7B	0,64			LIT
7C	0,62			LIT
7D				LIT
8A	0,74			LIT
8B	0,78			LIT
8C	0,85			LIT
8D	0,93			LIT
9A			0,50	GEO
9B			0,37	GEO
9C			0,47	GEO
9D			0,67	GEO
9E			0,63	GEO
9F			0,71	GEO

Tabelle 9 Fortsetzung.

	Faktor			Skala
	F1	F2	F3	
9G				LIT
9H			0,65	GEO
9I			0,85	GEO
9K		0,78		LIT

Anmerkung. N = 182, GEO = Geometrie und grafische Funktionen, LIT = mathematische Literalität, PROZ = prozedurales Rechnen. Leere Zellen stehen für Nullladungen, außer Item 7D und 9G, (Ladung ≤ 0,20). Faktorinterkorrelationen: r_{12} = 0,77, r_{13} = 0,68, r_{23} = 0,81. RMSR = 0,0143. Die Faktorladungen der dreifaktoriellen, explorativen Lösung finden sich in Anhang 12.1.2.

Auf den ersten Blick zeigt sich eine recht gute Passung, nur für Item 7D und 9G entstehen Ladungen kleiner a = 0,20, was bedeutet, dass diese Items mit einer derartigen Modellstruktur keine ausreichende Passung aufweisen. Das erfreuliche Bild wird jedoch dadurch getrübt, dass zwischen den Faktoren hohe Korrelationen bestehen (r_{12} = 0,77, r_{13} = 0,68, r_{23} = 0,81) und der GFI mit 0,88 schlecht ausfällt.

Einen möglichen Grund stellt die insgesamt sehr leistungsschwache Stichprobe dar, die - wie aus mündlichen Berichten der Testleiter hervorgeht - häufig Probleme hatte, den Test in der vorgegebenen Zeit abzuschließen. Inwiefern Zeiteffekte eine Rolle spielen kann durch die fehlende Kodierung von Missings leider nicht bestimmt werden (vgl. Abschnitt 4.4.2). Es zeigt sich deutlich, dass der Test nicht entwickelt wurde, um die vorgeschlagenen drei Facetten zu erfassen und das Hauptaugenmerk auf Vorhersagevalidität (Lienert & Raatz, 1994) lag.

4.5 Schlussfolgerungen

An dieser Stelle gilt es zu beurteilen, inwiefern die drei aufgestellten Hypothesen gemäß Abschnitt 4.1 bestätigt werden konnten. Sicher bestätigt werden konnte Hypothese H1, da sowohl DIMTEST, als auch eine exploratorische NOHARM-Lösung eindeutig eine mehrdimensionale Struktur nahe legen.

Hypothese H2 konnte ebenfalls bestätigt werden. Die aufgestellten Skalenkonzeptionen sind präzise genug, um die Aufgaben eines Mathematiktests den 4 (bzw. 3) Bereichen zuzuordnen. Dass praktisch keine Items, die zur Skala komplexes Rechnen passen, im Test enthalten waren, ändert daran nichts.

Ob Hypothese H3 erfüllt wurde, ist schwer zu beantworten. Bei einer Korrelation von r = 0,81 zwischen den Faktoren prozedurales Rechnen und Geometrie und grafische Fkt. fällt es schwer, von einer Trennbarkeit der beiden Bereiche auszugehen, wohingegen

die niedrigere Korrelation von r = 0,68 zwischen den Faktoren mathematische Literalität und Geometrie und grafische Fkt. eher eine Trennbarkeit nahe legt. Da jedoch gleichzeitig der GFI einen unzureichenden Wert annimmt, wird hier die Entscheidung getroffen, dass die Hypothese H3 nicht angenommen wird; sich die Skalen also nicht ausreichend trennen lassen. Hierfür gibt es mehrere mögliche Gründe, wie Eigenschaften des Instruments an sich, besondere Merkmale der Personenstichprobe (Leistung) und eventuell Datenqualität (Missingproblematik). Am wahrscheinlichsten scheint zum gegenwärtigen Zeitpunkt der erste Grund zu sein, da das Instrument nicht konstruiert wurde, um die 4 aufgestellten Skalen zu erfassen. Dies zeigt sich z.B. daran, dass Aufgaben die der Skala mathematische Literalität zugeordnet wurden, zwar am ehesten Textaufgaben darstellen, jedoch dafür relativ wenig Text enthalten, teils nur aus einem einzigen Satz bestehen. Auch ist eine vorhandene Zeichnung für einige der Geometrieaufgaben eher nebensächlich.
Die Ergebnisse der Reanalyse sprechen prinzipiell für den Versuch einen neuen Test, basierend auf den vier aufgestellten Skalen, zu konstruieren, womit im folgenden Abschnitt begonnen wird.

5 Erstellung einer neuen Testvorform

Die bisherigen theoretischen Annahmen zur Skalenstruktur und erste Vorab-Analysen werden in diesem Abschnitt integriert, um eine Vorform zu generieren, die im darauf folgenden Abschnitt 6 in der Zusammenstellung einer Endform resultiert.

5.1 Geltungsbereich und Zielgruppe

Der Altersbereich der Zielgruppe beginnt bei etwa 16 Jahren, was eine Orientierung an den Curricula der 9. und 10. Klasse von Haupt- und Realschule nahe legt. Da auch bereits vor der 9. Klasse wichtige Grundkenntnisse in Mathematik erworben werden (z.B. Division, Multiplikation, Prozentrechnung, etc.) spricht nichts dagegen, auch einige Aufgaben zu integrieren, die der 8. Klasse entsprechen, ebenso wie einige Aufgaben die eher in der 10. Klasse Gymnasium auftauchen (z.B die PQ-Formel zum Lösen quadratischer Gleichungen), wobei jedoch der Schwerpunkt stets auf den Inhalten der 9. und 10. Klasse liegen sollte.
Inwiefern eine Orientierung an Bildungsstandards und Lehrplänen zur Testkonstruktion möglich und sinnvoll ist, wird in den folgenden Abschnitten 5.2 und 5.3 geklärt.

5.2 Bildungsstandards und Lehrpläne

In einem Artikel aus dem Jahre 1998, mit dem Titel *Erasmus, Gates and the end of curriculum* sagte der britische Bildungsforscher William A. Reid das Ende der Curricula für das 20. Jahrhundert voraus. Reid (1998) stützte diesen Gedanken darauf, dass durch Internationalisierung und Modernisierung unserer Welt, er verweist als Beispiel auf E-Mails, Microsoft und IBM, nationale Lehrpläne immer mehr an Bedeutung verlieren, was im Endeffekt sämtliche Lehrpläne in ihrer Bedeutung zurückdrängte. Wie Hopmann (2000) herausarbeitet, ist jedoch diese Idee keineswegs neu, so wurden ähnliche Prophezeiungen bei diversen neuen Kommunikationstechnologien getätigt, die allesamt - gemessen an den Prophezeiungen - erschreckend wenig Einfluss geltend machten. Interessanterweise scheinen Lehrplanreformen eher selten zu spürbaren Einschnitten in der Unterrichtspraxis zu führen (Hopmann, 2000, S. 386). Ihre Wirkung bezieht sich vermutlich mehr darauf, Inhalte und Methoden, die als nicht mehr zeitgemäß angesehen werden, auszuschließen (Hopmann, 2000, S. 387). Hamburger, Horstkemper, Melzer und Tillmann (1999, S. 28) sprechen in diesem Kontext von einer Orientierungsfunktion, die Lehrpläne gegenüber den Lehrern erfüllen. Anhand einer repräsentativen Lehrerbefragung in Hessen zur damaligen Lehrplanrevision der Sekundarstufe I berichten die Autoren, dass Lehrplanreformen meist nicht freudig von den Lehrenden aufgenommen werden (Hamburger et al., 1999, S. 47). Schließlich werden die, teils über viele Jahre anhand von Unterrichtserfahrung aufgebauten, Curriculums-Skripten in Frage gestellt. Dazu passt, dass Lehrpläne anscheinend häufig innerhalb der Schulen (Hamburger et al. 2000 S. 150) in so genannte Arbeitspläne - die keineswegs curriculumskonsistent sein müssen - übersetzt werden.

Mittlerweile wurden, anscheinend als Reaktion auf die TIMSS und PISA-Untersuchungen in einigen Bundesländern die bisherigen inhaltszentrierten Lehrpläne durch so genannte Bildungsstandardpläne abgelöst (Wacker, 2008, S. 13). Für den Bereich Mathematik (ab 8. Klasse) mit den *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss* vom *15.10.2004* und *Den Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss* vom 4.12.2003, herausgegeben durch die Konferenz der Kultusminister [KMK]. Damit verbunden ist eine Orientierung am System-Output (nicht Input wie bei Lehrplänen), anhand der Benennung von konkreten Zielen. Der Hintergedanke ist es, sich dadurch die einzelschulische Autonomie zunutze zu machen, wobei auch für Bildungspläne noch Unsicherheit hinsichtlich ihrer

tatsächlichen (Steuerungs)-Wirkung bestehen (Wacker, 2008, S. 13). Zur Konzeption der Bildungsstandards ist nach Feltes und Paysen (2005, S. 35) zu sagen, dass die in diesem Rahmen postulierten *Kompetenzen* nichts Weiteres sind, als eine Abstraktion von Lerngegenständen. Es ist auch sicherlich kein Zufall, dass der bei TIMSS und PISA verwendete Kompetenzbegriff eine klare Passung zu dem Kompetenzbegriff im Rahmen der Bildungsstandards darstellt (Klieme et. al, 2003). Weiter heißt es bei Feltes und Paysen (2005, S. 156) Kompetenzen seien lediglich verallgemeinerte Beschreibungen dessen, was ein Schüler können sollte und für Fehldiagnosen zu abstrakt.

Inwiefern die Bildungsstandards im Bereich Mathematik für den Zweck der Testkonstruktion sinnvoll sind, kann letztendlich nur beurteilt werden, indem man die konkret vorliegenden Standards analysiert, was in den folgenden Abschnitten 5.2.1 und 5.2.2 der Fall ist. Ein Fazit zu den Bildungsstandards wird in Abschnitt 5.2.3 gezogen.

5.2.1 Bildungsstandards für den Hauptschulabschluss (Mathematik)

Die Bildungsstandards Mathematik für den Hauptschulabschluss der KMK (2005a) sollen "Anhaltspunkte" (S. 6) für die Gestaltung des Mathematikunterrichts liefern. Es wird von sechs allgemeinen, mathematischen Kompetenzen ausgegangen, deren Trennung jedoch mitunter schwer fällt. Als Beispiel seien Kompetenz eins *mathematisch argumentieren* und Kompetenz sechs *kommunizieren* genannt (KMK, 2005a, S. 7). Anhand der auszugsweisen Auflistung in Tabelle 10 wird deutlich, dass eine sehr starke Überlappung zwischen den Kompetenzbereichen anzunehmen ist.

Tabelle 10 Auszüge zweier mathematischer Kompetenzbereiche aus den Bildungsstandards für Mathematik (KMK, 2005a)

Kompetenzen	
mathematisch argumentieren	kommunizieren
Fragen stellen, die für Mathematik charakteristisch sind...	...Texte zu mathematischen Inhalten verstehen und überprüfen
Mathematische Argumentationen entwickeln...	Fachsprache adressatengerecht verwenden
Lösungswege beschreiben und begründen.	Überlegungen, Lösungswege bzw. Ergebnisse dokumentieren, verständlich darstellen und präsentieren....

Das Spektrum daraus ableitbarer Aufgaben ist vielschichtig, weshalb versucht wird, durch inhaltsbezogene mathematische Kompetenzen eine gewisse Konkretisierung herbeizuführen (KMK, 2005a, S. 9). Hier heißt es z.B. bei der Leitidee Raum und Form: "operieren gedanklich mit Strecken, Flächen und Körpern" (KMK, 2005a, S. 10). Demgegenüber heißt es bei der Leitidee Messen: "ermitteln Flächeninhalt und Umfang von Rechteck, Dreieck und Kreis..." (KMK, 2005a, S. 10). Es stellt sich die Frage, ob das Ermitteln des Flächeninhalts kein gedankliches Operieren mit dem Vorliegenden, z.B. Dreieck, erfordert. Es ließen sich noch weitere Beispiele anführen, doch es reicht hier festzustellen, dass durch die inhaltsbezogenen Kompetenzen eine gewisse Konkretisierung erreicht wird, diese jedoch für eine Testkonstruktion unzureichend erscheint. Die 15 gegebenen Aufgabenbeispiele enthalten genau zwei Aufgaben, die keine Einbettung in eine alltägliche Situation beinhalten. Ähnlich wie in Abbildung 15 enthalten fast alle Aufgaben einen substantiellen Anteil von Text, der gerade bei Hauptschülern ohne Deutsch als Muttersprache zu niedriger Beurteilung Mathematischer Kompetenz (im Sinne der KMK) führen kann.

(13) Mogelpackung

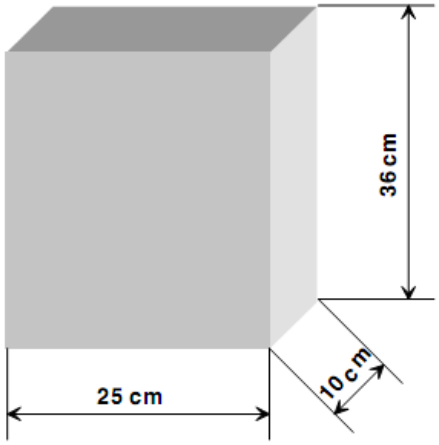
Aufgabenstellung:

Eine Cornflakes-Packung hat die in der Abbildung angegebenen Abmessungen.

a) Berechne das Fassungsvermögen einer solchen Packung.

Peter stellt fest, dass die gekaufte Packung Cornflakes nur zu $\frac{5}{6}$ der Höhe gefüllt ist.

b) Welche Gründe könnten für die größere Packung sprechen?



c) Wie viel Pappe könnte die Herstellerfirma pro Verpackung einsparen?

Abbildung 15 Beispielaufgabe der Bildungsstandards (Hauptschule).

Dies wäre nicht gravierend, wenn explizit zwischen textgebundenen und textfreien Aufgaben unterschieden würde, was jedoch nicht der Fall ist. Die textlastigen Aufgabenformate ziehen sich durch alle Inhaltsbereiche und alle Kompetenzen. Gravierender für eine Testkonstruktion, davon abgesehen, dass das Verhältnis von Höhe und Breite gemäß Zeichnung nicht dem Verhältnis der Zahlen entspricht (je nach Sichtweise müsste der Quader höher oder weniger breit sein), ist jedoch bei dem

Beispiel in Abbildung 15 die Teilaufgabe b. Die KMK (2005a, S. 29) sieht als richtige Lösung die Antworten *Vortäuschen eines großen Volumens* oder *technische Bedingungen beim Einfüllen der Cornflakes an.* Was, wenn ein Schüler als Antwort schreiben würde: Eine große Packung fällt im Regal mehr auf. Wäre diese Antwort noch korrekt oder bereits falsch? Dieser sehr praxisorientierte Aufgabentyp erinnert an einige TIMSS und PISA Aufgaben (vgl. Abschnitt 2.2). Derartige Aufgaben eignen sich jedoch kaum für einen Leistungstest, da die richtige Lösung zu ungenau definiert ist, wodurch das Kriterium der Aufgabenobjektivität nicht erfüllt ist (Lienert & Raatz, 1994, S. 29).

5.2.2 Bildungsstandards für den mittleren Schulabschluss (Mathematik)

Die Bildungsstandards für den mittleren Schulabschluss orientieren sich laut KMK (2004a, S. 4) an den einheitlichen Prüfungsordnungen in der Abiturprüfung. Die folgende Abbildung zeigt links die zu erwerbenden Kompetenzen mit dem Hauptschulabschluss am Ende der 9. Hauptschulklasse (KMK, 2005a, S. 7) und rechts die Kompetenzen die mit dem Erwerb des mittleren Schulabschlusses (KMK, 2004a, S. 7) bei den Schülern vorhanden sein sollten.

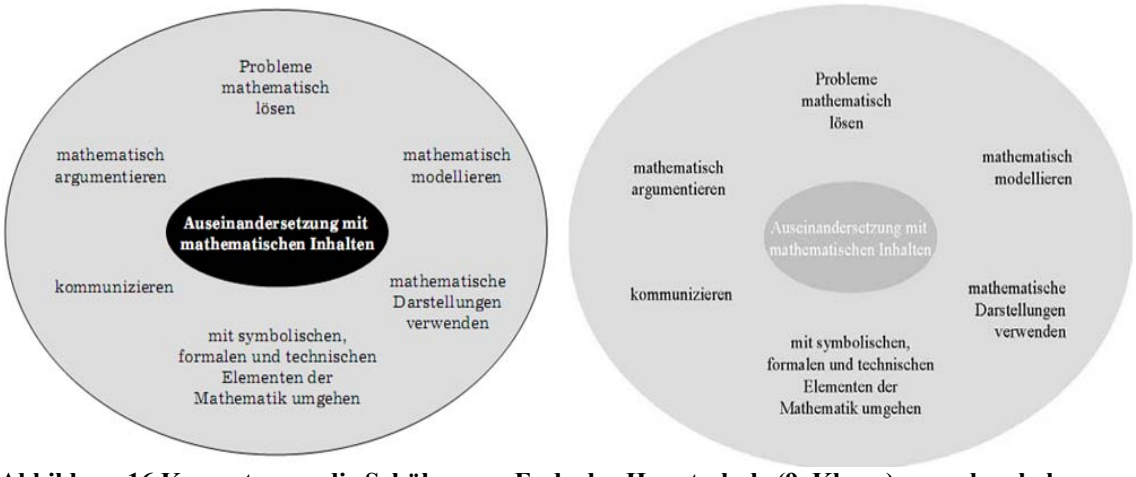


Abbildung 16 Kompetenzen die Schüler zum Ende der Hauptschule (9. Klasse) erworben haben sollten (links) und Kompetenzen die Schüler mit dem mittleren Schulabschluss erworben haben sollten (rechts). Quelle: KMK (2004a, 2005a).

Trotz der unterschiedlichen Konzeption der Bildungsstandards sind die Grafiken, abgesehen von ihrer Farbe, identisch. Ebenfalls wortgleich sind die jeweils zweiseitigen verbalen Umschreibungen der in Abbildung 16 dargestellten Kompetenzen (KMK, 2004a, S. 8 bis 9; KMK, 2005a, S. 7 bis 8). Dies heißt nichts anderes, als dass dieselben

Kompetenzen mit dem Hauptschulabschluss und dem mittleren Schulabschluss angenommen werden.

Der Unterschied zwischen den KMK-Konzepten zum mittleren- und Hauptschulabschluss ergibt sich einzig durch die Auflistung zu den fünf Leitideen *Zahl, Messen, Raum und Form, funktionaler Zusammenhang, Daten und Zufall (*KMK, 2004a). Die Leitideen existieren in beiden Konzepten unter gleichen Namen, enthalten jedoch teilweise unterschiedliche Beschreibungen. Für jede Leitidee sind im Falle des Bildungsstandards für den mittleren Schulabschluss mehr und augenscheinlich schwierigere Inhalte angeführt. Auch durch die Beispiele wird ersichtlich, dass eine höhere Fähigkeit erforderlich sein sollte. Die folgende Abbildung 17 zeigt exemplarisch eine der Beispielaufgaben.

(7) Holzbestand

Aufgabenstellung

Der Holzbestand eines Waldstückes beträgt 80000 m³. Er wächst jährlich um 2,5 %.

a) Berechnen Sie den Holzbestand nach zwei Jahren.

b) Stellen Sie die Entwicklung des Holzbestandes für die nächsten 20 Jahre mit Hilfe eines Tabellenkalkulationsprogramms dar.

Folgender Tabellenkopf ist dafür vorgegeben:

	A	B	C	D
1	Anzahl der Jahre	Holzbestand		
2	0	80000		
3	1			
4				
5				

Wie viele Jahre würde es dauern, bis sich der Holzbestand verdoppelt hat?

c) Geben Sie zur Beantwortung der Frage in Aufgabe b) eine weitere Lösungsmöglichkeit ohne PC an.

d) Die tatsächliche Entwicklung des Holzbestandes kann von der berechneten abweichen. Geben Sie dafür Gründe an.

Abbildung 17 Aufgabenbeispiel 7 aus den Bildungsstandards für den mittleren Schulabschluss. Quelle: (KMK, 2004a, S. 25).

Abgesehen von der textlastigen Einkleidung der Aufgabe (siehe auch Abschnitt 5.2.1) fällt für Teilaufgabe d auf, dass deren Lösung ähnlich wie im vorangegangen Beispiel in Abschnitt 5.2.1, nicht klar definiert ist. Die KMK (2004a, S. 276) schlägt als Beispiel

vor, dass ein gleichmäßiges Wachstum über einen längeren Zeitraum idealisiert wäre. Ebenso denkbar wäre es, die Forstwirtschaft mit einzubeziehen, die maßgeblichen Einfluss auf die Geschwindigkeit des Waldwachstums haben könnte. Kurzum, auch hier ist keine Objektivität (Lienert & Raatz, 1994) gegeben. Typisch für die meisten Aufgaben ist eine inhaltliche Einkleidung und - damit verbunden - ein hoher Textanteil.

5.2.3 Fazit zu den Bildungsstandards Mathematik

Zur Entwicklung von Bildungsstandards in Deutschland existiert eine seitens der KMK in Auftrag gegebene 228-seitige Expertise von Eckard Klieme et al. (2003). Für die Konstruktion eines psychologischen Leistungstests können Bildungsstandards dennoch als problematisch angesehen werden. So heißt es in der offiziellen Expertise von Klieme et al. (2003, S. 85), dass Bildungsstandards im Kontext der Testentwicklung Kriterien im Sinne einer kriteriumsorientierten Leistungsmessung darstellen. Weiter heißt es dort, dass es nicht darum geht, die Position eines Schülers - wie im Falle der normorientierten Messung (Lienert & Raatz, 1994) - im Vergleich zu anderen Testpersonen (Normstichprobe) festzustellen. Auf dieselbe Expertise stützen sich Ehmke, Leiß, Blum und Prenzel (2006, S. 222) in einem Artikel zur Entwicklung von Testverfahren für die Bildungsstandards (in Mathematik), indem die konkrete Testentwicklung daraus bestand, dass fünf Regionalgruppen (hauptsächlich Lehrer, S. 226) über 1000 Aufgaben entwickelten, die von einer Bewertungsgruppe (Fachdidaktiker und Erziehungswissenschaftler) auf einer vierstufigen Skala bewertet wurden. Die darauf folgenden Analysen werden beschrieben als Auswahl nach Trennschärfe und Schwierigkeit. Auf welcher Basis die Trennschärfen berechnet wurden, d.h. in Bezug zur jeweiligen Leitidee oder Kompetenzstufe, oder beidem bleibt unklar (vgl. Ehmke et al., 2006, S. 230).

In dieser Arbeit wird, gemäß Rost (2004a, S. 41) die Meinung vertreten, dass es sich bei der Normierung um ein wichtiges Gütekriterium handelt, da die Interpretation weniger von der subjektiven Festlegung von (kriterialen) Standards abhängig ist und objektiver wird. Es muss betont werden, dass es hierbei nicht um ein besser- oder schlechter-Urteil handelt, sondern für eine normorientierte Individualdiagnostik Bildungsstandards als zu subjektiv und allgemein angesehen werden. Rost (2004b) führt die Diskussion um Bildungsstandards auf die Diskussion um das so genannte Mastery Learning (Bloom, 1976) zurück, welches in den 70er Jahren an Popularität gewann, sicherlich auch, da die Quantifizierung von Leistung und Notengebung in dieser Zeit kritisiert wurde. Rost

(2004b) sieht das Hauptproblem in der teils nicht gegebenen Kompatibilität von qualitativen Standards und quantitativer Messung. Zwar gibt es prinzipiell im IRT-Bereich Methoden, die dieses Dilemma auflösen können (z.B: LLTM in Kombination mit Mixed-Rasch-Modellen), doch ist der zusätzlich Aufwand meist enorm. Drei Jahre später fasst Rost zusammen, dass Aufgaben zur Messung von Kompetenzen praktisch allen Maximen für die Entwicklung von Testaufgaben für Leistungstests widersprechen Rost (2007, S. 63). Solche Aufgaben haben häufig mehr als eine einzige Lösung, sind nicht homogen und meist nicht in kurzer Zeit zu bearbeiten. Dies führt dazu, dass die Information ob eine Aufgabe gelöst wurde oder nicht für die Auswertung nicht ausreicht (Rost, 2007, S. 72).

Interessanterweise werden durch Bildungsstandards Lehrpläne keineswegs überflüssig, sondern erhalten die Funktion von so genannten Kerncurricula. Als Ergebnis einer Kultusministerkonferenz des Jahres 2004 in Nordrheinwestfalen wird angegeben, dass die Schulen gerne wissen würden, welche Funktion eigentlich die neuen Kernlehrpläne haben und was sie von üblichen Lehrplänen unterscheidet. In einem Argumentationspapier der KMK (KMK, 2005b) heißt es hierzu, dass die Bildungsstandards nicht die ganze Breite eines Lernbereiches abdecken, sondern vielmehr fachliche und fachübergreifende Basiskompetenzen beschreiben.

Ungeachtet des teils deutlichen Unterschiedes von intendiertem, implementiertem und erreichtem Curriculum können Lehrpläne eine große Hilfestellung darstellen, um das so genannte Itemuniversum abzubilden (Rost, 2004a, S. 55). Damit ist gemeint, dass Lehrplaninhalte häufig konkreter sind und eine große Chance besteht, dass die daraus abgeleiteten Aufgaben eine gewisse Validität aufweisen.

In den folgenden Abschnitten 5.3.1 bis 5.3.4 werden die Lehrpläne der bevölkerungsreichsten Bundesländer Deutschlands analysiert und in Abschnitt 5.3.5 ein Fazit zur möglichen Verwendung im Rahmen einer Testkonstruktion gezogen.

5.3 Exemplarische Betrachtung vorhandener Curricula

Alle aktuell verfügbaren Lehrpläne der deutschen Bundesländer sind im Lichte der KMK-Beschlüsse (KMK, 2004a; KMK, 2005a) zu betrachten, die implizit eine Art Kerncurricula (Leitideen und deren Beschreibungen) vorschreiben.

Bayern hat sich in den bisherigen PISA-Studien im Fach Mathematik als Musterschüler gezeigt (PK, 2003). Dies könnte auch mit den Lehrplänen des Landes zu tun haben,

weshalb es nahe liegt, die Lehrpläne Bayerns genauer zu betrachten. Ein weiterer Aspekt geht aus der Bevölkerungsverteilung der Bundesländer hervor. Bayern, als flächenmäßig größtes Bundesland, wird nur noch von dem wesentlich kleineren Nordrhein-Westfalen (NRW), in Bezug auf die Bevölkerungszahl, übertroffen.

Auf die vier Bundesländer mit der höchsten Bevölkerungsanzahl (Bayern, Nordrhein-Westfalen, Baden-Württemberg und Niedersachsen) verteilen sich mit ca. 49,2 Millionen fast 60% der Bevölkerung Deutschlands. Abbildung 18 zeigt die Situation für die Anzahl der Schüler in Berufsschulen und allgemein bildenden Schulen für alle Stufen (da keine Auflistung nach Stufen vorlag).

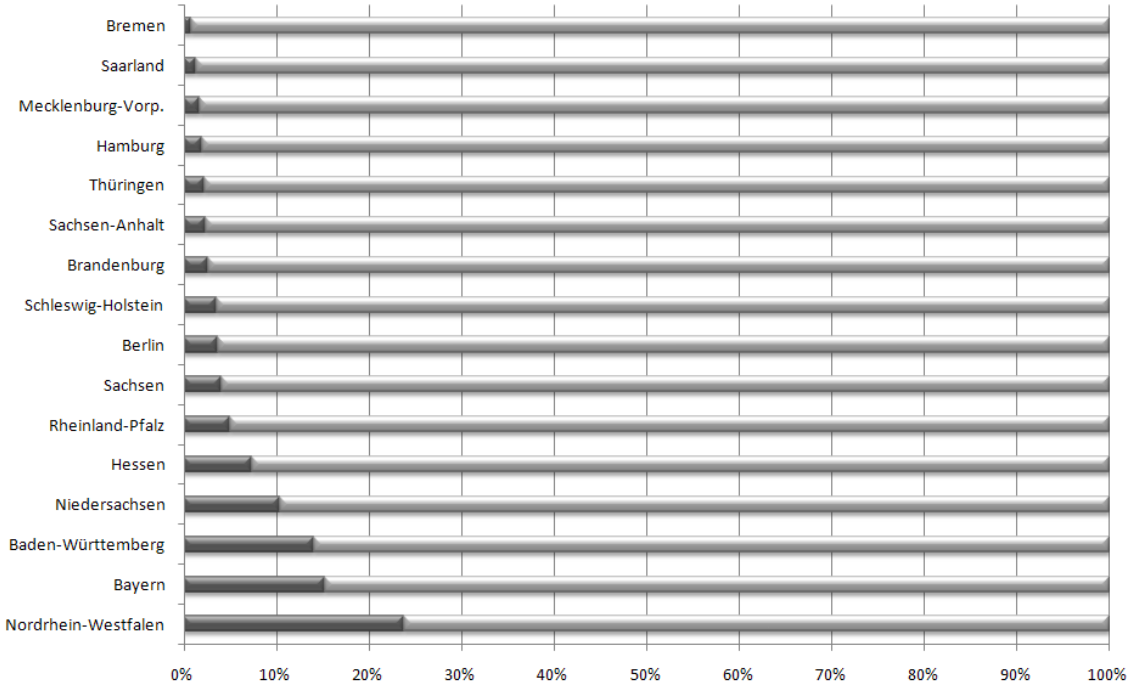


Abbildung 18 Verteilung der Schüler allgemeinbildender- und Berufsschulen auf die Bundesländer. Stand 2007, N ≈ 12,1 Millionen. Quelle: Statistisches Bundesamt.

Es befinden sich etwa 64% der Schüler Deutschlands in den vier schülerreichsten Bundesländern Deutschlands. Deswegen scheint es als Heuristik ausreichend zu sein, die Lehrpläne von NRW, Bayern, Baden-Württemberg und Niedersachen genauer für das Vorhaben der Testentwicklung zu prüfen. Das Hauptaugenmerk liegt in den folgenden Abschnitten auf den entsprechenden Lehrplänen der 9. und 10. Jahrgangsstufe für Haupt- und Realschulen, auszugsweise werden auch Gymnasiallehrpläne der 9. und 10 Klasse betrachtet.

5.3.1 Lehrpläne Nordrhein-Westfalens

Der Mitte 2008 aktuellste Lehrplan für die Hauptschule in NRW (NRW, 2004a) trat August 2005 in Kraft und wird in NRW als Kernlehrplan bezeichnet. In den

Kernlehrplänen wird direkt auf die Beschlüsse der KMK (2005a) zu den Bildungsstandards Bezug genommen. Gemäß des Kernlehrplans (NRW, 2004a, S. 13) sollten Schüler mit Erwerb des Hauptschulabschlusses (dort nach der 10. Klasse) über 8 Kompetenzen verfügen und zwar Argumentieren / Kommunizieren, Problemlösen, Modellieren, Werkzeuge (verwenden), Arithmetik / Algebra, Funktionen, Geometrie und Stochastik. Die Umschreibung dieser Kompetenzen ist teilweise etwas unscharf, so heißt es z.B. bei Argumentieren/Kommunizieren: "Sie nutzen verschiedene Arten des Begründens und Überprüfens (Plausibilität, Beispiele, Argumentationsketten)" (NRW, 2004a, S. 14). An anderen Stellen hingegen scheint eine Umsetzung hin zu einer Testaufgabe leichter realisierbar, so heißt es bei Geometrie "Sie schätzen und bestimmen Winkel, Längen, Flächeninhalte, Oberflächen und Volumina" (NRW, 2004a, S. 16). In dem Kernlehrplan werden explizit Kompetenzerwartungen für das Ende der 8. und 10. Klasse angegeben, nicht jedoch für die 9. Klasse. Für die 8. und 10. Klasse sind die Kompetenzerwartungen deutlich präziser als die Kompetenzbeschreibungen. Für Problemlösen finden sich zwar noch vage Beschreibungen wie "zerlegen Probleme in Teilprobleme" (S. 27), für andere Bereiche hingegen existieren Beschreibungen wie "vergrößern und verkleinern einfache Figuren maßstabsgetreu" (S. 30, Geometrie). Es werden für die 10. Klasse drei Aufgabenbeispiele gegeben, die allesamt in das Alltagsgeschehen eingekleidete Textaufgaben sind. Einmal geht es um eine Pizzeria-Speisekarte, dann um Legosteine für die Messungen und Berechnungen vorgenommen werden müssen und schließlich um einen Heißluftballon - umgeben von einer Gruppe Menschen, für den geschätzt werden muss, wie viel Volumen der Heißluftballon "in etwa" enthält (NRW, 2004a, S. 49). Die Benennung der acht geforderten Kompetenzen ist in den Lehrplänen für die Hauptschule und die Realschule wortgleich (NRW, 2004a, S. 11; NRW, 2004b, S. 11). Bei genauer Durchsicht gelang es in Bezug auf deren Beschreibungen vier Unterschiede zu identifizieren: Bei Arithmetik / Algebra sowie Funktionen wurden für die Realschule (NRW, 2004a) linear und quadratische Sachverhalte um exponentielle ergänzt, bei Geometrie der Satz von Thales und bei Stochastik die Laplace-Regel hinzugefügt. Von diesen Satzteilergänzungen abgesehen sind die Beschreibungen der Kompetenzen anscheinend wortgleich.

5.3.2 Lehrpläne Niedersachsens

Im Hauptschullehrplan Niedersachsens, herausgegeben vom niedersächsischen Kultusministerium (NK), wird explizit zwischen prozessbezogenen- und

inhaltsbezogenen Kompetenzen unterschieden (NK, 2006a, S. 6), was einen klaren Unterschied zu den Plänen NRWs in Abschnitt 5.3.1 darstellt.

Durch die Aufteilung in die *prozessbezogenen Kompetenzen* modellieren, problemlösen, argumentieren, kommunizieren, darstellen, symbolische, formale und technische Elemente sowie die *inhaltsbezogenen Kompetenzen* Zahlen und Operationen, Größen und Messen, Raum und Form, Funktionaler Zusammenhang und Daten und Zufall soll eine Konkretisierung der Ziele des Mathematikunterrichts erreicht werden (NK, 2006a, S. 8). Darüber hinaus wird zwischen drei Aufgabentypen unterschieden, den eher technischen- (z.B. kalkülhafte Durchführung, Typ I), rechnerischen Problemlöse- und Modellierungsaufgaben (vor allem Textaufgaben, Typ II) sowie den begrifflichen Problemlöse- und Modellierungsaufgaben (u.a. logisches Argumentieren, Typ III), die neben didaktischen Funktionen auch zur individuellen Leistungsfeststellung genutzt werden sollen (NK, 2006a, S. 10). Für den prozessbezogenen Aufgabenbereich lassen sich bestenfalls Tipps zum Erstellen von Testaufgaben ableiten, als Beispiel sei hier aus dem Bereich Argumentieren der Unterpunkt *hinterfragen mathematischer Aussagen* in Tabelle 11 präsentiert.

Tabelle 11 Auszug der Kernkompetenzumschreibung für den Prozessbereich *Argumentieren*, Unterkategorie: hinterfragen mathematischer Aussagen (NK, 2006a, S. 18).

Schuljahr	
8	9/10
Präzisieren Vermutungen, um sie mathematisch prüfen zu können	Unterscheiden zwischen experimentell gewonnen Vermutungen und logisch gewonnen Argumenten
Stellen Fragen „Gibt es Gegenbeispiele…?“, „Wie lautet die Umkehrung der Aussage…?“	Stellen die Frage „Gibt es Spezialfälle…?“

Um solche Prozessbereiche sinnvoll zu operationalisieren, müssen sie sich auf einen Inhaltsbereich beziehen. Dies würde ein zweidimensionales Modell voraussetzen, bei dem eine Dimension Prozesse beschreibt, die andere Inhalte, z.B. die Kombination argumentieren aus dem Prozessbereich und Daten und Zufall aus dem Inhaltsbereich. Dann könnte dieser Ansatz analog zum BIS-Modell mit den Dimensionen *Inhalte* und *Operationen* betrachtet werden (Jäger, 1982).

Im vorliegenden Fall scheitert dies jedoch, da die Inhalte größtenteils selbst prozesshaft sind. Damit ist gemeint, dass wie in Tabelle 12 dargestellt eine inhaltliche Erwartung für die Kernkompetenz Daten und Zufall beinhaltet, dass auf Daten basierende

Schlussfolgerungen geäußert (=Prozess) und begründet (=Prozess) werden (NK, 2006a, S. 34).

Tabelle 12 Auszug der Kernkompetenzumschreibung für den Inhaltsbereich *Daten und Zufall* (NK, 2006, S. 18), Unterkategorie interpretieren Daten (NK, 2006, S. 34).

Schuljahr	
8	9/10
Äußern auf Daten basierende Schlussfolgerungen und begründen diese	Beurteilen Daten und Grafiken in Medien auf mögliche Fehlschlüsse (Stichprobenrepräsentativität. Klassenbildung, grafische Verzerrung, Verteilungsschiefe)

Eine derartig starke Überlappung der Dimensionen ist nicht für alle Inhalts- und Prozessbereiche gegeben, doch erschwert sie die Verwendung des Lehrplans zur Testkonstruktion erheblich.

Für den Lehrplan der Realschule erübrigt sich eine getrennte Betrachtung insofern, da die Benennung der Inhalts- und Prozesskompetenzen identisch ist (NK, 2006b). Bei der tabellarischen Konkretisierung (wie in Tabelle 11 und Tabelle 12) sind alle Beschreibungen des Hauptschullehrplans (NK, 2006a) wortgleich im Realschullehrplan (NK, 2006b) enthalten, der darüber hinaus noch einige zusätzliche Beschreibungen enthält.

Für den Lehrplan des Gymnasiums werden die inhaltsbezogenen Kompetenzbereiche wortgleich von Haupt- und Realschule übernommen, die prozessbezogenen Kompetenzbereiche werden ebenfalls fast wortgleich übernommen, hier wird argumentieren zu mathematisch argumentieren, Problemlösen zu Probleme mathematisch lösen, usw. (NK, 2006c, S. 12). Auch wird nun explizit darauf hingewiesen, dass es Dopplungen zwischen prozess- und inhaltsbezogenen Kompetenzbereichen gibt, begründet mit besserer Lesbarkeit und besserem Verständnis (NK, 2006c, S. 12). Die inhaltsbezogenen Kompetenzen des Gymnasial-Lehrplans, sind teilweise bereits sehr konkret, wie "Kennen der Identität $\sqrt{a^2} = |a|$ „ (NK, 2006c, S. 25), mindestens genauso häufig jedoch sehr vage, wie z.B. "erkennen und begründen von Ähnlichkeiten" (NK, 2006c, S. 31) für die inhaltsbezogene Kompetenz *Raum und Form* für den Schuljahrgang 10. Leider sind keinerlei komplette Beispielaufgaben im Haupt-, Real- oder Gymnasiallehrplan enthalten.

5.3.3 Lehrpläne Baden-Württembergs

In Baden-Württemberg existieren derzeit vier verschiedene allgemeinbildende Schulen (außer der Grundschule) und zwar neben Haupt-, Realschule und Gymnasium die Werkrealschule. Bei letzterer handelt es sich um Hauptschulen in Baden-Württemberg (BW) die es ermöglichen nach der 10. Klasse einen Realschulabschluss zu erwerben (Oettinger, 2008, S. 6). Dieser Abschluss ist gleichwertig, jedoch nicht gleichartig, mit dem regulären Realschulabschluss (Oettinger, 2008, S. 7). Für alle vier Schultypen werden in Baden-Württemberg die Lehrpläne als Bildungsplan, Leitgedanken und Bildungsstandards bezeichnet (BW, 2004a, 2004b, 2004c, 2004d). Der Hauptschulbildungsplan enthält eine kurze Beschreibung der erwarteten Kompetenzen und den Hinweis, dass die Kompetenzen und Inhalte nach Leitideen strukturiert seien (BW, 2004b). Fünf der Leitideen entsprechen von ihrer Bezeichnung her exakt den Vorgaben zu inhaltsbezogenen Kompetenzen der KMK (2005a), die dort u.a. auch Leitideen genannt werden. Darüber hinaus kommt eine weitere Leitidee hinzu und zwar *modellieren.* Diese ist in den KMK-Bildungsstandards (KMK, 2004a; KMK, 2005a) jedoch keine inhaltsbezogene Kompetenz oder Leitidee, sondern eine allgemeine mathematische Kompetenz (vgl. Abbildung 16).

Das heißt die inhaltsbezogenen mathematischen Kompetenzen wurden mit den allgemeinen mathematischen Kompetenzen der KMK vermischt, allesamt Leitideen genannt und mit Kompetenzen und Inhalten überschrieben (BW, 2004b). Dieses Vorgehen erschwert eine Anwendung im Rahmen der Testkonstruktion erheblich. Es trifft auf alle besprochenen Lehrpläne Baden-Württembergs zu. Das Konzept der Leitidee soll hier nicht an allen Beispielen durchgearbeitet werden, um einen gewissen Überblick zu gewährleisten wurde die Leitidee Daten und Zufall für drei Schultypen in Tabelle 13 zusammengefasst.

Tabelle 13 Leitidee Daten und Zufall für Haupt-, Werkreal- und Realschule in Baden-Württemberg. Nach BW (2004a, 2004b, 2004c).

Hauptschule	Werkrealschule	Realschule
Tabellen und unterschiedliche grafische Darstellungen auswerten	Daten systematisch sammeln und mit geeigneten Hilfsmitteln übersichtlich darstellen	Daten systematisch sammeln und übersichtlich darstellen
Daten recherchieren, mit geeigneten Hilfsmitteln aufbereiten, in Tabellen erfassen und grafisch darstellen sowie die Wirkung der Darstellung beurteilen.	Wahrscheinlichkeitsaussagen verstehen und anwenden	Wahrscheinlichkeitsaussagen verstehen
	Statistiken nach vorgegebenen Kriterien analysieren und bewerten	Daten interpretieren
	Statistiken selbstständig analysieren und bewerten	verschiedene mathematische Darstellungen verwenden Beurteilen Daten und Grafiken auf mögliche Fehlschlüsse (Stichprobenrepräsentativität. Klassenbildung, grafische Verzerrung, Verteilungsschiefe
		Aussagen, die auf Datenanalysen basieren, reflektieren und bewerten
		Daten erfassen, entnehmen transferieren Wahrscheinlichkeiten bestimmen - zweistufige Zufallsversuche
		logisch schließen und begründen.

Anmerkung. Die Beschreibungen wurden nachträglich so angeordnet, dass ein möglichst leichter Vergleich zwischen den Schultypen möglich ist.

Der Lehrplan des Gymnasiums (BW, 2004d) für die 10. (und 8.) Klasse enthält nicht fünf, sondern zehn Leitideen, dargestellt nach dem Schema der drei anderen Schultypen Hier taucht unter anderem die Leitidee *Variable* auf, mit den Unterpunkten einfache

Terme lösen und elementare Gleichungen lösen die im Falle der Hauptschule und Realschule am ehesten bei der (im Gymnasiumsplan ebenfalls vorhanden) Leitidee *Zahl* zu finden ist.

5.3.4 Lehrpläne Bayerns

Die Bayerischen Lehrpläne wurden von dem Institut für Schulqualität und Bildungsforschung (ISB) in München entwickelt. Das Institut hat in einer 100-seitigen Expertise die Konsequenzen der KMK-Bildungsstandards (KMK, 2004a, KMK, 2005a) für die bayerischen Lehrpläne in allen Fächern herausgearbeitet, darunter auch im Fach Mathematik (ISB, 2005). Dazu gehört, dass - im Gegensatz zu Niedersachsen, NRW und Baden-Württemberg - präzise beschrieben wird, wo Passungen und Abweichungen von Lehrplänen und KMK-Standards liegen (ISB, 2005, S. 53). Um zu verdeutlichen wie präzise das ISB sich dieser Aufgabe annimmt, sei auf Tabelle 14 verwiesen, in der das ISB (2005) den bayerischen Lehrplan hinsichtlich Passung mit der Leitidee Raum und Form der KMK-Bildungsstandards (KMK, 2004a; KMK, 2005a) einschätzt.

Tabelle 14 Passung von Hauptschullehrplan und KMK-Bildungsstandards für die Leitidee Raum und Form laut ISB (2005, S. 28)

KMK-Wortlaut	ISB-Einschätzung
fertigen Netze, Schrägbilder und Modelle von ausgewählten Körpern an und erkennen Körper aus ihren entsprechenden Darstellungen	Das Erstellen von Netzen ist im Lehrplan nicht ausdrücklich erwähnt; es sollen jedoch Beziehungen zwischen Netz und Körper untersucht werden. Anstelle von Schrägbildern ist im Lehrplan von Schrägskizzen die Rede.
wenden Sätze der ebenen Geometrie bei Konstruktionen und Berechnungen an	Die Anwendung des Satzes von Pythagoras ist bei Berechungen vorgesehen. Im Zsg. mit Konstruktionen bzw. Zeichnen mit Zirkel und Lineal, wie es im Lehrplan heißt, werden jedoch keine Sätze der ebenen Geometrie thematisiert.
zeichnen und konstruieren geometrische Figuren unter Verwendung angemessener Hilfsmittel wie Zirkel, Lineal, Geodreieck oder dynamische Geometriesoftware	Dynamische Geometriesoftware ist im Lehrplan nicht ausdrücklich erwähnt.

Die Passung für den mittlere Reife Hauptschulzug (am Ende der Hauptschule kann der Realschulabschluss erworben werden), die Hauptschule und das Gymnasium mit den KMK-Standards (KMK, 2004; KMK, 2005) wird laut ISB (2005) bis auf wenige Ausnahmen als gut bezeichnet. Insbesondere für den mittlere Reife Hauptschulzug heißt es, dass sowohl die allgemeinen, als auch die inhaltlichen mathematischen Kompetenzen der KMK "nahezu vollständig verankert" seien (ISB, 2005, S. 59).
Hinzu kommt die konkrete Formulierung im Lehrplan als solchem, so ist der Lehrplan des mittlere Reife Hauptschulzweigs (9. Klasse, M-Klasse) nicht nach abstrakten Leitideen (Baden-Württemberg), Kernkompetenzbeschreibungen (Niedersachsen) oder Kompetenzen (NRW) geordnet, sondern nach Prozentrechnen und Zinsrechnung, Potenzen und Wurzeln, Geometrie, Funktionen und beschreibende Statistik und ebenso konkreten Unterpunkten "rein quadratische Gleichungen Lösen", "rationale Zahlen und Variablen quadrieren", "Fachbegriffe: Hypotenuse, Kathete" (ISB, 2004, S. 590).

5.3.5 Fazit zu den Lehrplänen

Der Kernlehrplan für die Hauptschule in NRW (2004a) ist bereits deutlich konkreter als die eingangs (Abschnitt 5.2) betrachteten Bildungsstandards, könnte für die Zwecke der Testkonstruktion jedoch deutlich präziser ausfallen. Diese Schlussfolgerung gilt auch für den Realschul-Kernlehrplan und den Gymnasial-Kernlehrplan (NRW, 2007) die sehr ähnlich aufgebaut sind.
Zu den Lehrplänen Niedersachsens lässt sich abschließend anmerken, dass sie durch ihre Multidimensionalität, bei der die einzelnen Dimensionen starke Überlappungen aufweisen eine Testkonstruktion schwieriger machen als dies nötig wäre. Der Gymnasiallehrplan ist im Vergleich zu NRW bereits relativ konkret, wobei noch großer Spielraum für Verbesserungen bestünde.
Für die Lehrpläne Baden-Württembergs lässt sich schlussfolgern, dass sie eigentlich vor allem eine teilweise Adaption der Bildungsstandards darstellen, bei der einige der Ursprungselemente vermischt und neu benannt wurden. Aus Sicht der Testkonstruktion weisen sie keine Vorteile gegenüber den bereits dargestellten Bildungsstandards auf. Probleme bereitet ferner die Auffächerung in vier Schultypen mit teils unterschiedlichen Leitideen. Dies macht die Phase der Testkonstruktion unnötig kompliziert, da für diverse Begriffe ihre Entsprechungen in den jeweiligen Lehrplänen gefunden werden müssen, ist z.B. *mit Variablen in Formeln rechnen* (Hauptschule) (BW, 2004b) äquivalent zu *Formeln nach einer Variable auflösen* (Werkrealschule) (BW, 2004a),

oder ist letzteres eine inhaltliche Teilmenge? Hier zeigt sich, dass es generell schwierig ist, die Unterschiede der einzelnen Lehrpläne herauszuarbeiten.

Die Tatsache, dass Bayern genaue Aussagen zur Passung der Lehrpläne macht, anstatt wie andere Bundesländer den bisherigen Lehrplan durch Formulierungen, die ebenso vage wie die Bildungsstandards anmuten, zu ersetzen, spricht für diese Lehrpläne, insbesondere jenen für den mittlere Reife Hauptschulzweig, als Arbeitsgrundlage zur Testkonstruktion. Es scheint praktisch keine Inhaltsbereiche zu geben, die in einem der anderen betrachteten Lehrpläne enthalten sind und im Bayrischen fehlen. Hauptunterschied von letzterem zu den anderen Lehrplänen ist die genaue Information zur Abdeckung hinsichtlich KMK-Standards und die sehr konkreten Inhaltsbeschreibungen. Es zeigte sich, dass im bayerischen Lehrplan eventuell fehlende Teile entweder bei der Testkonstruktion kaum umsetzbar sind (z.B. dritter Vergleich in Tabelle 14), inhaltlich enthalten aber nicht verbindlich mit Praxis verknüpft, (z.B. zweiter Vergleich in Tabelle 14), oder wie beim ersten Vergleich in Tabelle 14 die KMK-Formulierung vermutlich dasselbe meint wie der Lehrplan.

Wie eingangs bei Betrachtung der Bildungsstandards und ihrer Beziehung zu Lehrplänen herausgearbeitet wurde (Abschnitt 5.2), stellen Bildungsstandards alleine keine gute Grundlage für einen Leistungstest dar. Ob Aufgaben, die implizit an den Bildungsstandards orientiert sind (da es der bayerische Lehrplan auch ist), Eingang in die Endform haben hängt in dieser Arbeit ausschließlich von der empirischen Bewährung der Vorform ab.

5.4 Technische Konstruktionsprinzipien

Nachdem in Abschnitt 5.1 Geltungsbereich sowie Zielgruppe festgelegt wurden und in Abschnitt 5.2 bis 5.3 die Bedeutung von Bildungsstandards und Lehrplänen bestimmt wurde, gilt es nun sich den technischen Aspekten der Testkonstruktion zu widmen. Nach Krohne und Hock (2007, S. 35) stellt die Testkonstruktion einen mehrstufigen Prozess dar, bestehend aus der Konstruktdefinition, der Erstellung einer vorläufigen Itemmenge und deren Erprobung, Analyse, Bewertung und Revision. Die Konstruktdefinition wurde in Abschnitt 3.1.6 in Anlehnung an die Intelligenzdiagnostik vorgenommen, die Erstellung der vorläufigen Itemmenge wird in Abschnitt 5.5 beschrieben und die Erprobung, Analyse, Bewertung und Revision folgt darauf in Abschnitt 6.

5.4.1 Item-Benennungen in dieser Arbeit

In dieser Arbeit kann ein und dieselbe Aufgabe unter verschiedenen Benennungen auftauchen. Einmal als Aufgabe im studentischen Test, dann als Aufgabe in der Vorform und zum schließlich als Aufgabe in der Endform. Diese eine Aufgabe heißt aus organisatorischen Gründen in jeder Testform anders. Um das Kriterium der Nachprüfbarkeit zu gewährleisten, ist in Anhang 12.2 eine Tabelle hinterlegt. In dieser Tabelle sind alle Aufgabenbezeichnungen aufgelistet.

5.4.2 Antwortformat

Lienert und Ratz (1994) unterscheiden zwischen Richtig-Falsch-, Ergänzungs-, Mehrfachwahl-, Zuordnungs-, Umordnungs- und Kurzaufsatzaufgaben. Eine neuere Aufteilung von Jankisz und Moosbrugger (2008) unterscheidet sich davon im Wesentlichen durch das Hinzufügen der Beurteilungsaufgaben und die Einordnung aller Aufgaben in drei Typen und zwar Aufgaben mit freiem-, gebundenem- und atypischen-Antwortformat. Für den hier angestrebten Leistungstest kommen letztlich nur Mehrfachwahl- und Ergänzungsaufgaben in Frage, da Ökonomie der Auswertung und Objektivität gleichzeitig optimiert werden sollen.

Ergänzungsaufgaben in Mathematiktests nehmen insofern eine Sonderrolle ein, als dass sie sich ebenso objektiv auswerten lassen wie MC-Aufgaben (Raatz, 1980, S. 28). Sonst vorhandene Nachteile von Ergänzungsaufgaben, wie z.B. Probleme der Auswertungsobjektivität sollten bei Mathematiktests eine untergeordnete Rolle spielen.

5.5 Generierung der Testaufgaben

Fast alle Items aus dem in Abschnitt 4.4 verwendeten Test wurden in die neu zu prüfende Vorform integriert. Zur Generierung der neuen Testaufgaben fand eine Orientierung an den Lehrplänen des Bundeslandes Bayern statt, die im Abschnitt 5.3.4 erläutert wurden. Es wurden insgesamt deutlich mehr Aufgaben generiert als für die Endform vorgesehen, um eine Selektion der besten Items zu ermöglichen.

Basis stellten die Lehrpläne der Haupt- (8. und 9. Klasse) und Realschule (8-10. Klasse), die jedoch sehr große Überlappungen aufweisen, dar. Viele der Aufgaben ließen sich auch dem Gymnasiallehrplan zuordnen. Eine große Hilfe bei der Generierung der Testaufgaben stellten zudem Lehrbücher des Westermann Verlags für

die 9. Klasse der Hauptschule (Golenia & Neubert, 2007) und die 10. Klasse der Realschule (Dlugosch, Englmaier, Götz & Widl, 2006) dar. Diese Lehrbücher halten sich strikt an die Lehrpläne des Bundeslandes Bayern.

Das wichtige Kriterium der Inhaltsvalidität der Aufgaben wird nach Cronbach und Meehl wie folgt sichergestellt (1955, S. 281): „Content validity is established by showing that the test items are a sample of a universe in which the investigator is interested". Durch die Orientierung an den genannten Lehrplänen wird demnach hier die Inhaltsvalidität sichergestellt (Lienert & Raatz, 1994). Ein weiteres Kriterium für die Aufgabengenerierung war, für alle vier Skalenkonzeptionen aus Abschnitt 3.1.6 so viele Aufgaben zu erhalten, dass eine Selektion schlechterer Items möglich wird und dennoch genügend Items unterschiedlicher Schwierigkeit übrig bleiben. Es wurde stets versucht, gleichzeitig lehrplanvalide Aufgaben zu erstellen, die zusätzlich noch einer der vier aufgestellten Skalenkonzeptionen zugeordnet werden können. Dies ist natürlich nur tendenziell möglich, so enthalten Aufgaben der Skala mathematische Literalität nicht nur Text, sondern auch Berechnungen. Je stärker der Anteil an Berechnungen im Verhältnis zum Textanteil wird, desto eher ist die Aufgabe wiederum der Skala prozedurales Rechnen zuzuordnen – ein Problem, das natürlich für alle Skalen existiert und sich aus der Annahme eines Modells korrelierter Faktoren ergibt.

5.6 Zusammenstellung zweier Testvorformen

Einige Studenten wurden gebeten, den Test durchzuführen und auf missverständliche Itemformulierungen und mögliche Fehler in den Aufgaben zu achten, wie es z.B. auch von Lienert und Raatz (1994, S. 53) empfohlen wird. Die Aufgaben und eine Testanweisung einschließlich Aufgabenbeispielen sowie eine kurze Formelsammlung wurden von zwei Kollegen der Uni Mannheim gegen geprüft. Dieses Vorgehen resultierte letztlich in einem Itempool, der im kommenden Abschnitt die Basis für die Zusammenstellung der Endform darstellt. Insgesamt wurden aus Gründen der Zeitbegrenzung im Schulsetting zwei Testvorformen mit verschiedenen Aufgaben zusammengestellt, die von unterschiedlichen Probanden ausgefüllt wurden. Die zwei Testformen lauten: Form A und Form B. Beide liegen wiederum in unterschiedlicher Reihung der Items vor. Die Unterscheidung ist nur in Abschnitt 6.1 von Bedeutung, dort wird von Form A1 und Form A2 respektive Form B1 und Form B2 gesprochen. Form A enthält Items aus allen 4 Skalen, Form B hauptsächlich Aufgaben der Skalen Geometrie und grafische Funktionen sowie mathematische Literalität. Da die wenigen Aufgaben

der Form B, die nicht zu diesen beiden Skalen gehörten, nicht ihren Weg in die Endform fanden (B35, B26b-d), weil bereits genügend Aufgaben aus Form A für die Skalen vorlagen, werden sie im folgenden nicht mehr aufgeführt

6 Zusammenstellung der Endform

Es ist an dieser Stelle – mit Blick auf das Ziel dieser Arbeit (vgl. Einleitung) – nicht zielführend, jeden Selektionsschritt detailliert darzustellen. Schließlich umfassen beide Vorformen zusammen ganze 160 Items, von denen nur 75 in die Endform eingehen. Es erscheint zweckdienlicher, in erster Linie für jede der Skalen die wichtigsten vorläufigen Kennwerte der schließlich ausgewählten Aufgaben zu präsentieren und stets mindestens ein prototypisches Item zu präsentieren. Im Anhang 12.6 und 12.7 finden sich für alle Items der Vorformen die Kennwerte vor Zusammenstellung der Endform. Alle Items der Vorformen A und B, die letztlich nicht für die Endform verwendet wurden, sind unter der Adresse: http://www.psychologie.uni-mannheim.de/projekte/lims/restliche_aufgaben.zip zu wissenschaftlichen Zwecken als editierbares Word-Dokument verfügbar (Passwort: STARTm).

6.1 Stichprobe

Leider war es aus organisatorischen Gründen nicht möglich, die unterschiedlichen Reihungen in jeder Klassenstufe einzusetzen. Tabelle 15 zeigt die Verteilung der Personen des Vortests auf die unterschiedlichen Testvorformen und Klassen.

Tabelle 15 Testformen, Klassenstufen und Anzahl von Personen.

Klasse	Testform	Anzahl Personen
9	Form A1	28
9	Form B2	28
11+	Form A2	45
11+	Form B1	48

Die Stichprobe für den Vortest umfasste demnach insgesamt 149 Personen, die sich gleichmäßig auf die Klassenstufen und Testformen aufteilten. Die Erhebungen fanden in der Unterrichtszeit statt und waren auf eine Zeitstunde begrenzt. Erhebungsorte waren der Raum Berlin und München.

6.2 Zusammenstellung der Skalen der Endform

Für die endgültige Zusammenstellung der Skalen der Endform in diesem Abschnitt wurden zahlreiche Selektionen vorgenommen. Besonderer Wert wurde hierbei auf die folgenden fünf Aspekte gelegt.

1. Alle Aufgaben weisen ausreichende Trennschärfen auf
2. Das ganze Schwierigkeitskontinuum wird abgedeckt
3. Es werden Items gewählt die möglichst prototypisch (vgl. Abschnitt 3.1.6) für die Skalen sind
4. Die Testlänge von ca. 1 Stunde wird nicht überschritten
5. Alle Skalen enthalten genügend Items

6.2.1 Auswahl von Items für Geometrie und grafische Funktionen

Tabelle 16 zeigt die verbliebenen Items der Skala Geometrie und grafische Funktionen. Nach Möglichkeit wurde über den ganzen Schwierigkeitsbereich an jeder Stelle jenes Item ausgewählt, welches über die beste Trennschärfe verfügte. Es verbleiben 20 Items für die Skala.

Tabelle 16 Verbliebene Items Geometrie und grafische Funktionen

Item	P	SD	r_{it}	Cronbach's α nach Ausschluss
B27b	0,25	0,43	0,67	0,89
B27a	0,19	0,40	0,63	0,89
B28b	0,27	0,45	0,62	0,89
B28c	0,27	0,45	0,62	0,89
B27d	0,12	0,33	0,62	0,89
B29c	0,52	0,50	0,61	0,89
B29a	0,56	0,50	0,60	0,89
B21	0,38	0,49	0,59	0,89
B28a	0,30	0,46	0,58	0,89
B29b	0,56	0,50	0,58	0,89
B27c	0,15	0,36	0,58	0,89
B25b	0,12	0,33	0,56	0,89
B19	0,37	0,49	0,55	0,89
B25a	0,14	0,35	0,53	0,89
B15	0,48	0,50	0,45	0,90
B17	0,77	0,43	0,41	0,90
B22	0,81	0,40	0,40	0,90
B32	0,82	0,39	0,33	0,90
B13	0,77	0,43	0,31	0,90
A19	0,70	0,46	-	-

Anmerkung. Innerhalb der Testform sortiert nach r_{it}. $N = 73$.

Die korrigierte Trennschärfe variiert zwischen $r_{it} = 0,67$ und $r_{it} = 0,31$, was zusammen mit einer internen Konsistenz von $\alpha = 0,90$ für gute Skaleneigenschaften spricht (Lienert & Raatz, 1994). Zusätzlich wurde aus der kleinen Itemmenge der FORM A die den Bereich Geometrie und grafische Funktionen erfasst, Aufgabe A19 übernommen. Sie wurde ausgewählt, da gemäß Tabelle 16 mehr Aufgaben mit $p < 0,50$ als mit $p > 0,50$ enthalten sind und eine ausgewogene Schwierigkeit angestrebt wird. Wie in der Skalenkonzeption festgelegt (vgl. Abschnitt 3.1.6) sollen Items der Skala Geometrie und grafische Funktionen hauptsächlich reine Geometrie und die Verarbeitung einer grafischen Darstellung von Funktionen prüfen. Insofern stellt das Item 21a gemäß folgender Abbildung 19 ein prototypisches Item dar. Es wies unter den 9 Items der Form A für die Skala Geometrie akzeptable Eigenschaften mit $p = 0,60$ und $r_{it} = 0,30$ auf.

21. Betrachte zunächst die folgenden Kurven.

Die Kurven b, c und d entstehen alle durch einfache Umformung von Kurve a.
Kurve a ist eine Abbildung der
Funktion $y = 2^x$

Wähle die passende Umformung aus:
a.) Kurve b ist eine Abbildung der Funktion

- ☐ $y = 2$
- ☐ $y = 2^x + 29$
- ☒ $y = 2^x + 3$
- ☐ $y = 2^x - 3$
- ☐ $y = 2^{x+1}$

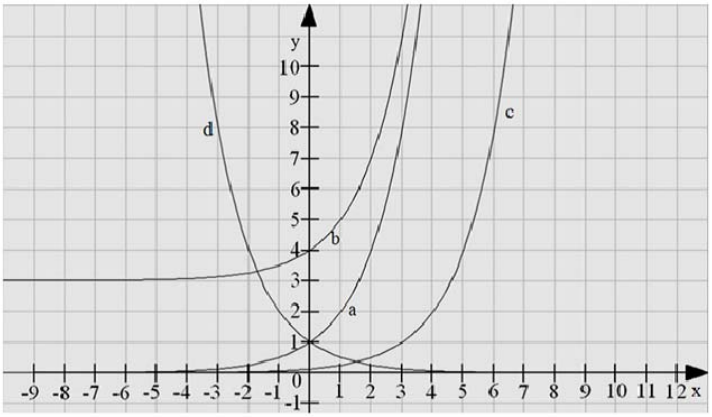


Abbildung 19 Prototypisches Item der Skala Geometrie und grafische Funktionen

Letztlich fand es keinen Weg in die Endform, da bereits genügend ähnliche Aufgaben mit gleich guten oder besseren Eigenschaften darin enthalten sind.

6.2.2 Auswahl von Items für prozedurales Rechnen

Die für die Skala prozedurales Rechnen gewählten Items variieren gemäß Tabelle 17 in ihrer korrigierten Trennschärfe zwischen $r_{it} = 0{,}10$ und $r_{it} = 0{,}68$.

Tabelle 17 Verbliebene Items prozedurales Rechnen.

Item	p	SD	r_{it}	Cronbach's α nach Ausschluss
A12b	0,22	0,42	0,68	0,91
A8c	0,51	0,50	0,68	0,91
A6d	0,53	0,50	0,67	0,91
A6a	0,64	0,48	0,64	0,91
A4g	0,45	0,50	0,64	0,91
A12c	0,33	0,47	0,62	0,91
A4f	0,47	0,50	0,62	0,91
A12a	0,40	0,49	0,59	0,91
A12d	0,12	0,33	0,58	0,91
A6b	0,70	0,46	0,57	0,91
A6c	0,73	0,45	0,55	0,91
A13c	0,19	0,40	0,55	0,91
a10b	0,59	0,50	0,55	0,91
A10c	0,42	0,50	0,52	0,91
a5c	0,30	0,46	0,50	0,91
A20	0,63	0,49	0,49	0,91
a5b	0,62	0,49	0,49	0,91
A13b	0,47	0,50	0,46	0,91
A13d	0,45	0,50	0,46	0,91
A2e	0,85	0,36	0,42	0,91
a5d	0,79	0,41	0,38	0,91
A22	0,55	0,50	0,38	0,91
A2d	0,86	0,35	0,37	0,91
a3a	0,93	0,25	0,35	0,91
A8a	0,95	0,23	0,33	0,91
A4a	0,95	0,23	0,32	0,91
a3e	0,52	0,50	0,31	0,91
a3b	0,68	0,47	0,29	0,91
A2a	0,95	0,23	0,10	0,91

Anmerkung. Sortiert nach r_{it}.

Auch die interne Konsistenz erreicht mit α = 0,91 einen guten Wert. Eine prototypische Aufgabengruppe, die für die Endform entfernt wurde, da zu viele Aufgaben vorlagen, zeigt die folgende Abbildung 20.

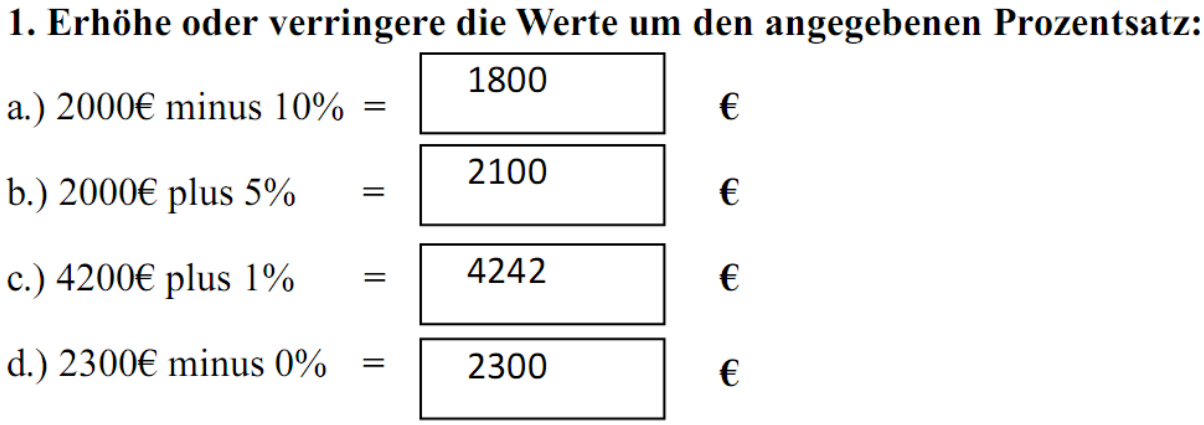


Abbildung 20 Prototypische Aufgabe der Skala prozedurales Rechnen A1a-d.

6.2.3 Auswahl von Items für komplexes Rechnen

Für die Endform wurden aus der Skala komplexes Rechnen der Form A, hauptsächlich aus Zeitgründen, nur die Aufgaben 9d und 9a entfernt. Bei diesen Items handelte es sich um ein sehr leichtes und ein eher schweres Item derselben Aufgabengruppe. Die nach der Selektion übrigen Items einschließlich Kennwerten sind der folgenden Tabelle zu entnehmen.

Tabelle 18 Verbliebene Items der Skala komplexes Rechnen, getrennt für Form A.

Item	p	SD	r_{it}	Cronbach's α nach Ausschluss
A9c	0,90	0,29	0,28	0,80
a11a	0,70	0,46	0,33	0,80
A18	0,63	0,48	0,42	0,79
A17b	0,60	0,49	0,42	0,79
A17a	0,52	0,50	0,59	0,77
A16c	0,52	0,50	0,50	0,78
a11b	0,48	0,50	0,58	0,77
A9f	0,32	0,47	0,51	0,78
A9g	0,25	0,43	0,63	0,77
A9e	0,22	0,41	0,39	0,79
A17c	0,12	0,33	0,41	0,79

Anmerkung. Sortiert nach r_{it}

Die beiden nicht ausgewählten Items A16a und A16b sind als prototypisch für diese Skala anzusehen, sie erfordern aus einer Auswahl von Lösungen die richtige Umformung einer Gleichung zu finden. Sie sind der beibehaltenen Aufgabe a16a sehr

ähnlich und im folgenden abgebildet (Abbildung 21). Weshalb sie deutlich niedrigere Trennschärfen aufwiesen bleibt unklar ($p_{a16a} = 0{,}86$, $p_{a16b} = 0{,}77$).

Die folgenden Gleichungen sollen nach x aufgelöst werden.
Wähle die richtige Lösung:

a.) Gleichung: $x^2+5=105$

- ☐ $x = \sqrt{10}$
- ☐ $x = \sqrt{110}$
- ☐ $x = 11$
- ☒ $x = 10$
- ☐ $x = \sqrt{-100}$

b.) Gleichung: $3x^2-27=0$

- ☒ $x = 3$
- ☐ $x = 9$
- ☐ $x = -\sqrt{\frac{27}{3}}$
- ☐ $x = 27 - 3x$
- ☐ $x = 30$

Abbildung 21 Prototypische Items der Skala komplexes Rechnen, A16a, A16b.

6.2.4 Auswahl von Items für mathematische Literalität

Tabelle 19 zeigt die ausgewählten Items der Skala mathematische Literalität. Aus dem kleinen Itempool der Form A zu dieser Skala wurden A27a und A27b ausgewählt. Vor allem, da sie gute Trennschärfen in der aus 5 Items bestehenden Mini-Skala der Form A aufwiesen ($r_{it} = 0{,}67$ und $r_{it} = 0{,}61$) und das Schwierigkeitskontinuum optimal ergänzen.

Tabelle 19 Items der Skala mathematische Literalität.

Item	p	SD	r_{it}	Cronbach's α nach Ausschluss
B10c	0,70	0,46	0,70	0,85
B5c	0,66	0,48	0,69	0,85
B5b	0,71	0,46	0,69	0,85
B5d	0,60	0,49	0,65	0,85
B4c	0,66	0,48	0,62	0,85
B4d	0,52	0,50	0,56	0,86
B4a	0,74	0,44	0,52	0,86
B7b	0,32	0,47	0,50	0,86
B2b	0,53	0,50	0,49	0,86
B2a	0,79	0,41	0,49	0,86
B10a	0,95	0,23	0,39	0,87
B10b	0,92	0,28	0,38	0,87
B12	0,14	0,35	0,24	0,87
A27a	0,29	0,45		
A27b	0,21	0,40		

Anmerkung. Innerhalb der Testform sortiert nach r_{it}

Die korrigierten Trennschärfen variieren von $r_{it} = 0{,}24$ bis $r_{it} = 0{,}70$, was zusammen mit der internen Konsistenz von $\alpha = 0{,}87$ als guter Wert anzusehen ist (Lienert & Raatz, 1994). Eine prototypische Aufgabe, die letztlich nicht in den Test aufgenommen wurde, ist Aufgabe B3b, mit folgendem Wortlaut:

Ein Seniorenwohnheim macht einen Ausflug und teilt die Bewohner in drei Gruppen ein. Die erste Gruppe ist 10% größer als die zweite. Die zweite Gruppe ist 50% kleiner als die dritte. In der dritten Gruppe sind 40 Personen. Wie viele Personen sind in der ersten Gruppe?(Lösung: 22 Personen)

Hierbei handelt es sich um eine typische Textaufgabe, bei der ein mathematisches Problem in Textform vorliegt und quasi in die Sprache der Mathematik übersetzt werden muss. Von diesem Typ befinden sich bereits mehrere sehr ähnliche Items in der Skala der Endform.

6.3 Weitere Veränderungen bis zur Endform

Bei abschließender Durchsicht der Aufgaben fiel auf, dass die Endform ein großes inhaltliches Feld abdeckt, jedoch eine inhaltlich nicht eingekleidete Divisionsaufgabe fehlt. Daher wurden kurzerhand zwei weitere Aufgaben (Endform: 14a, 14b) erstellt. Somit umfasst die vorläufige Endform mit ihren 4 Skalen 77 Items die im folgenden Abschnitt überprüft werden.

7 Passung der Endform gemäß Klassischer Testtheorie

In diesem Abschnitt soll geprüft werden, ob die Endform den notwendigen Bedingungen an Reliabilität und Validität gerecht wird, um im darauf folgenden Abschnitt 8 das aufgestellte Modell von vier korrelierten Faktoren und in Abschnitt 9 die psychometrische Bedeutung der taxonomischen Ordnung zu prüfen. Ein besonders wichtiger Aspekt stellt in diesem Zusammenhang die Konstruktvalidität (Cronbach & Meehl, 1955) dar, die in Abschnitt 7.4 geprüft wird.

7.1 Wieso klassische Testtheorie?

Bereits an mehreren Stellen dieser Arbeit wurde – insbesondere bei Methoden zur Prüfung der N-Dimensionalität eines Tests – auf Nachteile der klassischen Testtheorie (KTT) (Gulliksen, 1950; Lord & Novick, 1968) hingewiesen. Nun stellt sich die Frage, wieso die Endform des Tests zunächst anhand der KTT einer Kontrolle unterworfen

wird. Ein nicht erschöpfender Grund liegt darin, dass der Test auf Basis der KTT entwickelt wurde. Überhaupt wäre dies unter den vorherrschenden Bedingungen überhaupt nicht anders möglich gewesen, da die Stichprobengröße der Vorform (vgl. Abschnitt 6.1) recht klein war. Entscheidend ist jedoch, was bereits viele Autoren, unter ihnen Rost (1999), Kubinger (2000) und Moosbrugger (2008, S. 216) betonen, nämlich dass die probabilistische Testtheorie (PTT) keine Alternative, sondern ein komplementäres Modell zur KTT darstellt. Bei Verwendung der einfachsten Rasch-Modelle stellte Rost (1999, S. 141) fest, dass „aus Sicht des Testpraktikers bei einer Testanalyse mit dem Rasch-Modell in der Regel nichts anderes herauskommt als bei einer Analyse nach der klassischen Testtheorie“. Dazu passt auch die Replik von Klaus Kubinger (2000), in der es an einer Stelle heißt, dass die komplementäre Rolle der PTT zur KTT allein schon darin begründet sei, dass auch ein Test der Modellgeltung gemäß PTT aufweist mindestens die klassischen Gütekriterien Validität und Normierung erfüllen muss, um den Qualitätsansprüchen klassischen Diagnostizierens zu genügen. Die Normierung eines Tests ist ein wichtiger Aspekt, doch vor allem für angewandte, Diagnostik entscheidend (Lienert & Raatz, 1994, S. 12), weshalb dafür auf das Testmanual verwiesen wird (Jasper & Wagener, in Druck). Die Validität des Tests wird an einigen Beispielen in diesem Abschnitt untersucht, um überhaupt die folgenden Analysen zur N-Dimensionalität und Konstruktstruktur auf Itemebene anhand geeigneter (probabilistischer) Verfahren in Abschnitt 8.2-8.3 zu rechtfertigen. Der wirkliche Nutzen des Rasch-Modell erschließt sich erst durch seine zahlreichen Verallgemeinerungen, so schreibt Rost (1999, S. 149): „…daß sich der praktische Nutzen des Rasch-Modells bei Testanalysen nicht wesentlich von dem der klassischen Testtheorie unterscheidet. Dies ändert sich grundlegend, wenn man zu den Verallgemeinerungen des Rasch-Modells übergeht“. Demnach werden in dieser Arbeit Modelle auf Basis der PTT nur eingesetzt, wenn sie einen tatsächlichen Mehrwert bringen, was in diesem Abschnitt nicht der Fall wäre.

7.2 Testanalyse

Im folgenden wird zunächst die umfangreiche Stichprobe kurz beschrieben, um sich anschließend den Aspekten der Testreliabilität und (Konstrukt)validität zu widmen. Sämtliche Daten sind zu wissenschaftlichen Zwecken beim Autor erhältlich.

7.2.1 Stichprobe

Die Normstichprobe wurde zwischen Oktober 2008 und Februar 2009 erhoben. Der Großteil der Probanden stammt aus Schulen in der Gegend um Berlin. Ein kleiner Teil der Stichprobe besteht aus Studenten der Uni Mannheim (N = 59). Die Gesamtzahl der Teilnehmer betrug N = 1554. Sämtliche Probandenantworten erfolgten aus ökonomischen Gründen anhand eines Antwortbogens (Lienert & Raatz, 1994). Die Antworten wurden in ein Statistikprogramm übertragen und anhand eines Skriptes in die Kodierung richtig (1) und falsch (0) überführt. Zuvor wurde geprüft, ob mögliche Fehleingaben, z.B. ein Zahlenwert ungleich 1 bis 5 für eine MC-Aufgabe, oder das Auslassen einer Frage bei der Übertragung von Antwortbogen in die SPSS-Datei vorlagen. Die Analyse ergab, dass ca. 1% der Daten überprüft werden mussten, was auch geschah. Alle Analysen beziehen sich auf diese korrigierten Daten. Aufgrund der starken Aufgliederung des Schulsystems ist die Aufteilung der Stichprobe in bisher erreichten Abschluss von hoher Bedeutung und der folgenden Tabelle 20 zu entnehmen.

Tabelle 20 Maximal erreichter Schulabschluss der Probanden der Normstichprobe.

Schulabschluss	N	Prozent	cum Prozent
kein Abschluss/Hauptschulabschluss (9Jahre Schule)	39	3	3
erweiterter Hauptschulabschluss	181	12	14
Mittlerer Schulabschluss	830	53	68
Fachabitur	91	6	73
Allgemeine Hochschulreife	334	21	95
Keine Angabe	79	5	100
Total	1554	100	

Der jüngste Teilnehmer der Stichprobe war 14 Jahre alt, der älteste 41 bei einem Modus von 18 Jahren. Letztlich waren etwa 50% der Stichprobe jünger als 20 Jahre.

7.2.2 Reliabilitätsschätzungen

Der Mathematiktest besteht aus den vier Subskalen Geometrie und grafische Funktionen, prozedurales Rechnen, Mathematische Literalität und komplexes Rechnen. Insgesamt lassen sich auf Basis der internen Konsistenz befriedigende bis sehr gute Reliabilitätsschätzungen für die Skalen aufzeigen (vgl. Tabelle 21). Werte zwischen 0,82 und 0,95 für Cronbach‘s α sind angemessen und als gut zu bewerten (Lance, Butts & Michels, 2006; Nunally & Bernstein, 1994). Verglichen mit der kleinen Stichprobe der Vorform, fällt Conbach’s α in zwei Fällen etwa besser (prozedurales-, komplexes Rechnen) und in zwei Fällen schlechter (Geometrie und grafische Fkt., mathematische

Literalität) aus. Die Unterschiede in der internen Konsistenz zwischen Frauen und Männern sind derart gering (vgl. Tabelle 21), dass sie vernachlässigt werden können.

Tabelle 21 Reliabilitätsschätzungen für Skalen und Gesamtwert in der Gesamt-Stichprobe und getrennt nach Geschlecht.

Skala	Anzahl Items	Cronbach's α Alpha gesamt	Cronbach's α	
			männlich	weiblich
Geometrie und grafische Funktionen	20	0,87	0,86	0,87
Prozedurales Rechnen	31	0,88	0,88	0,87
Mathematische Literalität	15	0,82	0,81	0,82
Komplexes Rechnen	11	0,84	0,84	0,86
Gesamtskala	77	0,95	0,95	0,95

Anmerkung: $N_{gesamt} = 1554$; $n_{männlich} = 1048$, $n_{weiblich} = 482$, Rest ohne Geschlechtsangabe.

Die wichtigsten Kennzahlen für Analysen auf Itemebene im Rahmen der KTT stellen Schwierigkeit (p) und Trennschärfe dar (Nunally & Bernstein, 1994). Die folgende Tabelle 22 listet beides für alle Items der Endform auf.

Tabelle 22 Itemanalyse Gesamtstichprobe (*N* = 1554).

Geometrie und grafische Fkt.			Prozedurales Rechnen			Mathematische Literalität			Komplexes Rechnen		
Item	p	r_{it}	Item	p	r_{it}	Item	p	r_{it}	Item	p	r_{it}
1	0,86	0,21	13a	0,95	0,23	25b	0,86	0,55	31a	0,82	0,38
2	0,89	0,23	13b	0,71	0,20	25c	0,82	0,54	31b	0,22	0,44
3	0,81	0,34	13c	0,52	0,32	25d	0,80	0,55	31c	0,39	0,56
4	0,88	0,22	14a	0,95	0,21	26a	0,96	0,34	31d	0,26	0,55
5	0,66	0,32	14b	0,89	0,21	26b	0,84	0,40	32a	0,55	0,59
6a	0,59	0,50	14c	0,92	0,22	26c	0,68	0,52	32b	0,39	0,63
6b	0,60	0,52	15a	0,79	0,49	27a	0,81	0,31	33	0,58	0,48
6c	0,58	0,51	15b	0,81	0,51	27b	0,80	0,42	34a	0,40	0,60
7	0,36	0,36	15c	0,76	0,52	27c	0,77	0,47	34b	0,51	0,61
8	0,45	0,43	15d	0,67	0,54	27d	0,65	0,42	34c	0,16	0,40
9	0,35	0,42	16a	0,50	0,39	27e	0,57	0,45	35	0,45	0,50
10a	0,35	0,58	16b	0,38	0,47	28	0,43	0,43			
10b	0,32	0,61	16c	0,60	0,38	29	0,17	0,24			
10c	0,31	0,60	17a	0,55	0,24	30a	0,32	0,46			
11a	0,23	0,60	17b	0,28	0,15	30b	0,27	0,50			
11b	0,27	0,60	18	0,59	0,41						
11c	0,23	0,63	19	0,45	0,32						
11d	0,16	0,58	20a	0,49	0,55						
12a	0,09	0,49	20b	0,22	0,10						
12b	0,08	0,48	21a	0,64	0,55						
			21b	0,52	0,61						
			22a	0,66	0,40						
			22b	0,45	0,48						
			22c	0,55	0,61						
			23a	0,45	0,50						
			23b	0,15	0,41						
			23c	0,42	0,51						
			24a	0,42	0,51						
			24b	0,17	0,52						
			24c	0,33	0,54						
			24d	0,13	0,51						

Es ist ersichtlich, dass die ersten drei Skalen gemäß Tabelle 22 Aufgaben enthalten, die den ganzen Schwierigkeitsbereich abdecken, was die Voraussetzung für eine hohe Messgenauigkeit bei Personen unterschiedlicher Fähigkeit darstellt (Lienert & Raatz, 1994). Die Skala komplexes Rechnen schließlich weist - bis auf ein erstes Eisbrecher-Item (Lienert & Raatz, 1994) - tendenziell schwerere Items auf.

7.2.3 Abschließende Itemselektionen

Aufgrund der klassischen Analysen gemäß vorherigem Abschnitt war es nicht nötig Items zu entfernen. Es finden sich in der Endform lediglich zwei von 77 Einzelitems mit einer Trennschärfe $r_{it} < 0{,}20$. Da es sich bei diesen Aufgaben (20b und 17b) um schwere Items handelt, wurden sie beibehalten. Bis auf eine Multiple Choice Aufgabe (A18) wurden alle Aufgaben identisch aus den Vorformen A und B übernommen. Bei Aufgabe A18 geht es darum, wann eine Gerade g Tangente zu einer Parabel p ist. Die richtige Antwort, wenn es genau einen Schnittpunkt gibt, ist jedoch mathematisch nicht ganz korrekt. Dies trifft nämlich nicht zu, wenn die Gerade parallel zur y- oder x-Achse liegt. Zwar ist auch ohne diesen Zusatz diese Antwortalternative die einzig sinnvolle, doch schien es geboten, diese Einschränkung in die Aufgabe aufzunehmen.

7.3 Hypothesen II

Die folgenden Hypothesen ergeben sich augenscheinlich aus der Skalenkonzeption und den gewählten Verfahren. In den folgenden Abschnitten, die sich der Prüfung der Hypothesen widmen, finden sich zusätzliche Erläuterungen zur Plausibilität der Hypothesen.

Hypothese H4: Alle Skalen des Mathetests korrelieren untereinander. Der höchste Zusammenhang findet sich zwischen den Skalen komplexes Rechnen und prozedurales Rechnen.

Hypothese H5: Keine der Skalen weist einen signifikanten Zusammenhang mit Trait- oder State-Angst auf.

Hypothese H6: Die Skala mathematische Literalität weist den höchsten Zusammenhang mit der verbalen Intelligenz auf.

Hypothese H7: Alle Skalen korrelieren deutlich mit den IST 2000R-Subtests Rechenaufgaben und Zahlenreihen.

Hypothese H8: Es gibt einen deutlichen Zusammenhang zwischen Mathenote und allen Mathetest-Skalen.

Hypothese H9: Der Zusammenhang der Deutschnote mit den 4 Skalen ist stets geringer als jener der Mathenote.

7.4 Konstruktvalidität des Mathematiktests

Zur Konstruktvalidität heißt es bei Cronbach und Meehl (1955, S. 282) „Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator." Trochim und Donelly (2006) gehen so weit, Konstruktvalidität als übergreifende Idee anzusehen, die prädiktive-, konkurrente-, konvergente- und diskriminante Validität mit beinhaltet. Da rein rechnerisch die Reliabilität eine Bedingung für Validität darstellt (Horst, 1971) war es in jedem Fall nötig, im vorherigen Abschnitt zunächst die Frage der Messgenauigkeit zu beantworten. Konstruktvalidität soll hier vor allem als Einordnung in ein nomologisches Netzwerk betrachtet werden (Cronbach & Meehl, 1955).

Es konnte aus ökonomischen Gründen bei weitem nicht allen Versuchspersonen jeder zur Konstruktvalidierung verwendete Test vorgelegt werden, was die Anwendung von Strukturgleichungsmodellen (Kline, 2005) in diesem Kontext unmöglich macht. Sie finden jedoch in Abschnitt 8 und 9 ausgiebige Verwendung um die Skalenzusammenhänge (Abschnitt 7.4.1) und Struktur des Mathematiktests präziser zu prüfen. Bei den verwendeten Tests handelt es sich um einen Verbalen-Kurz-Intelligenztest (Anger, Mertesdorf, Wegner & Wülfing, 1980), ein Inventar zur Messung von Stait-Trait Angst (Laux, Glanzmann, Schaffner & Spielberger, 1981), Teile des IST-2000R (Liepmann et al., 2007) und natürlich Schulnoten in den Fächern Mathematik und Deutsch.

7.4.1 Zusammenhänge zwischen den Skalen der Endform

Die folgende Tabelle 23 gibt die Interkorrelationen zwischen den vier Testskalen wieder. Der höchste Zusammenhang auf Skalenebene findet sich zwischen prozeduralem- und komplexem Rechnen (r = 0,74), der niedrigste Zusammenhang zwischen *Geometrie und grafischen Fkt*. und mathematischer Literalität (r = 0,54). Diese Ergebnisse entsprechen der Erwartung eines Konstrukts der Mathematikfähigkeit, welches aus deutlich miteinander korrelierten Einzelkomponenten besteht.

Tabelle 23 Interkorrelationen der Mathetest-Skalen in der Gesamtstichprobe (*N* = 1554).

Skala	Geometrie und grafische Funktionen	Prozedurales Rechnen	Mathematische Literalität	Komplexes Rechnen
Prozedurales Rechnen	0,70			
Mathematische Literalität	0,54	0,60		
Komplexes Rechnen	0,68	0,74	0,59	
Gesamt	0,86	0,92	0,76	0,86

Anmerkung. Alle Korrelationen sind hochsignifikant ($p < 0,01$).

Wie erwartet, korrelieren alle Skalen hochsignifikant und der höchste Zusammenhang auf Skalenebene existiert zwischen prozeduralem Rechnen und komplexem Rechnen. Letztlich kann Hypothese 4 somit bestätigt werden.

7.4.2 Zusammenhänge mit Trait-State-Angst

Eine der ersten empirischen Untersuchungen, die sich der Testangst widmeten, entstand zu Beginn des 20. Jahrhunderts an Medizinstudenten (Folin, Demis & Smillie, 1914).
Seitdem wurden zahlreiche Theorien zur Entstehung von Angst, der Unterscheidung von State- und Trait Komponente der Angst und dem Zusammenhang von Angst und Leistung aufgestellt (Gärtner-Harnach, 1972; S. 55; Zeidner, 1998). Hembree (1988) fand in einer Meta-Analyse einen Zusammenhang von Angstkorrelaten und Leistung in Mathematik von r = -0,22 bei der Untersuchung von Stichproben die Schüler der Klassen 4 bis 12 umfassten. In einer weiteren Meta-Analyse fand Seipp (1990) einen Zusammenhang von Angst (verschiedene Maße) und Leistung im Fach Mathematik (alle Klassenstufen) von ebenfalls exakt r = -0,22.
Als Maß zur Erfassung der Angst wird in dieser Arbeit das Stait-Trait-Angst Inventar (STAI) von Laux et al. (1981) eingesetzt. Dieses Instrument basiert auf dem State-Trait Anxiety Inventory von Spielberger, Gorsuch und Lushene (1971). Das Instrument unterscheidet zwischen Zustandsangst und Angst als Eigenschaft, wobei Leistungsunterschiede vor allem bei hoher versus niedriger Zustandsangst erwartet werden (Laux et al., 1981). Da Seipp (1990) in ihrer Meta-Analyse bei einer Trennung nach State und Trait-Maß der Angst für beide Konzepte ähnliche Zusammenhänge zur Leistung (über alle Verfahren) fand (r_{STATE} = 0,19; r_{TRAIT} = 0,17) sind an dieser Stelle beide Maße von Interesse. Von den mittlerweile veralteten Normwerten abgesehen

wurde der STAI mehrfach als adäquates Instrument zur Erfassung von Zustandsangst und Ängstlichkeit bewertet (Debener, 2003, S. 163; Muthny, 1997).

Der STAI wurde 79 Probanden (Bachelor-Psychologie-Studenten) vor der Durchführung des Mathematiktests vorgelegt. Die aus 17 Männern und 62 Frauen bestehende Stichprobe erwies sich als hochängstlich bezüglich Statekomponente mit $\bar{x}_{\text{STATE, F}} = 50{,}9$ (*SD* = 4,27; Männer: $\bar{x} = 49{,}4$, *SD* = 5,83) und hochängstlich bezüglich Traitkomponente mit $\bar{x}_{\text{TRAIT, F}} = 50{,}5$ (*SD* = 3,24; Männer: $\bar{x} = 49{,}8$, *SD* = 4,72). Die studentische Teilstichprobe im STAI-Manual weist für beide Maße unter Normalbedingungen etwa 10 Punktwerte niedrigere Mittelwerte und gleichzeitig eine etwa doppelt so hohe Streuung auf (Laux et al., 1981, S. 27). Die demnach hohe Ängstlichkeit bei gleichzeitig niedriger Streuung scheint ein Artefakt der Stichprobe von Psychologiestudenten darzustellen. Dies muss der Fall sein, da der höchste von Laux et al. (1981, S. 27) für Studenten berichtete Wert die State-Angst unter Belastung bei Männern mit $\bar{x} = 46{,}2$ darstellt (was deutlich unter den hier vorliegenden Werten liegt) und gleichzeitig in der hier vorgenommenen Untersuchung Trait und State Angst praktisch gleich hoch ausfallen (die Anzahl der Fragen ist gleich). Die Korrelation der State- und Trait-Werte mit den verschiedenen Mathetest-Skalen ist der folgenden Tabelle 24 zu entnehmen.

Tabelle 24 Zusammenhang von der Mathematikskalen mit State und Trait-Angst (N = 79).

	Geometrie und grafische Funktionen	Prozedurales Rechnen	Mathematische Literalität	Komplexes Rechnen	Gesamt-score	Trait
State	0,15	0,14	0,12	0,14	0,16	0,62**
Trait	0,06	0,09	0,04	0,07	0,08	

Anmerkung. P < 0,01**

Demnach besteht der höchste Zusammenhang zwischen Gesamtscore und State-Angst mit r = 0,16. Keine der Korrelationen zu den Skalen des Mathetests ist signifikant (einseitig, 5% Niveau), was im Sinne der diskriminanten Validität (Lienert & Raatz, 1994) als positiv zu werten ist. Dementsprechend kann auch Hypothese 5 voll bestätigt werden.

7.4.3 Verbale Intelligenz

Insbesondere wegen dem hohen Textanteil einer der Skalen (Mathematische Literalität) liegt es nahe, den Zusammenhang der einzelnen Skalen mit den verbalen Fähigkeiten der Probanden zu untersuchen. Der Verbale Kurz-Intelligenztest (VKI) stellt ein von Anger et al. (1980) konstruiertes Verfahren dar, bei dem unter Zeitbegrenzung 20 Wörter jeweils einem von vier Bildern zugeordnet werden müssen. Er wird zusätzlich zur Deutschnote eingesetzt, da Noten grundsätzlich auch durch soziale Kompetenz zustande kommen, die bei dem verwendeten Mathetest keinen Einfluss hat. Der VKI wurde auf Basis der Wort-Bild-Tests entwickelt und soll in erster Linie verbale Intelligenz, aber auch allgemeines Urteilsvermögen erfassen (Frings, 2002, S. 239). Im vorliegenden Fall wurde der VKI vor dem Mathetest durch eine Teilstichprobe von 58 Studenten (15 Männer, 41 Frauen, 2 ohne Angabe) der Universität Mannheim bearbeitet. Der Mittelwert von $\bar{x} = 15,78$ (SD = 2,46) gelösten Aufgaben liegt über den Normen (Gesamtbevölkerung) des Manuals mit $\bar{x}_{NORM} = 14,2$, die Standardabweichung hingegen unter dem Normwert von SD_{NORM} = 4,2. Dieser Unterschied verwundert nicht sonderlich und ist sicherlich Ausdruck des Fähigkeitsprofils der Teilstichprobe. Die folgende Tabelle 25 zeigt den Zusammenhang des VKI-Rohwerts mit den vier Mathematikskalen und dem Gesamtscore.

Tabelle 25 Korrelation von VKI mit den Mathetestskalen und Gesmtscore , N = 58.

	Geometrie und grafische Funktionen	Prozedurales Rechnen	Mathematische Literalität	Komplexes Rechnen	Gesamt-score
VKI	0,19	0,09	0,40*	0,01	0,21

Anmerkung. p < 0,05*, einseitig. p < 0,01**, einseitig.

Lediglich die Skala mathematische Literalität weist eine signifikante Korrelation zum VKI-Wert auf, was den Erwartungen entspricht und demnach Hypothese 6 bestätigt.

7.4.4 Numerische Intelligenz

Der IST-2000R (Liepmann et al., 2007) erfasst in Form des Grundmoduls, als weitgehend eigenständige Komponenten anhand von 9 Subtests, die figurale, verbale und numerische Intelligenz (vgl. Abschnitt 3.1.5). Zwei der drei Aufgaben, die der numerischen Intelligenz zuzuordnen sind, stellen die Subgruppen Zahlenreihen und Rechenaufgaben dar. Liepmann et al. (2007) berichten für die numerische Intelligenz

den höchsten Zusammenhang zur Mathematiknote (r = -0,45) und den zweitniedrigsten zur Deutschnote (r = -0,04). Dementsprechend wird erwartet, dass alle Skalen des Mathetests signifikant mit beiden Aufgabengruppen und dem Gesamtscore korrelieren. Explizit für die Skalen prozedurales Rechnen und komplexes Rechnen werden, konzeptbedingt (vgl. Abschnitt 3.1.6.3 und 3.1.6.4), besonders hohe Korrelationen erwartet.

Insgesamt wurden 532 Teilnehmern neben dem Mathetest auch die Subtests Rechenaufgaben und Zahlenreihen des IST-2000R vorgelegt. Die folgende Tabelle Tabelle 26 zeigt die Interkorrelation, einmal für alle 532 Teilnehmer und zudem für jene Teilnehmer die im Summenwert aus den Subaufgaben Rechenaufgaben und Zahlenreihen einen Wert ungleich 0 erreichten. Dieses Vorgehen wurde gewählt, da es recht unwahrscheinlich erschien, dass 11,5% der Probanden (61 Personen) bei motivierter Bearbeitung der IST-Subtests einen Score von 0 in beiden Subaufgaben (Summe) erreichen. Im Falle des Mathetest-Gesamtscores erreichten lediglich 4 Personen (0,03%) einen Score von 0.

Tabelle 26 Korrelationen von Mathetestskalen mit IST-Subtests Rechenaufgaben, Zahlenreihen und deren Summe.

IST-Skala	Geometrie und grafische Fkt.	Mathematische Literalität	Komplexes Rechnen	Prozedurales Rechnen	Gesamt-score
		Alle Daten (n = 532)			
Rechen-aufgaben	0,22	0,46	0,40	0,40	0,46
Zahlenreihen	0,16	0,43	0,38	0,34	0,40
Beide	0,20	0,49	0,42	0,40	0,47
		IST-Gesamtscore≠0 (n = 471)			
Rechen-aufgaben	0,37	0,48	0,47	0,52	0,58
Zahlenreihen	0,27	0,42	0,42	0,41	0,47
Beide	0,37	0,53	0,52	0,54	0,61

Anmerkung. Ale Korrelationen sind hochsignifikant ($p < 0,01$, einseitig).

Wie erwartet, zeigen alle Skalen einen deutlichen, hochsignifikanten Zusammenhang zu den drei IST 2000R-Maßen. Nach Ausschluss der Personen mit einem Score von 0 in

beiden IST-Aufgabengruppen steigen diese Koeffizienten weiter an, zeigen jedoch dasselbe Bild. Der höchste Zusammenhang findet sich zwischen dem Gesamtscore und der Summe beider IST-Aufgaben, was mit gängigen Symmetrieprinzipien (Brunswik, 1955; Wittmann, 1985) in Einklang steht. Diese Ergebnisse berechtigen zur Annahme der Hypothese 7.

7.4.5 Schulnoten

Unter dem Gesichtspunkt der konvergenten Validität (Campbell & Fiske, 1959; Lienert & Raatz, 1994) sollten Zusammenhänge zwischen dem Mathetest und der korrespondierenden Schulnote bestehen. Die Zusammenhänge mit den Noten in anderen Schulfächern sollten im Sinne der diskriminanten Validität sehr niedrig ausfallen. Die folgende Tabelle 27 zeigt schließlich den Zusammenhang des Mathetests mit dem Mittel der letzten beiden Deutsch- und Mathenoten.

Tabelle 27 Zusammenhang der Skalen des Mathetests mit dem Mittel der letzten beiden Deutsch- und Mathenoten.

Noten	Geometrie und grafische Fkt.	Mathematische Literalität	Komplexes Rechnen	Prozedurales Rechnen	Gesamt-Score
Mathematik	-0,32	-0,32	-0,43	-0,39	-0,42
Deutsch	-0,13	-0,12	-0,21	-0,17	-0,18

Anmerkung. Alle Korrelationen sind hochsignifikant ($p < 0,01$). Niedrige Werte stehen für bessere Noten. $N = 1436$. Es wurden nur Fälle gewählt die Deutsch- und Mathenote enthielten.

Wie ersichtlich, korrelieren alle Skalen signifikant mit der Mathenote und der Deutschnote. Die beiden Noten korrelieren untereinander zu $r = 0,36$ ($N = 1436$, $p < 0,01$). Die Korrelation mit der Mathenote ist in jedem Fall doppelt so hoch wie mit der Deutschnote.

Steiger (1980) erarbeitete eine Formel, um die Unterschiedlichkeit zweier abhängiger Korrelationen zu prüfen, da in diesem Fall eine einfache Fisher-Z-Transformation (Cohen et al., 2003) nicht ausreichend ist. Die Formel (Steiger, 1980, S. 45) wurde in ein SPSS-Skript übersetzt und findet sich in Anhang 12.3. Das Vorgehen ist dem hierachischen F-Test konzeptuell sehr ähnlich und ergibt für alle Korrelationen in

Tabelle 27 einen hochsignifikanten Unterschied. Demzufolge können auch Hypothese 8 und 9 angenommen werden.

7.5 Schlussfolgerungen

Der vorangegangene Abschnitt 7 hat sich der Passung der Endform gemäß KTT gewidmet und in diesem Rahmen die Reliabilität (7.2.2) und (Konstrukt)validitäten (7.4) geprüft. Letzteres geschah anhand von 6 a-priori aufgestellten Hypothesen (H4-H9). Alle Hypothesen konnten bestätigt werden, was die Konstruktvalidität im Sinne der diskriminanten und konvergenten Validität sicherstellt. Auch die Reliabilitäten erreichten zufriedenstellende Werte, die recht gut mit den vorläufigen Ergebnissen der Vorform (6.2) übereinstimmen.

8 Konfirmatorische Prüfung der theoretischen Annahmen

Nachdem grundlegende Anforderungen an Reliabilität und Validität im Rahmen der KTT gesichert wurden, soll in diesem Abschnitt 8 eine konfirmatorische Prüfung der Modellstruktur erfolgen. Dies gliedert sich in zwei Abschnitte. Zum einen den konfirmatorischen Nachweis der vier Inhaltsbereiche (Skalen gemäß Abschnitt 3.1.6) und zum anderen die taxonomische Ordnung der Endform.

Der Nachweis der vier Inhaltsbereiche ist unterteilt in Untersuchungen auf Item- und Parcel-Ebene und orientiert sich methodisch an den Ausführungen der Abschnitte 4.2 und 4.3. Der Einsatz der Lernzieltaxonomie umfasst die Frage nach der Reliabilität der Zuordnungen der Items zu den 6 Taxonomiestufen und der Übereinstimmung zwischen dem, was Lehrer als besonders wichtig empfinden und dieser Zuordnung.

Aus den konfirmatorischen Untersuchungen dieses Abschnitts ergeben sich einige weiterführende Fragestellungen, die schließlich in Abschnitt 9 geprüft werden. Zunächst werden jedoch im folgenden Abschnitt 8.1 einige weitere a-priori Hypothesen aufgestellt, um das soeben zusammengefasste konkret und prüfbar darzustellen.

8.1 Hypothesen III

Hypothese H10: Der Test ist mehrdimensional.

Hypothese H11: Die Struktur entspricht in ausreichendem Maß der postulierten Skalenstruktur, d.h. vier korrelierte Faktoren, die sich tendenziell trennen lassen.

Weitere Hypothesen betreffen die Frage, inwiefern sich die Items den Taxonomiestufen nach Anderson & Krathwohl (2001) reliabel zuordnen lassen und inwiefern der Test jene Stufen erfasst, die Lehrer für besonders wichtig halten (vgl. Abschnitt 3.3).

Hypothese H12: Die Experten beurteilen nicht alle Taxonomiestufen als gleich stark im Test vertreten

Hypothese H13: Eine ausreichend reliable Zuordnung von Items zu den Stufen ist möglich.

Hypothese H14: Der Test erfasst aus Lehrersicht vor allem jene Stufen die für wichtig bei Berufseinsteigern angesehen werden.

8.1.1 N-Dimensionalität der Inhalte: DIMTEST – DETECT

Bei Anwendung der DIMTEST-Prozedur (Stout, 1987) auf die Endform, ergibt sich ein DIMTEST T von T = 10,23 (p = 0,00), womit die Hypothese 10 direkt angenommen werden kann. Somit kann zur DETECT Prozedur übergegangen werden, die für 4 Dimensionen den besten Fit der Statistik D(P) mit D(P) = 0,41 ergibt, was als schwacher Hinweis auf homogene, unterscheidbare Itemcluster zu betrachten ist (Gierl & Wang, 2005) und etwas schlechter als der entsprechende Wert aus der Expra-Reanalyse ausfällt (D(p) = 0,47) Die Aufteilung in 4-Cluster zeigt folgende Abbildung 22.

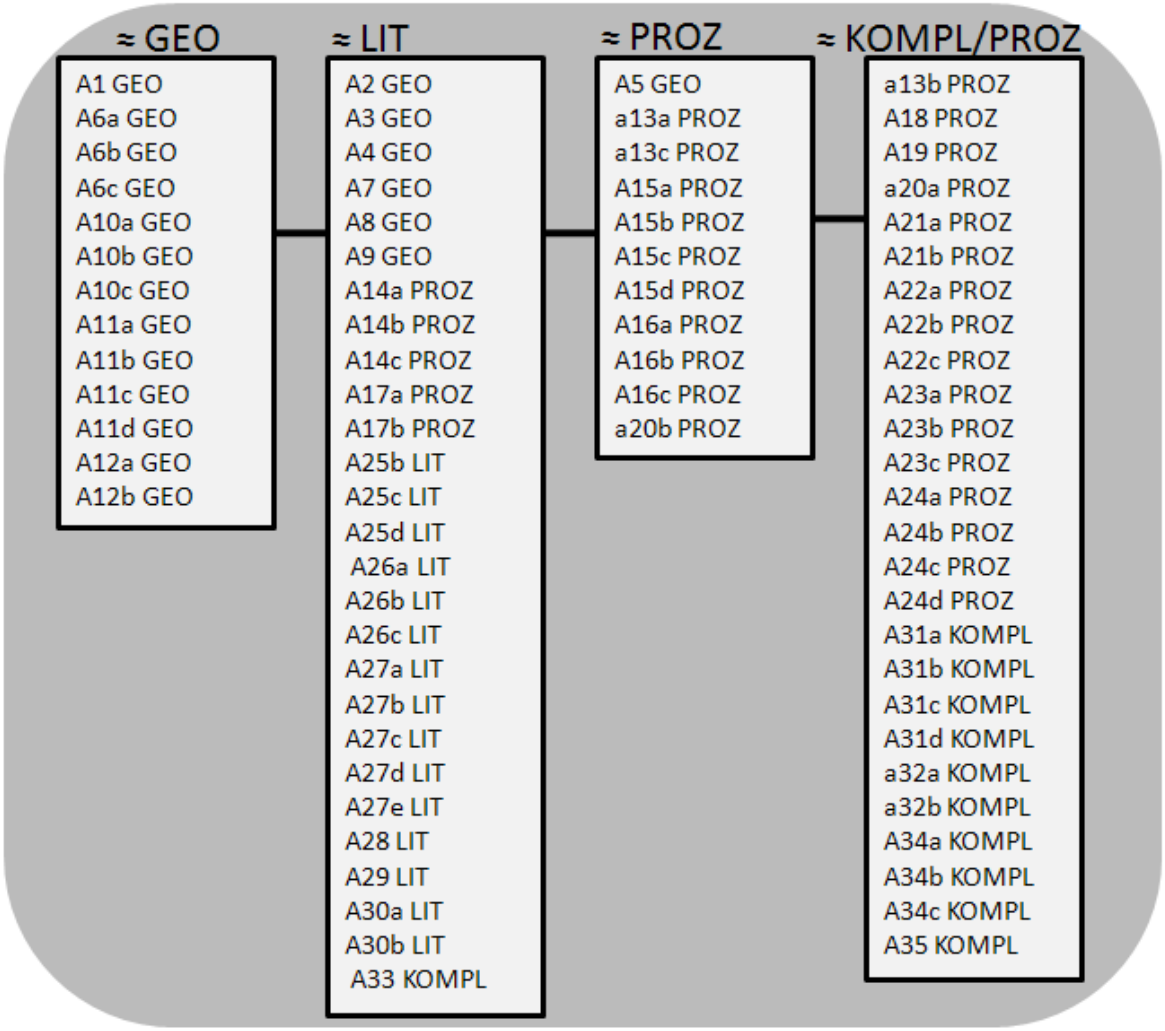


Abbildung 22 Exploratorische DETECT-Lösung der Endform, *N* = 1554.

Ein Cluster besteht praktisch ausschließlich aus Items der Skala Geometrie und grafische Funktionen, der zweite Cluster (von links) enthält auch andere Items, jedoch in auffälliger Weise alle Items der Skala mathematische Literalität. Der Dritte Cluster enthält mit einer Ausnahme (A5) ausschließlich Items der Skala prozedurales Rechnen und der 4. Cluster schließlich eine Mischung aus den Skalen prozedurales Rechnen und komplexes Rechnen. Die Lösung ist dahingehend zu beurteilen, dass einzelne Skalen die Cluster zu dominieren scheinen, jedoch eine Trennung von komplexem und prozeduralem Rechnen schwer erscheint. Ins Bild passt der r_{max} Wert (Zhan & Stout, 1999), der das Ausmaß von Einfachstruktur beschreibt und mit r_{max} = 0,72, einen Hinweis auf eher schwache Einfachstruktur gibt, was nach den bisherigen Ergebnissen nicht überrascht, aber deutlich besser als der Wert aus der Reanalyse des Expra-Tests mit r_{max} = 0,51 ausfällt.

Umso interessanter ist das Ergebnis einer DETECT Lösung, bei der eine Aufteilung in 4 Cluster, die exakt den 4 Mathetestskalen entsprechen, erzwungen wird

(konfirmatorischer Modus). Es ergibt sich ein DETECT-Wert von D(P) = 0,32, (r_{max} = 0,56) der zwar als relativ schlecht anzusehen ist (geringer Hinweis auf Multidimensionalität) (Zhang & Stout, 1999), jedoch nicht wesentlich schlechter als die rein exploratorische Lösung ausfällt. Ein Grund liegt sicherlich darin, dass die angenommene Struktur keineswegs einer Einfachstruktur entsprechen kann und soll, was die Schätzungen von DETECT verzerren könnte. So heißt es bei Zhang & Stout (1999) für den Fall eines r_{max} Wertes unter r_{max} = 0,80: „in particular, it can still locate relatively dimensionally homogeneous clusters; however, there is no longer a unique best or correct partition to be found by DETECT because there will be little to no separation between some of the clusters found" (S. 215).

Verglichen mit anderen Ergebnissen im selben Bereich (Mathematiktests), wie bei Gierl und Wang (2005) oder bei der Analyse des Vortests in Abschnitt 4.4.2.1 sind die Ergebnisse befriedigend. Die gefundene Struktur entspricht einigermaßen den Erwartungen, die unklaren Eigenschaften von DETECT bei nicht vorliegender Einfachstruktur (Zhang & Stout, 1999) lassen das Ergebnis insgesamt gut erscheinen. Die Ergebnisse der primär datengetriebenen Verfahren DIMTEST und DETECT sprechen dafür detailliertere Strukturanalysen mit dazu besser geeigneten Methoden, wie NOHARM, klassischer Faktorenanalyse und SEM-Ansätzen, durchzuführen. Erst dann kann Hypothese H11 mit ausreichender Sicherheit beantwortet werden.

8.2 Strukturanalysen der Inhaltsfacetten auf Itemebene

Um zu zeigen, dass für den vorliegenden Datensatz eine lineare Faktorenanalyse auf Itemebene ungeeignet ist (vgl. Abschnitt 4.2.4), wird in diesem Abschnitt zunächst das Problem von Schwierigkeitsfaktoren praktisch dargestellt.

Es sei vorweggenommen, dass die Anwendung des unter Abschnitt 4.3.2 vorgestellten WLSMV-Schätzverfahrens auf Itemebene, d.h. mit 77 Einzelvariablen, für die vorliegenden Daten nicht möglich war. MPLUS (Muthén & Muthén, 2007) nutzt tetrachorische Korrelationen (Cohen et al., 2003) im Rahmen der Berechnungen und gab zahlreiche Warnmeldungen über extrem hohe Korrelationen zwischen einzelnen Items an. Im Output zur G-Faktorlösung heißt es unter anderem INFORMATION FROM THESE VARIABLES CAN BE USED TO CREATE ONE NEW VARIABLE bei sehr hohen tetrachorischen Korrelationen. Zwar konvergierte die Lösung, doch ist

ihr Nutzen unter diesen Bedingungen fraglich. Daher wird die WLSMV-Technik unter Abschnitt 8.3.2 nur auf Parcels angewendet.

8.2.1 Faktorenanalyse

Betrachtet man die Pattern-Matrix der 3-faktoriellen schiefwinklig rotierten (Gorsuch, 1983) Faktorenanalyse der 4 Skalen auf Itemebene gemäß Tabelle 28 wird das Problem der linearen Faktorenanalyse bei binären Items (vgl. im Detail Abschnitt 4.3) ersichtlich. Der Mathematiktest ist so aufgebaut, dass für alle Skalen außer komplexem Rechnen die Schwierigkeit zum Skalenende hin zunimmt. Für komplexes Rechnen ist dies nicht der Fall, da die Skala (bis auf A31a) generell eher schwierige Aufgaben enthält.

Tabelle 28 Oblimin Pattern-Matrix und Itemschwierigkeiten der Endform (*N* = 1554).

Skalenzuordnung	Schwierigkeit	Faktoren		
		1	2	3
Geometrie und grafische Fkt.				
A1	0,86			
A2	0,89			**0,22**
A3	0,81			**0,25**
A4	0,88			
A5	0,66			**0,27**
A6A	0,59	0,20		**0,54**
A6B	0,60	0,20		**0,55**
A6C	0,58	0,20		**0,54**
A7	0,36	0,28		
A8	0,45	**0,26**		0,22
A9	0,35	**0,32**		
A10A	0,35	**0,53**		
A10B	0,32	**0,54**		
A10C	0,31	**0,55**		
A11A	0,23	**0,71**		
A11B	0,27	**0,68**		
A11C	0,23	**0,74**		
A11D	0,16	**0,71**		
A12A	0,09	**0,66**		
A12B	0,08	**0,68**		
Prozedurales Rechnen				
A13A	0,95		**0,22**	**0,22**
A13B	0,71			
A13C	0,52			**0,26**
A14A	0,95		**0,35**	
A14B	0,89		**0,31**	
A14C	0,92		**0,41**	
A15A	0,79			**0,78**
A15B	0,81			**0,79**
A15C	0,76			**0,72**
A15D	0,67			**0,66**
A16A	0,50			**0,38**
A16B	0,38			**0,45**
A16C	0,60			**0,40**
A17A	0,55			
A17B	0,28			
A18	0,59	0,23		**0,24**
A19	0,45	**0,24**		
A20A	0,49	**0,34**		0,26
A20B	0,22	0,24		

Tabelle 28 Fortsetzung.

Skalenzuordnung	Schwierigkeit	Faktoren		
		1	2	3
A21A	0,64	0,23		**0,33**
A21B	0,52	0,31		**0,33**
A22A	0,66			**0,24**
A22B	0,45	**0,30**		0,23
A22C	0,55	0,30		**0,37**
A23A	0,45	**0,46**		
A23B	0,15	**0,47**		
A23C	0,42	**0,49**		
A24A	0,42	**0,54**		
A24B	0,17	**0,72**		
A24C	0,33	**0,62**		
A24D	0,13	**0,71**		
Mathematische Literalität				
A25B	0,86		**0,74**	
A25C	0,82		**0,74**	
A25D	0,80		**0,76**	
A26A	0,96		**0,47**	
A26B	0,84		**0,47**	
A26C	0,68		**0,51**	
A27A	0,81		**0,39**	
A27B	0,80		**0,45**	
A27C	0,77		**0,48**	
A27D	0,65		**0,37**	
A27E	0,57		**0,40**	
A28	0,43	0,22	**0,31**	
A29	0,17	**0,27**		
A30A	0,32	**0,47**	0,20	
A30B	0,27	**0,49**	0,21	
Komplexes Rechnen				
A31A	0,82		**0,32**	
A31B	0,22	**0,38**		
A31C	0,39	**0,43**		
A31D	0,26	**0,54**		
A32A	0,55	**0,34**	0,27	
A32B	0,39	**0,46**	0,21	
A33	0,58	**0,27**	**0,27**	
A34A	0,40	**0,51**		
A34B	0,51	**0,42**		
A34C	0,16	**0,60**		
A35	0,45	**0,51**		

Anmerkung. Oblimin, Gamma = 0. Die jeweils größte Ladung ist hervorgehoben. Ladungen kleiner 0,20 werden nicht angezeigt. Faktorinterkorrelationen: $r_{f1f2} = 0{,}31$, $r_{f2f3} = 0{,}38$, $r_{f1f3} = 0{,}47$.

Komplexes Rechnen, als schwerste Skala, lädt zusammen mit den jeweils schwersten Items der anderen drei Skalen auf Faktor 1. Generell weisen die schwereren Items ($p < 0,30$) nur auf Faktor 1 substantielle Ladungen auf. Dieser ist somit klar als Schwierigkeitsfaktor zu betrachten. Dies heißt letztlich, dass hier von konfirmatorischen Analysen der Struktur auf Itemebene anhand klassischer faktorenanalytischer Techniken abzuraten ist.

8.2.2 NOHARM

Das in NOHARM (McDonald, 1999) realisierte Modell der konfirmatorischen Lösung entspricht 4 Faktoren, denen jeweils die Items einer Skala zugeordnet sind. Eine Korrelation zwischen den Faktoren ist möglich. Das Modell-Prinzip ist demnach identisch mit den Strukturgleichungsmodellen die noch in Abschnitt 8.3.2 getestet werden, mit dem Vorteil, dass hier das binäre Itemformat konzeptbedingt kein Problem dargestellt (IRT-Ansatz, siehe Abschnitt 4.2.6.). Neben einer 4-faktoriellen Lösung wird auch eine dreifaktorielle Lösung geprüft, in der die oberflächlich ähnlichen Skalen prozedurales- und komplexes Rechnen einen gemeinsamen Faktor bilden (vgl. Abschnitt 3.1.6). Um den Unterschied im Fit verschiedener Modelle besser einschätzen zu können wurde darüber hinaus eine G-Faktormodell-Lösung berechnet. Die Kennwerte finden sich in der folgenden Tabelle 29.

Tabelle 29 Tanaka (GFI)-Index und RMSR Fit-Index für 1, 3 und 4-Faktorielle Lösungen (*N* = 1554).

Anzahl Faktoren	Tanaka (GFI)	RMSR
1	0,94	0,012
3	0,93	0,012
4	0,95	0,011

Gemäß McDonald's (1999) Faustregel weist nur die 4-Faktorielle Lösung einen guten Fit auf (GFI ≥ 0,95), wenngleich die Werte für eine Ein- und Drei-Faktorlösungen sehr nahe beieinander liegen. Gleiches gilt für die Residuen, die im RMSR zum Ausdruck kommen und alle einen deutlich besseren Wert aufweisen als von McDonald und Fraser (1988) gefordert. Sie legten als Daumenregel $RMSR \leq 4 \cdot 1/\sqrt{N}$ fest, was in diesem Beispiel einen Wert von $RMSR = 0,101$ ergibt, der für einen guten Fit zu unterschreiten

ist. Da die 4-Faktorielle Lösung den besten Fit aufweist und am ehesten der postulierten Theorie entspricht, wurden die Faktorladungen in der Tabelle 30 abgetragen.

Tabelle 30 Faktorladungen der konfirmatorischen, obliquen NOHARM Lösung (*N* = 1554).

Itemposition	Skala			
	Geometrie und grafische Fkt.	Mathematische Literalität	Prozedurales Rechnen	Komplexes Rechnen
1	0,40	0,80	0,55	0,63
2	0,54	0,71	0,28	0,65
3	0,62	0,71	0,43	0,78
4	0,45	0,75	0,52	0,79
5	0,49	0,60	0,41	0,79
6	0,57	0,72	0,51	0,84
7	0,59	0,39	0,70	0,72
8	0,57	0,61	0,73	0,79
9	0,60	0,68	0,69	0,79
10	0,67	0,61	0,69	0,70
11	0,66	0,65	0,47	0,66
12	0,79	0,68	0,60	
13	0,83	0,53	0,46	
14	0,83	0,86	0,32	
15	0,84	0,92	0,19	
16	0,82		0,57	
17	0,90		0,46	
18	0,87		0,78	
19	0,90		0,16	
20	0,94		0,75	
21			0,81	
22			0,56	
23			0,65	
24			0,80	
25			0,66	
26			0,70	
27			0,69	
28			0,73	
29			0,90	
30			0,81	
31			0,95	

Anmerkung. Es existieren in der konfirmatorischen Lösung keinerlei Nebenladungen. Die Nummerierung bezieht sich aus Platzgründen auf die Position in der Skala der Endform. Tabelle 28 zeigt die Klarnamen der Items.

Es zeigt sich, dass alle Ladungen eine substantielle Höhe aufweisen, jedoch geht dieser Effekt deutlich auf Kosten einer hohen Interfaktorkorrelation. Die Korrelation der

Faktoren variiert zwischen $r = 0{,}65$ (mathematische Literalität, Geometrie und grafische Fkt.) und $r = 0{,}87$ (prozedurales-, komplexes Rechnen). Die Tatsache, dass die höchste Korrelation zwischen prozeduralem- und komplexem Rechnen auftritt entspricht den Erwartungen. Abschließend zeigt Tabelle 31 sämtliche Faktorinterkorrelationen der 4-Faktor NOHARM-Lösung.

Tabelle 31 Intekorrelationen der 4 Faktoren einer obliquen NOHARM-Lösung ($N = 1554$).

Skala	Geometrie und grafische Fkt.	Prozedurales Rechnen	Mathematische Literalität
Prozedurales Rechnen	0,79		
Mathematische Literalität	0,65	0,70	
Komplexes Rechnen	0,79	0,87	0,73

8.2.3 Allgemeine Schlussfolgerungen aus der NOHARM-Lösung

Die recht hohen Interkorrelationen mögen auf den ersten Blick entmutigend wirken, sind jedoch deutlich niedriger als jene, die Gierl und Wang (2005, S. 16) bei einer Reanalyse des SAT-Mathematikteils fanden. Sie unterschieden zwischen Algebra, Arithmetik, Geometrie und Sonstiges, was zu Korrelationen zwischen 0,95 und 1 führte. Das heißt die kleinste Korrelation einer vergleichbaren Analyse (Gierl verwendete auch NOHARM) war deutlich höher als die größte Interkorrelation gemäß Tabelle 31. Dies spricht für die hier getestete 4-faktorielle Lösung. Ein Problem liegt sicherlich darin, dass die Korrelation zwischen den Faktoren dadurch immens ansteigt, dass in NOHARMS konfirmatorischem Modus keine frei zu schätzenden Nebenladungen vorgesehen sind. Andererseits würden vorhandene Nebenladungen, analog zu einer linearen Faktorenanalyse, die Interfaktorkorrelation nur scheinbar verringern und subjektiv mehr Unabhängigkeit zwischen den Faktoren implizieren, als tatsächlich vorhanden ist.

8.3 Strukturanalysen der Inhaltsfacetten auf Parcel-Ebene

Das prinzipielle Vorgehen, ebenso wie Vor- und Nachteile des Parceling wurden bereits unter Abschnitt 4.3.1 erläutert. Die im folgenden verwendete Parceling-Methode ist an Jägers (1982) Strukturmodell der Intelligenz und hier speziell die Inhaltsfacetten, angelehnt. Dementsprechend werden zunächst innerhalb jeder Skala Parcel gebildet, die

zunächst das leichteste und das schwerste Item enthalten, anschließend das zweitleichteste und das zweitschwerste, bis alle Items einer Skala in den Parcels aufgegangen sind. Die Zuordnung der endgültigen Items des Tests zu den Parcels ist der folgenden Tabelle 32 zu entnehmen.

Tabelle 32 Schwierigkeitsbasierte Parcelbildung der Endform auf Skalenebene.

Parcel Nr.	Geometrie und grafische Fkt.	Prozedurales Rechnen	Mathematische Literalität	Komplexes Rechnen
		Items der Endform		
1	A2,A12b	A13a, A14a, A24d	A26a, A25b, A29	A31a, A33, A34c
2	A4, A12a	A14c, A23b	A26b, A30b	A32a, A31b
3	A1,A11d	A14b, A24b	A25c, A30a	A34b, A31d
4	A3, A11c	A15b, A20b	A27a, A28	A35, A32b
5	A5, A11a	A15a, A17b	A25d, A27e	A34a, A31c
6	A6b, A11b	A15c, A24c	A27b, A27d	
7	A6a, A10c	A13b, A16b	A27c, A26c	
8	A6c, A10b	A15d, A24a		
9	A8, A10a	A22a, A23c		
10	A7, A9	A21a, A23a		
11		A16c, A22b		
12		A18, A19		
13		A17a, A20a		
14		A22c, A16a		
15		A13c, A21b		

Für jede Skala die eine ungerade Anzahl von Items enthielt wurde das erste Parcel aus den zwei leichtesten und dem schwersten Item gebildet. Bei allen folgenden Analysen werden die Parcels nach ihrer Nummer und der Abkürzung benannt. Dass 3. Parcel der Skala komplexes Rechnen enthält z.B A34b sowie A31d und wird KOMPL3 genannt.

8.3.1 Faktorenanalyse

Bei Faktorenanalysen auf Parcelebene zeigte sich, dass es nicht zuverlässig gelingt eine Trennung der Skalen prozedurales Rechnen und komplexes Rechnen zu erreichen. Da durch einen zusätzlichen vierten Faktor weder ein besser interpretierbares Muster (d.h. die Items einer Skala laden am stärksten auf jeweils einem Faktor) noch eine bedeutsame zusätzliche Varianzaufklärung (49% Varianzaufklärung, anstelle von 45%) zu erreichen war, wird hier nur die dreifaktorielle Lösung berichtet. Tabelle 33 stellt die Pattern-Matrix (Gorsuch, 1983) samt Varianzaufklärung und Interfaktorkorrelation dar.

Tabelle 33 Pattern Matrix einer schiefwinkligen Faktorenanalyse der Parcels .

Parcel	Faktor 1	Faktor 2	Faktor 3
Geo1			**0,34**
Geo2		0,21	**0,42**
Geo3			**0,56**
Geo4	0,22		**0,51**
Geo5			**0,55**
Geo6			**0,91**
Geo7			**0,93**
Geo8			**0,94**
Geo9			**0,54**
Geo10	0,20	**0,30**	0,21
Proz1	**0,34**	0,26	
Proz2	0,25	**0,28**	
Proz3	0,23	**0,26**	0,21
Proz4	**0,61**		
Proz5	**0,51**		
Proz6	**0,68**		
Proz7	**0,59**		
Proz8	**0,67**		
Proz9	**0,62**		
Proz10	**0,71**		
Proz11	**0,62**		
Proz12	**0,39**		
Proz13	**0,43**	0,21	
Proz14	**0,74**		
Proz15	**0,63**		
Lit1		**0,72**	
Lit2		**0,64**	
Lit3		**0,74**	
Lit4		**0,61**	
Lit5		**0,81**	
Lit6		**0,65**	
Lit7		**0,74**	
Kompl1	**0,33**	**0,33**	
Kompl2	**0,44**		
Kompl3	**0,46**		
Kompl4	**0,44**		
Kompl5	**0,47**		

Anmerkung. Oblimin, Gamma =0 . Varianzaufklärung Faktor 1 = 18%, Faktor 2 = 13%, Faktor 3 = 14%. Interfaktorkorrelationen: $r_{F1\,F2} = 0{,}55$ $r_{F2\,F3} = 0{,}48$ $r_{F1F3} = 0{,}58$. Ladungen kleiner 0,20 wurden ausgeblendet.

Von den Faktoren in Tabelle 33 bildet Faktor 1 am ehesten prozedurales Rechnen und komplexes Rechnen ab, Faktor 2 mathematische Literalität und Faktor 3 schließlich Geometrie und grafische Funktionen. Sicherlich ist die Faktorlösung noch verbesserungswürdig, doch lässt sich tendenziell die Struktur des Mathematiktests wieder erkennen. Dass, falls Skalen einen gemeinsamen Faktor bilden, es sich um komplexes- und prozedurales Rechnen handeln würde ist aufgrund ihrer Ähnlichkeit plausibel. Detailliertere Strukturanalysen zum Testaufbau bedingen andere Testmethoden und werden im folgenden Abschnitt vorgenommen.

8.3.2 Strukturgleichungsmodelle

Dies ist der erste Abschnitt dieser Arbeit, in dem Strukturgleichungsmodelle angewendet werden. Um Bewertungsmaßstäbe für den Modellfit zu erhalten, werden im folgenden zunächst einige Fit-Indizes dargestellt und der Sinn von Cut-Off Kriterien hinterfragt, bevor die eigentlichen Modelle geprüft werden.

8.3.2.1 Sinn und Nutzen von Cut-Off Kriterien

Mittlerweile existiert eine Fülle von Fit-Indizes für Strukturgleichungsmodelle, von denen die meisten so genannte deskriptive Indizes darstellen (Kline, 2005). Die meisten Kennwerte basieren auf dem Vergleich des aufgestellten Modells zu einem Basis-Modell, das keinerlei Pfade enthält und bei dem alle Varianzen und Kovarianzen 0 sind (worst case szenario). Der GFI beispielsweise ist formuliert als $GFI = 1 - (F_{\min} / F_0)$, wobei F_{min} das Ergebnis einer Diskrepanzfunktion (die vom Schätzverfahren abhängt, z.B. ML) ist, die die Abweichung der modellinduzierten Kovarianzmatrix $\sum(\theta)$ von der tatsächlichen Kovarianzmatrix $\sum$ darstellt (Bollen & Long, 1993). Der Wert von F_{min} ist nicht normiert und daher für sich genommen kaum interpretierbar, er muss mit einem anderen Kennwert ins Verhältnis gesetzt werden. Hierzu dient F_0 welches die Abweichung der Matrix $\sum(\theta \mid \theta = 0)$ von der tatsächlich vorhandenen $\sum$ Matrix angibt. Im Fall des NFI wird ein ähnliches Vorgehen angewandt, nur basiert dieser Kennwert auf zwei χ^2-Werten: $NFI = 1 - \chi^2_{Modell} / \chi^2_0$ (Bentler, 1990). χ^2_{Modell} stellt hier eine einfache Umformung von F_{min} dar und zwar $\chi^2_{Modell} = F_{\min} \cdot (N-1)$, χ^2_0 hingegen das Analogon zu F_0. Zeitgenössische Lehrbücher geben häufig gewisse Anhaltspunkte, welcher Fit noch als gut zu bezeichnen ist. So schlagen Backhaus, Erichson, Plink und Weiber (2006) Werte für GFI und NFI größer 0,90 als gut vor.

Diese pauschale Aussage ist jedoch problematisch, da die Indizes selbst auch von der Art des zu testenden Modells abhängen. Die eben erwähnten Indizes weisen zum Beispiel einen umso besseren Fit auf, je mehr Pfade eingezeichnet werden, so dass sie alle bei einem *just-identified model* einen perfekten Fit von 1 aufweisen würden (Loehlin, 2004). Dies kann nicht zweckmäßig sein, weshalb eine weitere Familie von Fit-Maßen entwickelt wurde, die *parsimonious fit indices* genannt werden, sparsame Modelle im Fit belohnen und nicht sparsame bestrafen (Schumacker & Lomax, 2004). Die Indizes NNFI und AGFI gehören zu dieser Familie von Fit-Maßen und stellen Erweiterungen der bereits erwähnten Kennwerte NFI und GFI dar.

Ein weiteres Problem der bereits besprochenen Indizes ist, dass sie (zumindest implizit) davon ausgehen, dass - gegeben das aufgestellte Modell ist in der Population gültig - $F_{\min}$ (und damit auch χ^2_{Modell}) den Wert Null annehmen müsste. In diesem Fall würde die Verteilung von $F_{\min}$ einer zentralen χ^2-Verteilung folgen (McCallum, Browne & Hazuki, 1996). Dies ist jedoch in jeder empirischen Anwendung, auch wenn die Population komplett getestet werden könnte, höchst unwahrscheinlich. Der Grund liegt darin, dass ein Modell stets ein vereinfachtes Abbild der Realität ist. Raykov und Marcoulides (2006) fassen dieses Problem zusammen: „…by its very nature, a model cannot be correct because then it would have to be an exact copy of reality and therefore useless“ (S. 45). Daher wurde ein so genanntes Non-Zentralitätsmaß eingeführt, das definiert ist als $\tau_{Modell} = (\chi^2_{Modell} - df)/(N-1)$ und einen Kennwert für das Ausmaß, in dem das Modell nicht stimmt, darstellt (Loehlin, 2004). Praktisch bedeutet dies, dass ein χ^2-Wert von z.B. χ^2_{Modell} = 200 bei konstantem N zu einem höheren τ führt, wenn es sich um ein wenig sparsames Modell handelt (viele Pfade, wenige df bleiben frei) als wenn ein sehr sparsames Modell postuliert wurde (wenig Pfade, viele df frei). Auf diesem realistischerem Maß bauen z.B. der RMSEA-Index und der CFI-Index auf, für die Backhaus et al. (2006) Werte von RMSEA < 0,05 bzw. CFI > 0,90 als gut bezeichnen.

Während der CFI das Prinzip des NFI (siehe oben) auf den Nonzentralitätsparameter überträgt, mit $CFI = 1 - \tau_{Modell} / \tau_0$, ist der RMSEA anders aufgebaut und zwar in Form von $RMSEA = \sqrt{\tau_{Modell} / df}$ (Hu & Bentler, 1998). Dieser Index stellt sozusagen den Missfit pro Freiheitsgrad dar und bietet darüber hinaus den Vorteil, dass es möglich ist ein Konfidenzintervall zu bestimmen (MacCallum et al., 1996).

Da zahlreiche Empfehlungen zur Interpretation von Fit-Indizes existieren (vgl. Backhaus et al., 2006; Loehlin, 2004), könnte dies den falschen Eindruck vermitteln, es

bestünde Einigkeit in Bezug auf deren Interpretation. So empfehlen beispielsweise Hu und Bentler (1998, S. 449) für (ML-basierte) CFI und RMSEA-Werte Cut-Offs von 0,95 (CFI) und 0,06 (RMSEA) als Indikatoren für einen guten Fit, was von den bereits erwähnten Empfehlungen abweicht. In einer sehr umfangreichen Simulationsstudie, die auf der Fähigkeit von Fit-Indizes zwischen falsch und richtig spezifizierten Modellen zu unterscheiden basierte, bekräftigen die Autoren (Hu & Bentler, 1999) diese Empfehlung und kritisieren zugleich ältere Daumenregeln ob ihrer nicht ausreichenden Begründbarkeit: „researchers often question the adequacy of these conventional cutoff criteria due to the lack of empirical evidence and compelling rationale for these rules of thumb“ (S. 4). Doch auch dieser - quasi empirische - Versuch bessere Richtlinien zu etablieren, wurde von Marsh, Hau und Wen (2004) deutlich kritisiert. Letztere fanden heraus, dass bei Hu Bentler‘s (1999) Vorgehen ein simpler Chi-Quadrat Wert die besten Ergebnisse (besser als alle Fit-Indizes) liefern würde und empfehlen solche Daumenregeln keinesfalls als goldene Regeln zu betrachten, sondern vielmehr theoretische Überlegungen (z.B. Plausibilität des Modells) sowie den Vergleich verschiedener konkurrierender Modelle als maßgeblich bei einer Entscheidung für oder gegen ein Modell zu betrachten. Äußerst interessant ist auch der Ansatz von Beauducel und Wittmann (2005), die eine Monte Carlo Studie durchführten, in der der Fokus auf einem Abschneiden der Indizes bei, für die psychologische Forschung typischen, niedrigeren Hauptladungen und deutlicheren Nebenladungen als in der Arbeit von Hu und Bentler (1999), (dort lagen die Hauptladungen zwischen 0,70 und 0,80) lag. Sie kamen zu wichtigen Schlussfolgerungen, von denen einige hier aufgezählt werden sollten (Beauducel & Wittmann, 2005):

- Hohe Heterogenität der verschiedenen Fit-Indizes, 4 Faktoren in einer PCA der Fit-Indizes (S. 58)
- Kleine Abweichungen von der Einfachstruktur werden von inkrementellen Indizes (z.B. CFI, NNFI) sowie im GFI bestraft (S. 41) jedoch weniger von RMSEA/ SRMR
- “most of the models based on salient loadings of .60 and .80 were accepted on the basis of the > .90 threshold and most of these models were rejected on the basis of the > .95 threshold (S. 59)

Was die konkrete Höhe eines möglichen Cutoffs für den RMSEA-Index angeht kamen jüngst, im Jahre 2008, Chen et al. zu der Schlussfolgerung: „..there is little empirical support for the use of .05 or any other value as universal cutoff values to determine

adequate model fit..“ (S. 462). Was bedeuten nun diese, eher entmutigenden, Ergebnisse für diese Arbeit und für konfirmatorische Tests der aufgestellten Modelle?

Letztlich läuft alles darauf hinaus, dass Cutoff-Werte nur eine gewisse Orientierung darstellen können. So ist es vielleicht fraglich, ob ein CFI (als Beispiel für ein inkrementelles Maß) von 0,95, 0,90 oder auch 0,89 noch für einen ausreichenden Modellfit spricht, doch kaum ein Forscher wird einen CFI von 0,65 für akzeptabel halten. Ebenso verhält es sich mit dem RMSEA (als Beispiel für ein populationsbasiertes Maß), für den man anscheinend auch nur einen gewissen Orientierungspunkt angeben kann.

Entscheidend muss es daher sein, ein theoretisch begründetes Modell aufzustellen und am Besten (so es von der Theorie her Sinn ergibt) alternative Varianten gegenüber zu stellen. Im Falle von *nested-models* (Loehlin, 2004) ergibt sich hier zudem der Vorteil, die Unterschiede der Modelle tatsächlich vergleichen zu können. Dies ist für *non-nested* Modelle, für die kein χ^2 – Differenz-Test durchgeführt werden kann, nicht möglich. Für letztere muss demnach gelten, dass sehr große Unterschiede in den Fit-Indizes eine Bedeutung haben und bei geringen Unterschieden das Prinzip der Sparsamkeit und theoretischen Passung gelten muss.

Im folgenden werden vor allem der RMSEA und CFI sowie der χ^2 – Wert (der praktisch allen Maßen implizit zugrunde liegt) herangezogen. Die anderen Indizes werden nur detailliert berichtet, falls sie zu unterschiedlichen Schlussfolgerungen führen würden, da es nicht sinnvoll erscheint 10 Fit Indizes für jedes Modell darzustellen und zu interpretieren. Ferner muss wie bereits angesprochen (siehe Abschnitt 4.3.2 oben) zwischen den Maßen gemäß unterschiedlichen Schätzverfahren (ML versus WLSMV) unterschieden werden.

8.3.2.2 Modelle mit einem G-Faktor

Es wurde an keiner Stelle dieser Arbeit erwartet, dass ein G-Faktor Modell den Bereich der Mathematikfähigkeiten ausreichend abdeckt. Um der Kritik vorzubeugen, dass ein einfaktorielles Modell vielleicht die beste Lösung dargestellt hätte wurde dennoch eine mögliche Passung geprüft. Es ergab sich ein Fit von RMSEA = 0,096 ($RMSEA_{WLSMV}$ = 0,146) und ein CFI von CFI = 0,68 (CFI_{WLSMV} = 0,74). Um weiter sicherzustellen, dass es sich nicht um ein Artefakt des Skalen- (und innerhalb der Skala schwierigkeitsbasierten) Parcelings handelt, wurden 38 neue Parcel gebildet, die nach

dem selben Prinzip wie jene gemäß Tabelle 32 zustande kamen, jedoch auf der Annahme basierten, es gäbe nur eine Skala. Auch in diesem Fall erreichen der CFI mit CFI = 0,74 (CFI_{WLSMV} = 0,74) ebenso wie der RMSEA mit RMSEA = 0,083 ($RMSEA_{WLSMV}$ = 0,117) Werte die klar gegen eine einfaktorielle Lösung ohne skalenspezifische Komponenten sprechen.

8.3.2.3 Modell mit drei Inhaltsfaktoren

Wie bereits mehrfach erwähnt, sind Modelle mit 3 Faktoren, also einem gemeinsamen Faktor für die Skalen prozedurales- und komplexes Rechnen, als auch Modelle mit 4 Faktoren denkbar. Die Lösung mit drei oliquen Faktoren, basierend auf den bereits dargestellten Parceln, lieferte einen CFI von 0,78 (CFI_{WLSMV} = 0,87) und einen RMSEA von 0,079 ($RMSEA_{WLSMV}$ = 0,096). Die Interkorrelationen der Faktoren (ML-Schätzung) reichten von r = 0,47 (Geometrie und grafische Fkt. mit mathematischer Literalität) über r = 0,64 (Geometrie und grafische Fkt. mit prozeduralem-/komplexem Rechnen) bis hin zu r = 0,72 (prozedurales/komplexes Rechnen mit mathematischer Literalität). Auffällig im Vergleich zu der unter 8.3.1 berichteten Faktorenanalyse ist die deutlich höhere Korrelation zwischen prozeduralem/komplexem Rechnen und mathematischer Literalität, ansonsten ist festzustellen, dass beide Fit Indizes eher schlechte Werte aufweisen, was gegen eine Gültigkeit des Modells spricht. Genauso verhält es sich mit einer Entscheidung nach χ^2 / df, die hier mit $\chi^2 / df = 6680 / 626 = 10,7$ ebenfalls gegen das Modell spricht.

Die Faktorladungen der Parcel zeigen ein durchweg unauffälliges Bild, sind alle positiv und variieren je nach Faktor und Parcel zwischen 0,16 und 0,71 (siehe Anhang 12.4 der Arbeit).

8.3.2.4 Modell mit 4 Inhaltsfaktoren

Ein Modell mit den vier, entsprechend ursprünglicher Theorie postulierten, Faktoren weist - trotz der Schwierigkeiten eine Trennung von prozeduralem- und komplexem Rechnen via obliquer Faktorenanalyse zu erreichen - einen besseren Fit auf, mit einem CFI von 0,80 (CFI_{WLSMV} = 0,88) und einem RMSEA von RMSEA = 0,076 ($RMSEA_{WLSMV}$ = 0,092).

Das 3-Faktormodell stellt ein Nested-Modell (Loehlin, 2004, S. 64) des 4-Faktormodells dar. Fixiert man die Interkorrelation der Faktoren komplexes- und

prozedurales Rechnen auf 1 und setzt man die Korrelation eines der Faktoren zu den restlichen beiden (*Geometrie und grafische Fkt*. und mathematischer Literalität) gleich 0 so entsprechen die Modelle einander. Daher ist es möglich einen χ^2 Differenztest (Kline, 2005; Loehlin, 2004) zu berechnen. Es ergib sich hierfür ein $\chi^2_{\text{Diff}} = 6680 - 6227 = 453$ mit $df_{\text{Diff}} = 626 - 623 = 3$ Freiheitsgraden, also ein hochsignifikanter Unterschied zugunsten der 4-Faktorlösung. Aus diesem Grund werden hier nur die Pfadkoeffizienten für die 4-Faktorversion dargestellt. Sie befinden sich zusammen mit der Skalenzugehörigkeit und dem Standardfehler in der folgenden
Tabelle 34.

Tabelle 34 Standardisierte Pfadkoeffizienten der 4 Faktorlösung.

Parcel Nummer	Geometrie und grafische Fkt.	Mathematische Literalität	Prozedurales Rechnen	Komplexes Rechnen
1	0,16	0,40	0,29	0,59
2	0,18	0,44	0,23	0,52
3	0,23	0,51	0,29	0,60
4	0,32	0,36	0,25	0,61
5	0,33	0,49	0,26	0,63
6	0,57	0,38	0,52	
7	0,70	0,46	0,33	
8	0,71		0,55	
9	0,50		0,48	
10	0,30		0,54	
11			0,45	
12			0,39	
13			0,42	
14			0,52	
15			0,49	

Anmerkung. Alle Parameter sind hochsignifikant ($p < 0,00$) von 0 verschieden. Die Zusammensetzung der Parcel lässt sich Tabelle 32 entnehmen. Schätzmethode ML.

Die Interkorrelation der 4 Faktoren schließlich ist Tabelle 35 zu entnehmen. Es ist ersichtlich, dass alle Interkorrelationen (sowohl bei ML als auch bei WLSMV) recht hoch ausfallen.

Tabelle 35 Interkorrelationen der 4 Faktoren.

	Geometrie und grafische Fkt.	Prozedurales Rechnen	Komplexes Rechnen
Prozedurales Rechnen	0,63 (0,72)		
Komplexes Rechnen	0,60 (0,73)	0,85 (0,84)	
Mathematische Literalität	0,47 (0,58)	0,69 (0,68)	0,69 (0,71)

Anmerkung. Alle Korrelationen sind hochsignifikant ($p < 0,00$). In Klammern: Korrelation bei Verwendung der WLSMV-Schätzmethode, sonst ML.

Ein Nebeneffekt der WLSMV-Methode scheint zu sein, dass die Korrelationen tendenziell gleich groß oder größer ausfallen. Der Fit des 4 Faktormodells ist zwar besser als jener des Einfaktormodells, jedoch nach wie vor verbesserungswürdig; vor allem auffällig sind die nach wie vor hohen Korrelationen zwischen den 4 Faktoren. Darüber hinaus finden sich die beiden einzigen Ladungen kleiner 0,20 bei den Geometrieparceln. Auf Basis dieser Erkenntnisse kann Hypothese H11 nicht angenommen werden.

8.4 Taxonomische Passung der Endform

Dieser Abschnitt widmet sich der Prüfung der Hypothesen H13 bis H14, d.h. es wird geprüft für welche Stufen wie viele Zuordnungen getroffen wurden (H13, Abschnitt 8.4.3.3), ob die Zuordnungen zwischen Ratern reliabel sind (H14, Abschnitt 8.4.3.4) und ob Passung zwischen dem was der Test aus Lehrersicht erfasst und dem was Lehrer als wichtig ansehen besteht (H15, Abschnitt 8.4.3.5).

8.4.1 Rekrutierung

Zur Probandengewinnung wurden ca. 50 Berufs- und Realschulen im Raum Mannheim, Ludwigshafen und Worms antelefoniert und um eine Kooperation im Rahmen einer Expertenbefragung zur Einschätzung der Aufgaben eines Mathetests durch Mathematiklehrer gebeten. Sämtliche Termine wurden individuell vereinbart. Eine Entlohnung fand nicht statt.

8.4.2 Durchführung

Bei Crone-Todd, Pear und Read (2000) zeigte sich, dass die Interrater-Reliabilität bei Einordnung von Testaufgaben durch Vorgabe von sukzessiv optimiertem Begleitmaterial (ein Flow-Chart und eine Tabelle) zur Taxonomie nach Bloom et al. (1956) deutlich erhöht werden konnte. Daher wurde die bereits in Abschnitt 3.2.2.1 vorgestellte Taxonomietabelle erweitert und allen Testanden vorgelegt. Darüber hinaus wurde zu jeder einzelnen Testaufgabe die Lösung samt einfachstem Lösungsweg angegeben.

Sämtliche Antworten der Probanden erfolgten auf einem speziellen Antwortblatt. Es wurde gebeten die Taxonomiestufe für jede Aufgabe zu notieren, die zur Lösung notwendig ist. Dabei sollte davon ausgegangen werden, dass ein typischer Berufsanfänger ab ca. 16 Jahren die Aufgaben vorgelegt bekäme. Dieser Berufsanfänger sollte die Möglichkeit gehabt haben, laut Lehrplan, die Lösungsprinzipien (z.B. Satz des Pythagoras) gelernt zu haben.

Im daran anschließenden Teil der Erhebung wurden die Teilnehmer gebeten auf einer 6-stufigen Likert-Skala mit den Polen *trifft überhaupt nicht zu – trifft voll und ganz zu* einzuschätzen, inwiefern sie der Meinung sind, dass jede der 6 Stufen für einen Berufseinsteiger wichtig ist. Alle Antworten erfolgten schriftlich und ohne Angabe von Name und Anschrift auf einem Antwortbogen.

8.4.3 Ergebnisse

Ein Aspekt der Auswertung, der vorab dargestellt werden muss, ist die Schwierigkeit der Integration von Ratings zu Gesamtscores. Damit ist gemeint, dass es eigentlich keinen Sinn ergibt so etwas wie ein mittleres Rating zu berechnen. Denn was sollte ein mittleres Rating eines Raters über alle Items von z.B. 2,37 bedeuten? Dieses Rating kann auf unterschiedlichste Art und Weise zustande gekommen sein. Durch mehrere Ratings für die erste und vierte Stufe genauso wie durch Ratings in der 2 und 3 Kategorie. Dies wäre nur sinnvoll, wenn man die Stufen als Indikatoren zur Messung von Komplexität ansieht und zudem noch davon ausgeht, dass der Abstand zwischen z.B. Evaluieren und Kreieren jenem zwischen Erinnern und Verstehen entspräche, wofür es weder eine Theorie noch eine empirische Rechtfertigung gibt.

Im folgenden werden zwecks Auswertung mittlere Ratings nur verwendet um zu prüfen, ob die Ratings der Lehrer etwas anderes abbilden als die Schwierigkeit der Items. Dies

ist nicht zu verwechseln mit der mittleren Häufigkeit mit der eine Zuordnung zu Kategorie x und Rater y getroffen wird, was in Abschnitt 8.4.3.5 eine Rolle spielt.

8.4.3.1 Stichprobe

Insgesamt nahmen 17 Lehrer an der Befragung teil. Tabelle 36 listet die wichtigsten Kennwerte der Lehrereinschätzungen auf. Es ist auf den ersten Blick ersichtlich, dass die Kategorie 6, kreieren, nur bei zwei von 77 Aufgaben (bei a26c zweimal, bei a33 einmal) als angemessen angesehen wird.

Tabelle 36 Minimal- und Maximalwert, Mittelwert, Streuung und Anzahl der Einschätzung aller Aufgaben durch Realschullehrer auf einer Skala von 1 = erinnern bis 6 = kreieren.

	N	Minimum	Maximum	Mittelwert	SD
A1	17	1	3	1,59	0,80
A2	17	1	3	2,41	0,80
A3	17	1	4	2,53	0,87
A4	17	1	3	2,35	0,86
A5	17	1	4	1,88	0,99
A6a	17	1	4	2,71	0,92
A6b	17	1	4	2,76	0,90
A6c	17	1	4	2,76	0,90
A7	17	1	4	2,94	0,75
A8	16	2	4	3,06	0,77
A9	17	1	3	2,53	0,72
A10a	17	1	4	2,71	0,92
A10b	17	2	4	2,88	0,78
A10c	17	2	4	2,88	0,78
A11a	17	1	5	2,71	1,26
A11b	17	1	5	2,71	1,26
A11c	17	1	5	2,71	1,10
A11d	17	1	5	2,71	1,10
A12a	17	1	5	2,29	1,40
A12b	17	1	5	2,29	1,40
a13a	17	1	3	2,47	0,80
a13b	17	1	3	2,53	0,80
a13c	17	1	3	2,53	0,80
A14a	17	1	3	2,06	0,90
A14b	17	1	3	2,06	0,90
A14c	17	1	3	2,06	0,90
A15a	17	1	3	2,35	0,70
A15b	17	1	3	2,35	0,70
A15c	17	1	3	2,35	0,70
A15d	17	1	3	2,41	0,62
A16a	17	1	4	2,47	0,72
A16b	17	1	4	2,47	0,72
A16c	17	1	4	2,47	0,72

Tabelle 36 Fortsetzung.

	N	Minimum	Maximum	Mittelwert	SD
A17a	17	1	3	2,47	0,80
A17b	17	1	3	2,41	0,80
A18	17	1	4	1,71	1,10
A19	17	1	4	2,06	1,34
a20a	16	1	3	2,50	0,63
a20b	16	1	4	2,56	0,73
A21a	16	1	4	2,63	0,72
A21b	16	2	4	2,88	0,50
A22a	17	1	3	2,24	0,75
A22b	17	1	3	2,35	0,79
A22c	17	1	3	2,35	0,79
A23a	16	1	5	2,63	0,89
A23b	16	1	5	2,69	0,87
A23c	16	2	5	2,75	0,77
A24a	16	1	4	2,75	1,06
A24b	16	1	4	2,75	1,06
A24c	16	1	4	2,75	1,06
A24d	16	1	4	2,88	0,96
A25b	17	2	5	3,47	0,94
A25c	17	2	5	3,53	1,01
A25d	16	2	5	3,44	0,96
A26a	17	1	5	2,41	1,33
A26b	17	2	5	2,82	1,01
A26c	17	2	6	3,41	1,00
A27a	17	2	4	3,24	0,83
A27b	17	3	4	3,35	0,49
A27c	17	2	5	3,47	0,72
A27d	17	2	5	3,41	0,94
A27e	17	2	5	3,59	0,71
A28	17	2	5	3,94	0,75
A29	17	1	5	3,35	1,27
A30a	17	1	5	3,71	1,10
A30b	17	2	5	3,65	1,00
A31a	17	1	4	2,76	0,66
A31b	17	1	4	2,82	0,64
A31c	17	1	4	2,82	0,64
A31d	17	1	4	2,82	0,64
a32a	17	2	4	2,82	0,64
a32b	17	1	4	2,82	0,73
A33	17	1	6	3,65	1,54
A34a	17	1	4	2,71	0,77
A34b	17	1	4	2,65	0,70
A34c	17	2	4	3,18	0,73
A35	17	1	3	2,76	0,56

Insgesamt weisen Items einer Aufgabengruppe (z.B. A10a, b, c) praktisch immer sehr ähnliche mittlere Einschätzungen auf und werden anscheinend, trotz nicht vorhandener direkter Abhängigkeit der Aufgaben voneinander, als zusammengehörig empfunden.

8.4.3.2 Zusammenhang von Einschätzung und Itemschwierigkeit

Eine Spearman-Rang Korrelation zwischen mittlerem Stufenlevel und der Schwierigkeit des Items in der Normstichprobe ergibt ein *r* von -0,23 (p = 0,02, einseitig, *N* = 77). Das heißt mit steigendem Taxonomielevel sinkt der Schwierigkeitsindex p, d.h. die Aufgabe wird schwieriger. Der gefundene Zusammenhang ist eher gering, was zum einen in der teils geringen Varianz der Lehrerratings begründet ist und zum anderen mit der Tendenz Aufgaben einer Aufgabengruppe sehr ähnlich zu beurteilen zusammenhängen könnte.

Daher wurden die Items der Tabelle 36 auf Aufgabenlevel gemittelt, auf Seite der Taxonomieeinschätzung ebenso wie auf Ebene der Testantworten in der Gesamtstichprobe. Die Spearman Rangkorrelation steigt dadurch etwas an auf *r* = -0,33 (p = 0,06, einseitig, *N* = 35).

8.4.3.3 Bedeutung der 6 Taxonomiestufen

Abbildung 23 vermittelt einen Eindruck davon, welche der kognitiven Prozesse die 17 Realschullehrer als besonders wichtig, oder eher unwichtig für den typischen Berufsanfänger ansehen und ist nicht zu verwechseln mit der Einschätzung der Items des Mathetests.

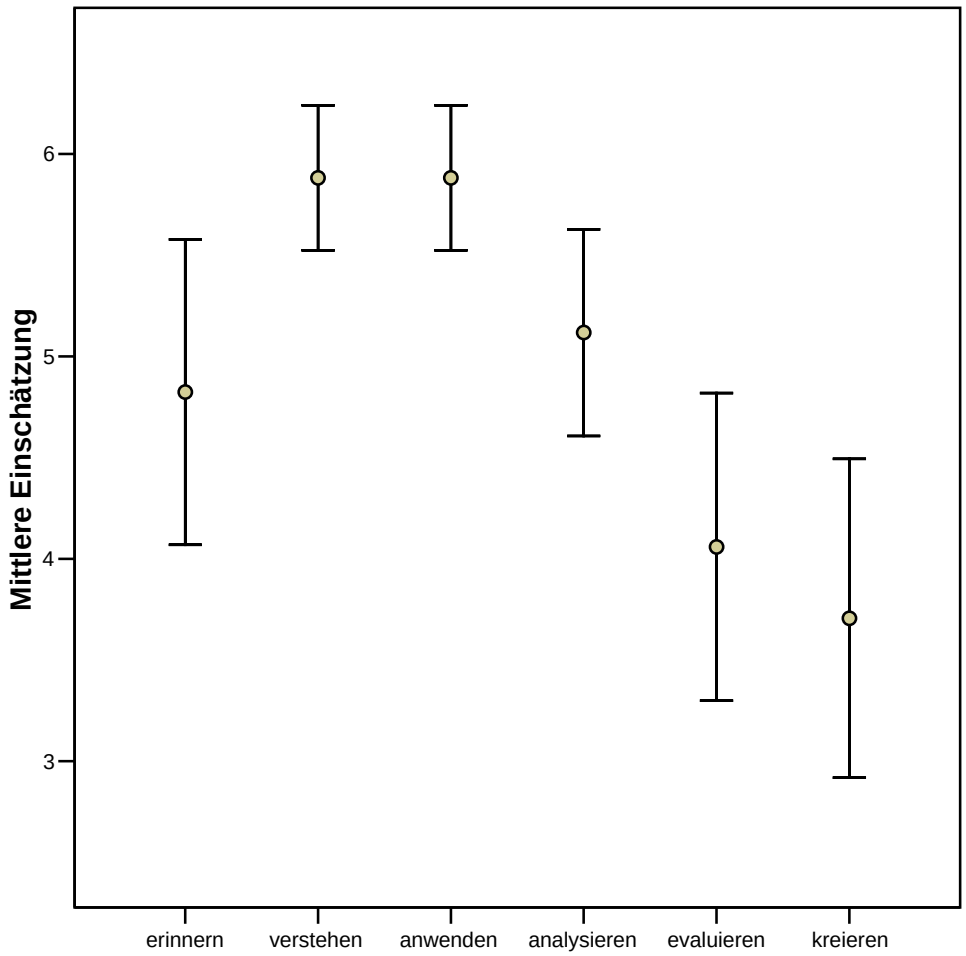


Abbildung 23 Mittelwerte und Konfidenzintervalle (95%) zur Einschätzung der Wichtigkeit der 6 kognitiven Prozesse nach Anderson und Krathwohl (2001) von 17 Realschullehrern.

Aufgrund der geringen Stichprobengröße von $N = 17$ überlappen sich die Konfidenzintervalle der meisten Mittelwerte. Dennoch sticht deutlich heraus, dass die Lehrer verstehen und anwenden für den typischen Berufsanfänger mit Realschulabschluss als besonders wichtig empfinden. Auch ist deutlich zu erkennen, dass sie Evaluieren und Kreieren als weniger bedeutsam ansehen.

8.4.3.4 Rater-Übereinstimmung

Bereits die Berechnung der Spearman Rangkorrelation in Abschnitt 8.4.3.2 kann wegen der Berechnung von einem mittleren Stufenlevel sehr kritisch hinterfragt werden. Dort ging es jedoch vor allem darum, ob die Ratings der Lehrer überhaupt etwas anderes als die wahrgenommene Schwierigkeit der Items darstellten. Dies konnte durch das Vorgehen weitestgehend ausgeschlossen werden.

Nun stellt sich die Frage, welches Maß zur Bestimmung der Interrater-Reliabilität zum Einsatz kommen soll. Im wesentlichen ist hier zu unterscheiden zwischen Maßen für nominalskalierte Kategorien wie Cohens Kappa (Cohen et al., 2003), Maßen für ordinal geordnete Kategorien wie Kendalls W (als Verallgemeinerung von Spearmans rho für mehrere Rater) (Siegel, 1956) und Ansätzen für intervallskalierte Maße wie die Intra-Klassen-Korrelation (ICC) (Shrout & Fleiss, 1979). Die Konzeption der Taxonomiestufen nach Anderson und Krathwohl (2001) legt nahe, eine zumindest ordinale Ordnung anzunehmen. Ferner berichten Wirtz und Caspar (2002), dass die ICC relativ robust gegenüber Verletzungen ihrer Annahmen ist. Dies ist insbesondere deshalb interessant, weil zum einen allein aufgrund der Skalenbeschränkung (1-6) keine allzu extremen Ausreißer möglich sind und zum anderen eine ICC die Möglichkeit beinhaltet, die absoluten Ratings als Maß zur Berechnung der Abweichung einzubeziehen (Nichols, 1998). Damit ist gemeint, dass hier die allgemeine Tendenz eines Raters zu strengen oder milden Urteilen nicht herausgerechnet wird. Würde ein Rater A z.B. stets die Kategorien 3 und 4 wählen, ein anderer Rater B bei diesen Items übereinstimmend stets die Kategorien 4 und 5, so wäre dies eine allgemeine Tendenz zu höheren Urteilen von Rater B. Dieser Effekt soll hier bewusst nicht aus der ICC herausgerechnet werden. Demnach ergibt sich für die 17 Rater und 77 Items ein mittlerer ICC Wert von ICC = 0,82 was als guter Wert betrachtet werden kann (Wirtz und Caspar, 2002), jedoch auch deutlich von der relativ hohen Anzahl der Rater abhängt. Dieser Wert beschreibt die Genauigkeit eines mittleren Raters analog dem Vorgehen mittels Spearman-Brown Formel bei Testverlängerungen (Wirtz & Caspar, 2002).

8.4.3.5 Taxonomielevel des Mathematiktests

Die Frage nach dem Taxonomielevel des Mathematiktests ist vor allem deswegen interessant weil sie, mit dem was die Lehrer für besonders wichtig halten, verglichen werden kann. Letztlich wurden von 17 Ratern für 77 Items Angaben gemacht (einige wenige missings ausgenommen, vgl. Tabelle 36). Die mittlere Häufigkeit, also am Beispiel der Stufe 1 erinnern, lautet $\overline{x}_{\text{Stufe 1}} = \frac{1}{17}\sum_{i=1}^{17}\sum_{k=1}^{77} x_{ik}$, mit $x_{ik} = 1$ falls Rater i bei Item k die Taxonomiestufe 1 wählte, sonst $x_{ik} = 0$ und ist für alle 6 Stufen in Abbildung 24 abgetragen. Diese Werte geben an, wie häufig der durchschnittliche Lehrer bei allen 77 Items eine bestimmte Kategorie wählte.

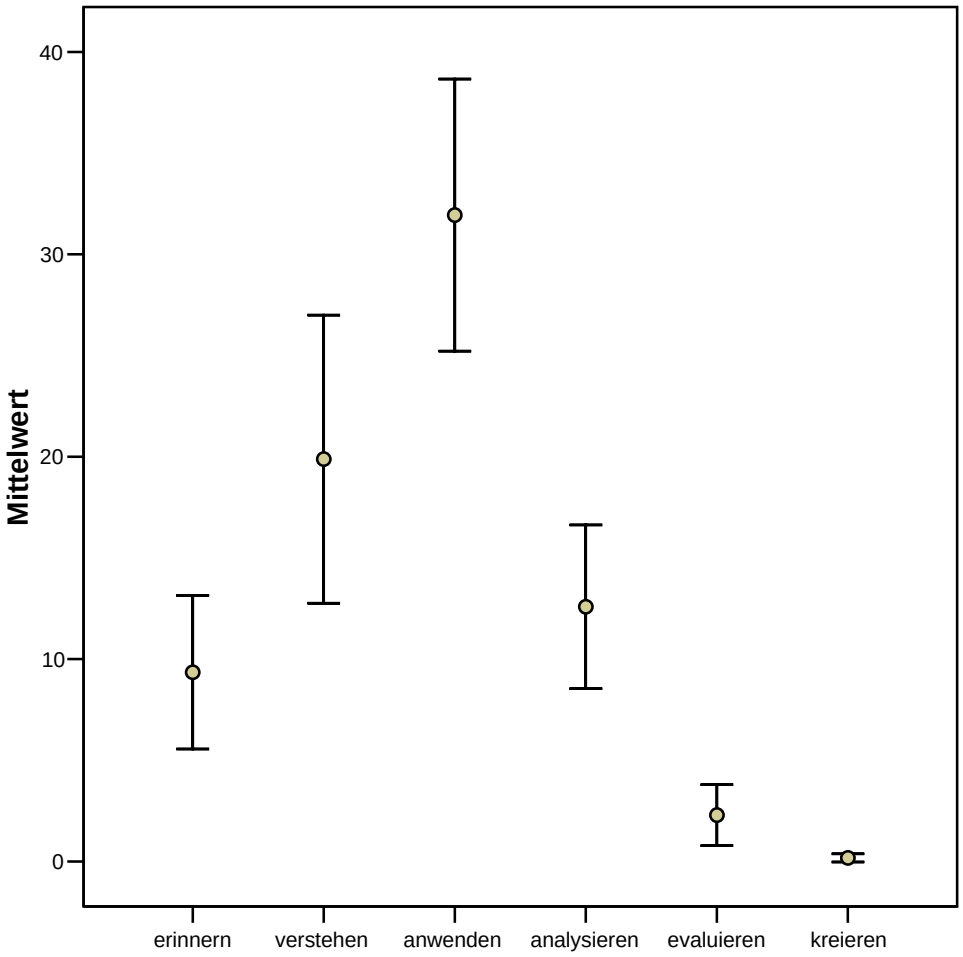


Abbildung 24 Mittlere Anzahl von Ratings für eine der 6 Taxonomiestufen einschließlich Standardfehler (95%), die Mittelwerte summieren sich zur Anzahl der Items (77).

Schlusslicht dieses Vergleichs bilden klar die Stufen evaluieren und kreieren, mit $\overline{x}_{kreieren} = 0,18$ und $\overline{x}_{evaluieren} = 2,3$. Da maximal ein Wert von 77 erreicht werden kann (=Anzahl der Items) scheint die Stufe kreieren praktisch bedeutungslos zu sein. Über die genaue Bedeutung von evaluieren ließe sich sicherlich streiten, jedoch ist sie verglichen mit den verbliebenen 4 Stufen klar untergeordnet. Die im Mittel am nächst häufigsten gewählte Stufe, wurde mehr als 4 mal so häufig gewählt ($\overline{x}_{erinnern} = 9,4$). Interessant ist, dass eine klare Passung bezüglich Reihung der Stufen nach Wichtigkeit in Abschnitt 8.4.3.3 und der Häufigkeit des Urteils im Mathetest besteht. Das heißt die Lehrer fanden in der Befragung Verstehen und Anwenden für einen Berufseinsteiger am wichtigsten und trafen die häufigsten Zuordnungen von Items zu eben diesen beiden Kategorien. Dies ist ein klarer Hinweis darauf ist, dass der Mathetest zu einem wesentlichen Anteil jene Bereiche am stärksten (mit am meisten Items) erfasst, die als am wichtigsten für den typischen Berufsanfänger angesehen werden. Dass Evaluieren

und Kreieren aus Lehrersicht mit dem Test praktisch überhaupt nicht erfasst werden, verliert an Dramatik, da Lehrer diese Bereiche auch als am wenigsten wichtig ansehen.

8.5 Schlussfolgerung

Am Ende dieses Abschnitts gilt es eine Bilanz zu ziehen: Konnten die Hypothesen 10 bis 14 (Abschnitt 8.1) bestätigt werden?

H10 konnte sicherlich bestätigt werden, der Test scheint eindeutig nicht eindimensional zu sein. In Bezug auf H11 hieß es im Anschluss an die Tests mit DIMTEST und DETECT (8.1.1), dass noch weitere Analysen auf Parcel und Itemebene notwendig wären um die Frage nach ausreichender Passung der Daten mit dem aufgestellten Modell zu beantworten. Nachdem dies durchgeführt wurde, kann im Sinne einer konservativen Herangehensweise Hypothese 11 nicht angenommen werden. Der Gesamtfit des 3- und 4 Faktormodells ist über alle verwendeten Methoden hinweg nicht ausreichend. Im weiterführenden Teil dieser Arbeit (Abschnitt 9) wird daher versucht werden ein Modell zu postulieren, das sowohl eigenständige Gruppenfaktoren (den 4 Skalen entsprechend) als auch einen G-Faktor enthält. Dies erscheint notwendig, da für sich genommen weder ein G-Faktor, noch ein Modell korrelierter Gruppenfaktoren einen zufrieden stellenden Modellfit erbrachten.

In Bezug auf die taxonomische Ordnung ist festzuhalten, dass die 6 Taxonomiestufen in unterschiedlichem Ausmaß im Test enthalten sind (8.4.3.5) und somit Hypothese H12 angenommen werden kann. Was Hypothese H13, die Reliabilität der Zuordnung durch die Rater angeht, so kann sie zwar bestätigt werden, doch ist dies sich auch in der geringen Varianz der Ratings und der großen Anzahl von Ratern (für eine solche Untersuchung) begründet.

Auch Hypothese H14, bei der es um die Passung zwischen dem, was aus Lehrersicht wichtig ist und dem, was der Test erfasst geht, kann angenommen werden. Auch wenn die Ergebnisse nicht repräsentativ sind, kann zumindest bei der vorliegenden Stichprobe davon ausgegangen werden, dass eine ausreichende Passung vorliegt (8.4.3.5).

9 Weiterführende Betrachtungen

Neben dem schon angekündigtem Versuch eine bessere Modellpassung in Bezug auf die 4 Skalen zu erreichen haben sich im Laufe der Arbeit noch einige weiterführende Fragen ergeben, die in diesem Abschnitt geprüft werden sollen.

9.1 Ein Schmid-Leiman Modell

Das so genannte Schmid-Leiman (SL) Modell wurde von seinen Autoren (Schmid & Leiman, 1957) vor über 50 Jahren aufgestellt, um Faktoren höherer Ordnung zu orthogonalisieren. Der Hintergedanke besteht daraus, dass in hierarchischen (konfirmatorischen) Faktorenanalysen die Interpretation der Faktoren häufig Probleme bereitet. Neben hierarchischen Faktormodellen, für die die Transformation ursprünglich gedacht war, ist die Anwendung jedoch auch auf oblique Strukturmodelle wie jenes in Abbildung 25 möglich.

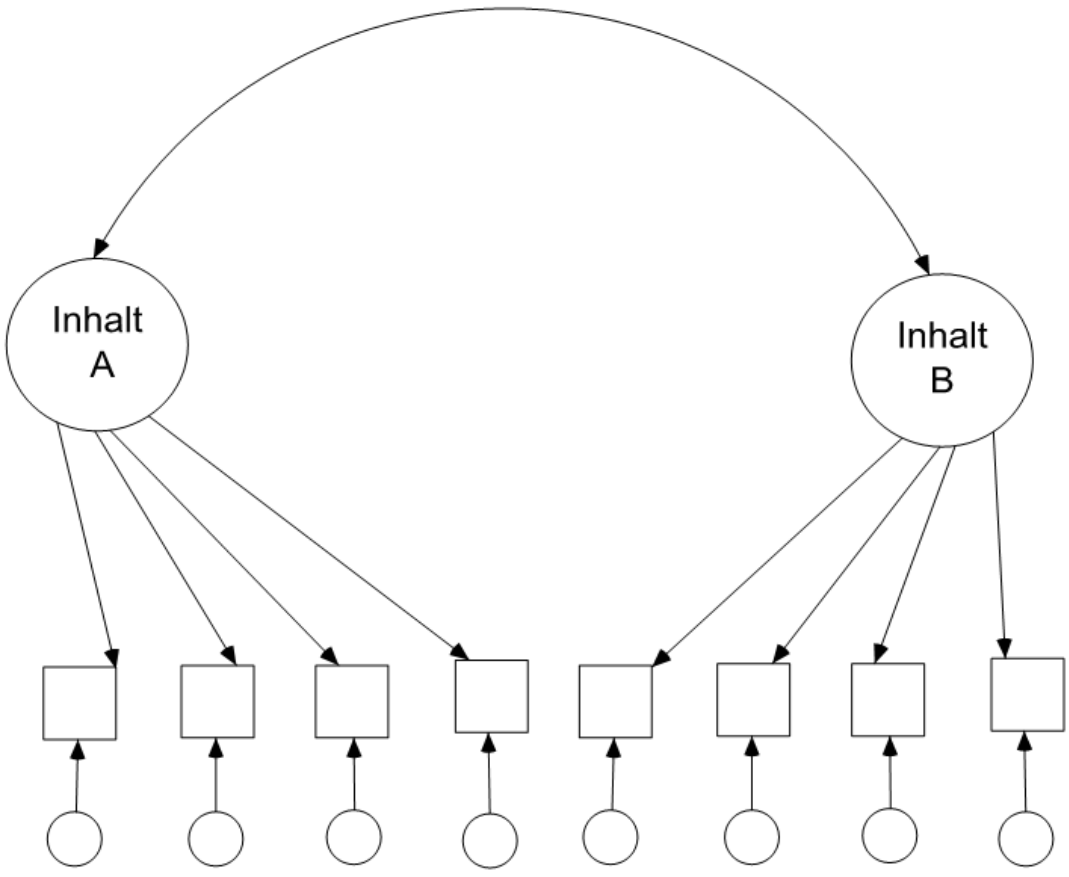


Abbildung 25 Strukturmodell zweier korrelierter Faktoren (Inhalt A und Inhalt B).

Das Hauptproblem an Modellen wie jenem gemäß Abbildung 25 besteht darin, dass die Ladungsmuster schwer interpretierbar sind, da sie stets erstens von dem Zusammenhang des Indikators und des zugehörigen latenten Faktors abhängen und zweitens ebenfalls von den anderen latenten Faktoren abhängen (indirekt). Überführt man ein solches

Modell via Schmid-Leiman Transformation in die Variante gemäß Abbildung 26, so hat dies den Vorteil, dass nun die Faktoren höherer Ordnung (im Beispiel Inhalt A und Inhalt B) orthogonal sind.

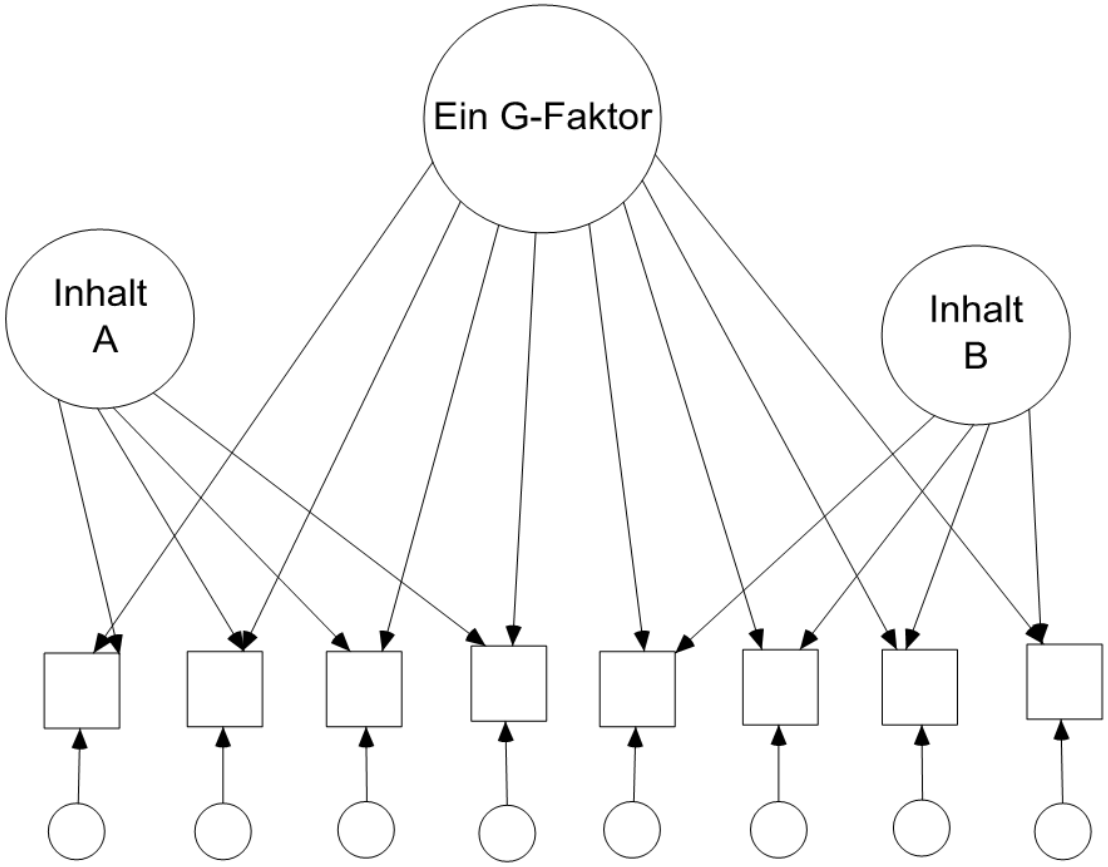


Abbildung 26 Schmid-Leiman Transformation des Modells gemäß Abbildung 25.

Inhaltlich heißt dies, dass zunächst die Varianz sämtlicher Variablen auf einen gemeinsamen G-Faktor zurückgeführt wird und die zwei Faktoren Inhalt A und Inhalt B jene Varianz aufklären, die darüber hinaus spezifisch für die latenten Konstrukte hinter Inhalt A und Inhalt B sind. Trotz seines Alters ist die SL-Transformation nach wie vor hoch aktuell. Erst vor wenigen Jahren wurde in der Zeitschrift *Behavior Research Methods* ein Artikel veröffentlich, der die Transformation darstellt und Syntax-Codes zur einfachen Durchführung in SPSS und SAS enthält (Wolff & Preising, 2005). Die zahlreichen Anwendungen umfassen beispielsweise Studien zur Struktur des 16PF (Chernyshenko & Stark, 2001), des Wechsler-Intelligenztests für Erwachsene (Rijsdijk, Vernon & Boomsma, 2002) und jüngst des Beck Angst-Inventars (Steer, 2009). Ein weiteres Beispiel stellt die Studie zur Struktur des Berliner-Intelligenz-Strukturmodells und der Integration von fluider und kristalliner Intelligenz in das Modell dar (Beauducel & Kersting, 2002). Da die in Abschnitt 8.3.2 bereits getesteten Modelle erstens einen nur mäßigen Fit aufweisen und zweitens auch dort die eingangs erwähnte

Interpretationsproblematik existiert wird nun ein SL-Modell für 3- und 4 Faktoren angewendet.

9.2 Schmid-Leiman-Modell versus oblique-Modelle

Wie bereits im theoretischen Teil der Arbeit dargelegt, ist eine Trennung der Skalen prozedurales Rechnen und komplexes Rechnen wünschenswert, doch aufgrund der inhaltlichen Ähnlichkeit der beiden Skalen nicht zwingend.
Demnach stehen ebenso wie bei den obliquen Modellen, die in Abschnitt 8.3.2 geprüft wurden, für die SL-transformierten Modelle sowohl Varianten mit drei als auch mit vier Inhaltsfaktoren zur Debatte. Die folgende Tabelle stellt die bereits berechneten obliquen 3- und 4 Faktormodelle, ein G-Faktormodell und die SL-Transformationen gegenüber.

Tabelle 37 Gegenüberstellung von bereits getesteten obliquen-Modellen, einem G-Faktor Modell und zwei Schmid-Leiman Modellen (*N* = 1554).

Modellvariante	χ_{ML}^2 (df)	RMSEA ML	CFI ML	CFI WLSMV	RMSEA WLSMV	AIC
4 Faktoren korreliert	6227 (623)	0,076	0,80	0,88	0,092	96027
3 Faktoren korreliert	6680 (626)	0,079	0,78	0,87	0,096	96474
Generalfaktor	7708 (665)	0,083	0,74	0,74	0,117	98585
Schmid-Leiman Modell mit 3 Faktoren	4173 (592)	0,062	0,87	0,92	0,073	94035
Schmid-Leiman Modell mit 4 Faktoren	3871 (592)	0,060	0,88	0,93	0,070	93733

Anmerkungen. Schätzmethode: Maximum Likelihood (ML), Weighted Least Squares Mean and Variance adjusted (WLSMV), RMSEA: Root Mean Square Error of Approximation, CFI = comparative Fit Index. AIC = Akaikes Information Criterion. In 3-Faktormodellen bilden prozedurales- und komplexes Rechnen einen Faktor.

Gemäß obiger Tabelle wiesen beide Schmid-Leiman Modellvarianten sowohl im parsimonious Fit-Index (RMSEA) als auch im inkrementellen CFI-Index einen deutlich besseren Fit als die obliquen Modelle und das G-Faktormodell auf. Bei dem Modell mit

dem besten Fit handelt es sich um ein SL-Modell mit 4 Faktoren, welches schematisch in der folgenden Abbildung 27 dargestellt ist.

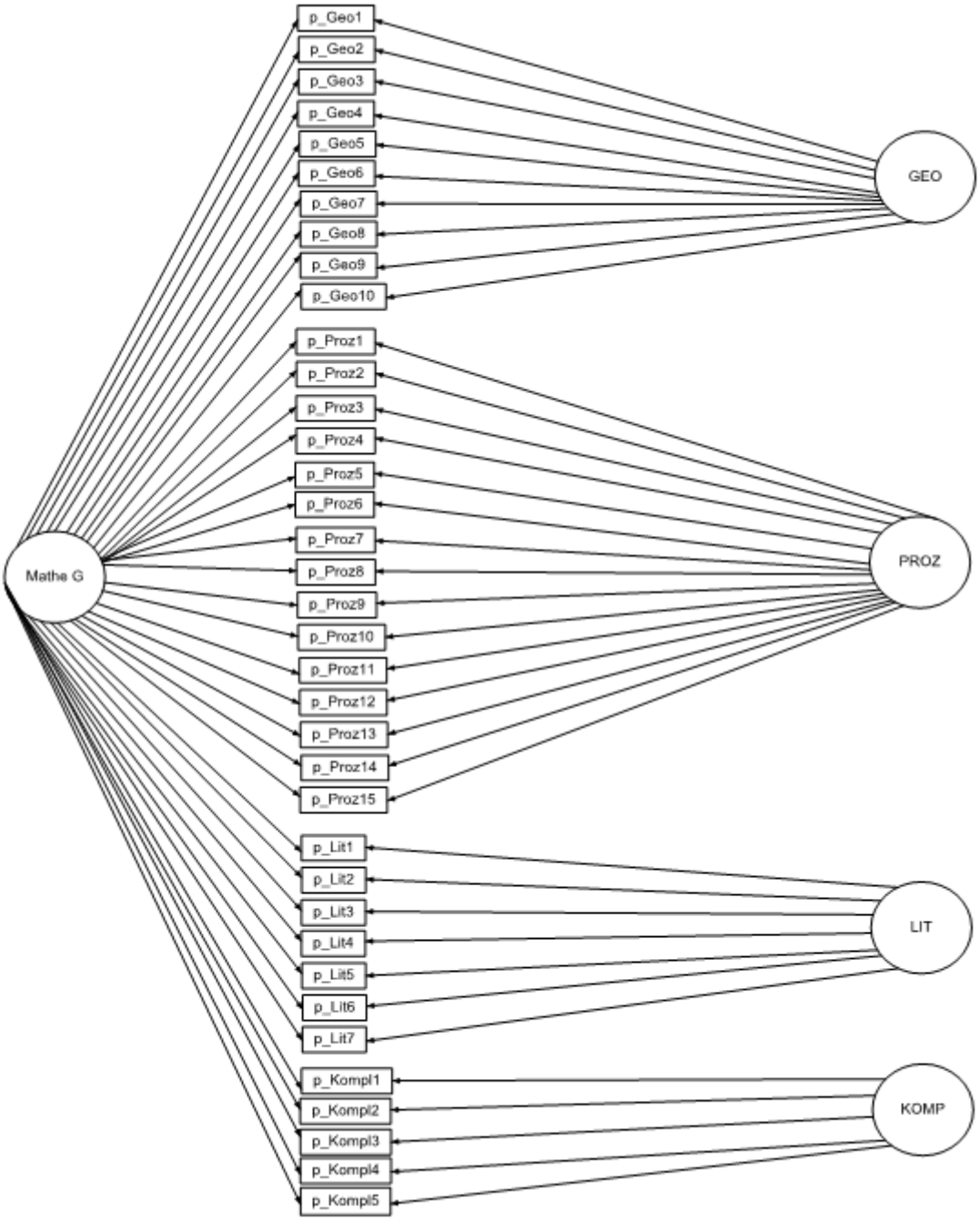


Abbildung 27 Darstellung des, finalen SL–Modells. Jeder manifesten Variable ist ein Messfehler zugeordnet, der aus Platzgründen nicht in der Abbildung aufgeführt ist. LIT = mathematische Literalität, PROZ = prozedurales Rechnen, KOMPL = komplexes Rechnen, GEO = Geometrie und grafische Fkt.

Das 3-Faktor SL-Modell stellt keine *nested* Variante (Loehlin, 2004) des 4 Faktormodells dar. Beide Modelle weisen genau dieselbe Anzahl von Freiheitsgraden auf. Fügt man eine Korrelation in Form eines Doppelpfeils zwischen den latenten Faktoren prozedurales- und komplexes Rechnen ein und fixiert sie auf den Wert 1, so

entspricht das 3-Faktor Modell zwar der 4-Faktorvariante, doch handelt es sich lediglich um alternative Modelle da man einen zuvor auf 0 gesetzten Pfad (die Korrelation der Faktoren prozedurales- und komplexes Rechnen) nun auf 1 fixiert, jedoch keinen vorher frei geschätzten Pfad auf einen beliebigen Wert fixiert.

Handelt es sich bei zwei Modellen um alternative non-nested Modelle, die jedoch genau dieselben Variablen enthalten, bietet sich nach Rust, Lee und Valente (1995) Akaikes Information Kriterium (AIC) an (Akaike, 1973). Es gibt mehrere Formulierungen des AIC, wobei es MPLUS (Múthen & Múthen, 2007) in Form von $AIC = -2Log(Likelihood) + 2df_{rest}$ formuliert. Mit diesem Wert können nun die konkurrierenden 3- und 4-Faktor-SL-Modelle (Tabelle 37) verglichen werden, wobei das Modell mit dem kleineren AIC (ein Nebeneffekt der Logarithmierung) zu bevorzugen ist, es jedoch keine Signifikanzprüfung für diesen Unterschied gibt (Mulaik, 2009, S. 348).

Gemäß diesen Erörterungen weist das 4-Faktorielle SL-Modell den besten Fit auf, da es den niedrigsten AIC liefert, d.h. niedriger als die 3-Faktor SL-Variante und niedriger als die untransformierten 3- und 4-Faktormodelle. Die folgende Tabelle 38 zeigt dementsprechend die standardisierten Faktorladungen der finalen Schmid-Leiman Lösung.

Tabelle 38 Standardisierte Pfadkoeffizienten der 4 Faktor-SL-Lösung der Mathetest-Parcel.

Parcel Nummer	G-Faktor	Mathematische Literalität	Prozedurales Rechnen	Komplexes Rechnen	Geometrie und Grafische Fkt.
1	0,23	0,31	-0,04	0,14	0,00 n.s.
2	0,23	0,27	-0,01 n.s.	0,20	0,03
3	0,28	0,37	-0,07	0,34	0,04
4	0,42	0,19	0,13	0,27	0,03 n.s.
5	0,41	0,36	0,13	0,34	0,05
6	0,43	0,22	0,16		0,36
7	0,41	0,29	0,19		0,59
8	0,41		0,20		0,61
9	0,49		0,23		0,23
10	0,45		0,26		-0,01 n.s.
11	0,32		0,30		
12	0,24		0,10		
13	0,32		0,11		
14	0,21		0,37		
15	0,24		0,23		
16	0,48				
17	0,29				
18	0,50				
19	0,43				
20	0,48				
21	0,39				
22	0,38				
23	0,41				
24	0,44				
25	0,44				
26	0,27				
27	0,34				
28	0,36				
29	0,30				
30	0,34				
31	0,30				
32	0,35				
33	0,57				
34	0,46				
35	0,52				
36	0,54				
37	0,55				

Anmerkung. Zusammensetzung der Parcel siehe Tabelle 32. Schätzmethode: ML. Varianz aller Faktoren = 1. Alle Koeffizienten außer n.s. hochsignifikant ($p < 0,01$).

Es ist ersichtlich, dass alle der 37 Parcels einen signifikanten Anteil an G-Varianz enthalten. Über die Hälfte der Parcels die der Skala Geometrie und grafische

Funktionen zuzuordnen sind werden durch die Transformation bedeutungslos (Ladung < 0,10), im Falle der Skala prozedurales Rechen handelt es sich um 3 Parcel. Eine Lösung anhand des WLSMV-Verfahrens (Múthen & Múthen, 2007) liefert sehr ähnliche Ergebnisse, doch fallen die Ladungen ebenso wie die CFI-Werte größer aus (siehe Anhang 12.5).

9.3 Trennbarkeit der Skalen prozedurales- und komplexes Rechnen

Zwar wies bei allen bisher geprüften Modellen eine Variante mit jeweils einem eigenen Faktor für prozedurales- und komplexes Rechnen einen besseren Fit auf, doch traten ebenso regelmäßig Probleme dabei auf die Skalen zu trennen.

Das einfachste Vorgehen zu prüfen, ob eine bessere Trennung der beiden Skalen prozedurales- und komplexes Rechnen möglich ist, besteht in der Durchführung einer Faktorenanalyse nur dieser beider Skalen, und zwar - im Gegensatz zu bisherigen Analysen – für Personen unterschiedlicher Fähigkeit. Dies ist gewissermaßen eine Teilung der Stichprobe an einem inneren Kriterium und stellt eine Übertragung von Spearmans law of diminishing returns (siehe Abschnitt 4.3.3; Spearman, 1904) auf den Bereich der Mathematik-Diagnostik dar.

Eine weitere Variante ergibt sich, indem man prüft, welche der Skalen des Tests am besten zwischen den Probanden verschiedener Schultypen differenzieren. Dies war einer der ursprünglichen Gedanken bei der theoretischen Konzeption dieser beiden Skalen (siehe Abschnitt 3.1).

9.3.1 Faktorenanalytisch

In der vorliegenden Untersuchung der Normstichprobe ließen sich die Bereiche prozedurales Rechnen und komplexes Rechnen am ehesten für die schlechtere Hälfte der Stichprobe (Gesamtscore < Md) trennen. Die Ergebnisse einer Faktorenanalyse, welche sich nur auf diese beiden Skalen bezieht ist Tabelle 39 zu entnehmen.

Tabelle 39 Pattern Matrix der Mathetest-Parcels für die schlechtere Hälfte der Stichprobe (Gesamtscore < 39, *N* = 787).

Parcel	Faktor 1	Faktor 2
Proz1	**0,28**	
Proz2	**0,23**	
Proz3	**0,29**	
Proz4	**0,73**	
Proz5	**0,75**	
Proz6	**0,71**	
Proz7	**0,32**	
Proz8	**0,59**	
Proz9		**0,41**
Proz10		**0,52**
Proz11	**0,36**	0,22
Proz12		**0,41**
Proz13	**0,24**	
Proz14	**0,44**	0,28
Proz15	0,27	**0,38**
Kompl1		**0,48**
Kompl2		**0,59**
Kompl3		**0,51**
Kompl4		**0,56**
Kompl5		**0,50**

Anmerkung. Rotation: Oblimin (Gamma=0). Varianzaufklärung: Faktor 1 = 18%, Faktor 2 = 9%. Interfaktorkorrelationen: $r_{F1\ F2} = 0{,}28$ Ladungen kleiner 0,20 wurden ausgeblendet. Die jeweils höchste Ladung eines Parcels ist hervorgehoben.

Die Trennbarkeit der Skalen ist demnach zumindest für die schlechtere Hälfte der Stichprobe ansatzweise möglich. Hervorzuheben ist in diesem Zusammenhang auch die eher moderate Korrelation der beiden Faktoren von $r = 0{,}28$.

9.3.2 Diskrininanzanalyse

In Abschnitt 3.1.6, wo die 4 Skalen des Mathetests umschrieben wurden, gab es bereits den Hinweis, dass insbesondere für die Skala komplexes Rechnen ein deutlicher Unterschied zwischen verschiedenen Klassenstufen und Schultypen zu erwarten ist. Die Diskriminanzanalyse (Cohen et al., 2003) soll nun genutzt werden, um zu prüfen anhand welcher Skalen sich Personen mit Abitur (einschließlich Fachabitur) und ohne Abitur am Besten trennen lassen. Dazu wird als erster Prädiktor komplexes Rechnen aufgenommen, als zweiter Prädiktor prozedurales Rechnen da er teils ähnliche Inhalte (numerische) enthält, gefolgt von mathematischer Literalität und Geometrie und grafischen Fkt.. Letzterer wird erst am Ende aufgenommen, da er im Gegensatz zu

mathematischer Literalität relativ sprachfrei ist und vermutlich der Einfluss des Schulsystems auf diese – an figurale Intelligenz angelehnte – Komponente des Tests geringer ist.

Bei der Diskriminanzanalyse wird versucht, gleichzeitig die Varianz zwischen den zu trennenden Gruppen zu maximieren und die Varianz innerhalb der Gruppen möglichst gering zu halten (Backhaus et al., 2006). Im Falle einer Gruppe mit nur zwei Ausprägungen entspricht die erste kanonische Korrelation der Diskriminanzanalyse dem *R* der multiplen Regression. Der wesentliche Vorteil dieses Verfahrens im Gegensatz zu einem Vergleich der einfachen Mittelwerte der beiden Gruppen auf den 4 Skalen besteht darin, dass komfortabel geprüft werden kann, ob durch Hinzufügen der anderen Skalen relevante Verbesserungen in der Trennung der Gruppen erreichbar sind. Die Ergebnisse der hierarchischen Diskriminanzanalyse sind in Tabelle 40 abgetragen.

Tabelle 40 Trennbarkeit von Personen mit und ohne Abitur anhand hierachischer Diskriminanzanalyse.

Schritt	Reihenfolge	Wilk‘s Lambda	R	R^2	F	df_A	df_2
1	KOMPL	0,74	0,515	0,265225			
2	KOMPL, PROZ	0,73	0,52	0,2704	13,94**	1	1471
3	KOMPL, PROZ, LIT	0,73	0,525	0,275625	2,79**	2	1469
4	KOMPL, PROZ, LIT, GEO	0,72	0,526	0,276676			
1	PROZ	0,81	0,437	0,191			
1	LIT	0,86	0,381	0,145			
1	GEO	0,83	0,411	0,169			

Anmerkung. Bei den Prädiktoren handelt es sich jeweils um die Summenwerte der Skalen. Bei Schritt 4 ist Validitätszuwachs so gering, dass kein F berechenbar (Division durch 0). $P < 0,01$**. KOMPL = komplexes Rechnen, PROZ = prozedurales Rechnen, GEO = Geometrie und grafische Fkt., LIT = mathematische Literalität.

Es zeigt sich, dass - nimmt man jede Skala für sich - tatsächlich komplexes Rechnen die beste Trennung der Gruppen ermöglicht. Durch Hinzufügen der restlichen 3 Skalen in dargestellter, plausibler Reihenfolge lässt sich praktisch keine Verbesserung mehr erzielen. Berechnet man hierarchische F-Tests (Cohen et al., 2003), um die zusätzliche Aufklärung auf Signifikanz zu prüfen, ergibt sich in zwei Fällen zwar ein signifikanter Unterschied, doch sind die Verbesserungen dermaßen minimal, dass man sie als weitestgehend bedeutungslos und nur durch die große Personenstichprobe hervorgerufen, bezeichnen kann. Die Skala komplexes Rechnen eignet sich demnach vor allem für die Konstruktion einer leichten und schweren Kurzform des Tests. Dieser

Aspekt soll hier jedoch nicht weiter vertieft werden und wird an anderer Stelle (Jasper & Wagener, in Druck) behandelt.

9.4 Geschlechterunterscheide

Wie Wittmann (2004) zusammenfasst, stellen Gruppenunterschiede im Bereich der Intelligenz und verwandten Messungen ein zugleich schwieriges aber auch wichtiges Thema dar. So heißt es dort: „Group differences is a most controversial topic in psychology and social sciences, in which a researcher can easily fall into booby traps, ruin or endager his or her academic career, or at least get a finger burnt." (Wittmann, 2004, S. 223).

Bei einer Analyse des schwedischen SAT kamen er und Kollegen unter anderem zu dem Schluss, dass es in dem anscheinend zweidimensionalen Test (verbale und reasoning-Komponenten) Profilunterschiede zwischen den Geschlechtern gibt, wobei der nonverbale Teil Männer bevorzugen könnte (Fremer, Lohman & Wittmann, 2002). Dies ist jedoch nur die halbe Wahrheit, so zeigte sich sowohl beim schwedischen SAT als auch bei einer Reanalyse von PISA-Daten (Fremer et al., 2002; Wittmann, 2004), dass viele Frauen mit hohes Scores auf den nonverbalen Testteilen existieren, was die Frage nach dem Grund ihrer Unterrepräsentierung im Bildungs- und Arbeitssystem aufwirft. In der mit anderem Fokus bereits dargestellten TIMSS 2007 Untersuchung (Abschnitt 2.2.1) variierten, bei einem Mittelwert von 500 und einer Standardabweichung von 100, die Differenzen zwischen Mädchen und Jungen (8. Klasse) je nach Land zwischen 0 (Malta) und 54 Punkten (Oman) (IEA, 2008, S. 59). Daraus resultiert für 2007 ein mittlerer Unterschied über alle Länder von 5 Punkten zu Gunsten der Mädchen, der zwar signifikant ist, jedoch in Relation zur Standardabweichung von 100 sehr gering ausfällt. Auch wenn man Länder betrachtet, die Deutschland ähnlich sind, wie z.B. Schweden (4 Punkte zugunsten der Frauen, n.s.) ergibt sich kein großer Unterschied. Etwas deutlicher - und in die andere Richtung - geht der Unterschied im PISA 2003-Mathematikteil für Mädchen und Jungen in Deutschland mit 15 Punkten zugunsten der Jungen (OECD, 2004). Dies entspricht jedoch - gemäß PISA Skalierung – nur 1,5 Zehntel einer Standardabweichung.

Ohne diese Frage nun ausschweifend diskutieren zu wollen, scheint geboten, sowohl im Sinne der Testfairness, als auch in Hinblick auf Zwecke der Beratung von Personen, Geschlechterunterschiede im Bereich Mathematik mit zu untersuchen. Wie Abbildung 28, zeigt weisen Männer in drei der vier Skalen visuell erkennbare Unterschiede in ihrer

Leistung auf. Die Standardfehler wurden unterschlagen, weil sie aufgrund der sehr großen Stichprobe extrem klein ausfallen.

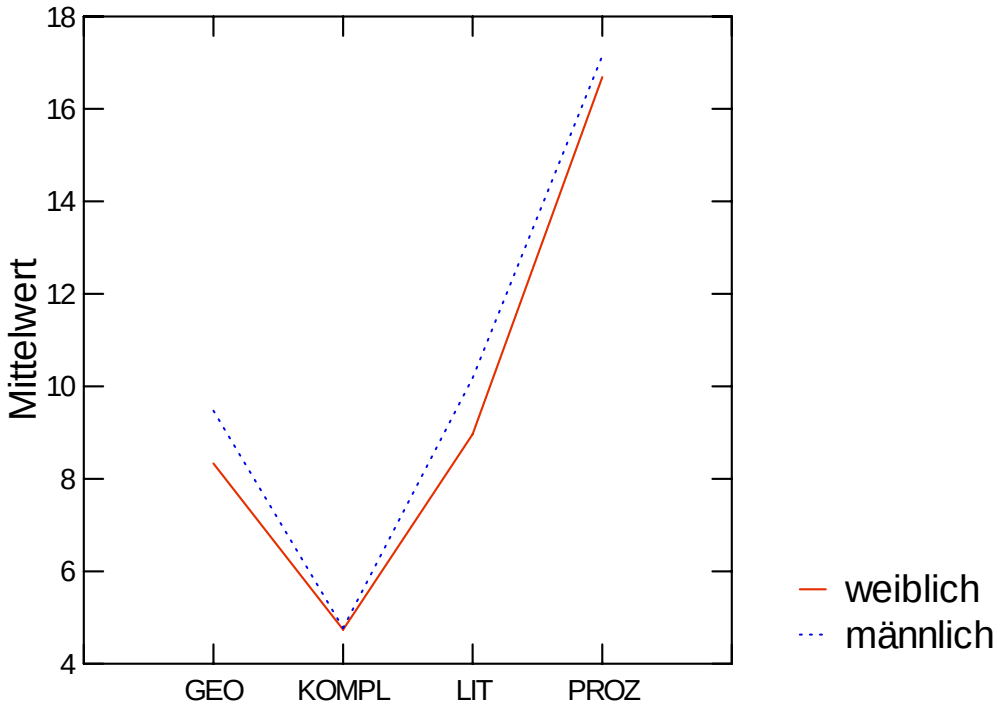


Abbildung 28. Unterschiede in den Mittelwerten aller Skalen getrennt für Männer und Frauen. N = 1554.

Die Tatsache, dass diese Unterschiede nur für Geometrie und grafische Funktionen und mathematische Literalität signifikant werden, spricht eher gegen wirklich bedeutsame Unterschiede zwischen den Geschlechtern, ist jedoch möglicherweise auch ein Ausdruck der starken G-Sättigung aller Skalen, die Profilunterschiedliche verwischen könnte. Die folgende Tabelle 41 listet die Mittelwertsdifferenzen einschließlich Cohens D (Cohen, 1992; Cohen et al., 2003) detailliert auf.

Tabelle 41 Mittelwerte und Mittelwertsunterschiede für Männer (*N* = 1048) und Frauen (*N* = 482) der Stichprobe.

	$\overline{x}_{\text{Männer}} - \overline{x}_{\text{Frauen}}$	$\overline{x}_{\text{Männer}}$	$\overline{x}_{\text{Frauen}}$	Cohen's d[1]
Geometrie und grafische Fkt.	1,14*	9,47	8,33	0,25
Prozedurales Rechnen	0,46	17,15	16,69	0,07
Mathematische Literalität	1,22*	10,18	8,96	0,39
Komplexes Rechnen	0,04	4,77	4,74	0,01

Anmerkung. P < 0,01* (t-test für unabhängige Stichproben). [1]Die Standardabweichung der Männer wurde gewählt, für keine der Skalen zeigte der Levene-Test eine Verletzung der Varianzhomogenität auf 1% Niveau.

Keine der Differenzen gemäß Tabelle 41 erreicht den Wert von d = 0,50 den Cohen als mittleren Effekt ansieht. Dies entspricht auch der subjektiven Empfindung, dass die

größten Geschlechterunterschiede (mathematische Literalität und Geometrie und grafische Fkt.) jeweils nur (etwa) einer durch die Männer zusätzlich gelösten Aufgabe entsprechen.

9.5 Profildiagnostik im Einzel- und Gruppenfall

Angenommen die Stelle eines Buchhalters in einem mittelständischen Unternehmens, das spezialisiert ist auf die Installation von Solaranlagen, ist zu vergeben und drei Bewerber kommen in die engere Auswahl. Wichtige Aufgaben für diese Stelle sind die Finanzbuchhaltung, Bezahlung von Lieferanten, Kundentransaktionen, Vorbereitung der Steuererklärung und das Erstellen eines jährlichen Abschlussberichts. In Abbildung 30 wurden die Profile dreier Bewerber abgetragen, die alle auf denselben Gesamtscore (allgemeine Mathematikfähigkeit) zurückgehen (Rohwert = 50, Z = 100) und somit zeigen, dass eine Betrachtung auf Skalenebene bedeutsame Mehrinformationen bringen kann.

Am deutlichsten sticht das Profil von Bewerber A ins Auge. Seine Stärke liegt eindeutig im Bereich der Mathematischen Literalität, also bei realitätsnahen, in Alltagskontexte eingebundenen Textaufgaben. Auch Geometrie beherrscht der Bewerber ähnlich gut wie seine Konkurrenten, doch fehlen ihm die Basiskenntnisse aus dem Bereich des prozeduralen Rechnens. Demnach wären für diese Stelle eher die Bewerber B und C zu bevorzugen, da ein sicheres Beherrschen aller Grundrechenarten und komplexerer Rechnungen unverzichtbar ist. Zwischen diesen beiden Stellenkandidaten fallen die Unterschiede wiederum eher gering aus. Letztlich fällt die Wahl auf Bewerber B. Er weist leicht überdurchschnittliche Fähigkeiten (Z > 100) in prozeduralem- und komplexem Rechnen auf. Zwar ist seine Leistung im Bereich Geometrie und grafische Funktionen noch etwas schlechter als jene von Bewerber C, doch immer noch durchschnittlich (Z = 100) und für die beschriebene Stelle sollten die Fähigkeiten in diesem Bereich nicht ausschlaggebend sein.

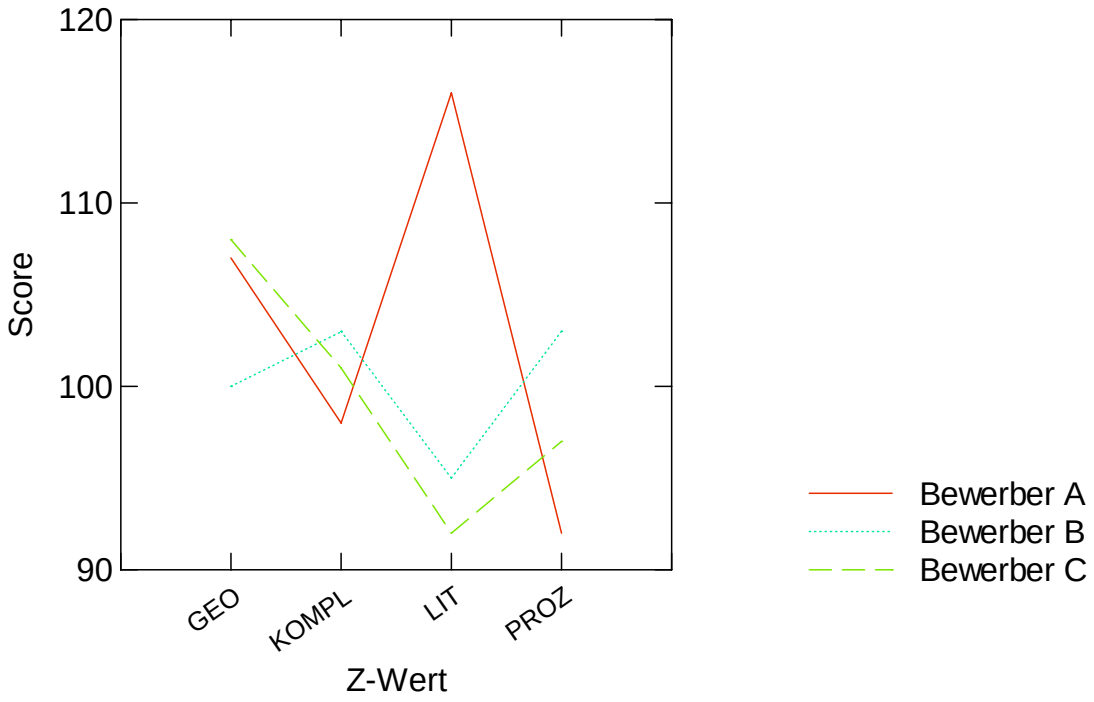


Abbildung 29 Standardwerte von 3 Personen der Normgruppe Gymnasial,über 20 Jahre alt. Alle Personen weisen denselben Gesamtscore auf (Z = 100, Rohwert = 50). Die kritische Differenz und die Normtabellen sind Jasper und Wagener (in Druck) zu entnehmen.

Neben dem dargelegten Beispiel zur Personalauswahl auf Einzelfallebene ist es auch möglich, generalisierende Aussagen zu Profilunterschieden auf Gruppenebene abzugeben. Um Profilunterschiede auf Skalenebene zu akzentuieren wurde zunächst eine Varimax-Faktorenanalyse der Parcels für komplexes Rechnen und mathematische Literalität durchgeführt (Tabelle 42).

Tabelle 42 Varimax-Rotierte Faktorladungsmatrix der Parcels für komplexes Rechnen und mathematische Literalität

	Faktor 1 (LIT)	Faktor 2 (KOMPL)
ParcelLit1	0,69	0,20
ParcelLit2	0,60	0,36
ParcelLit3	0,70	0,34
ParcelLit4	0,59	0,19
ParcelLit5	0,77	0,16
ParcelLit6	0,63	0,17
ParcelLit7	0,71	0,20
ParcelKompl1	0,36	0,65
ParcelKompl2	0,21	0,75
ParcelKompl3	0,23	0,80
ParcelKompl4	0,21	0,79
ParcelKompl5	0,23	0,79

Anmerkung. Die prozentuale Varianzaufklärung der Faktoren: F1: 29%, F2: 28%. N = 1554.

Die resultierenden Faktorwerte wurden genutzt, um ein Profilunterschiedlichkeitsmaß zu generieren, das sich zusammensetzt aus: *tilt = Faktorscore F1 – Faktorscore F2*. Dieses Maß gibt Auskunft über den tilt (Wittmann, 2004) des Profils. Werte größer Null stehen für einen tilt in Richtung mathematische Literalität (sprachgebunden), Werte kleiner Null hingegen für einen tilt in Richtung komplexes Rechnen (sprachfrei). Dieser Score wurde zusammen mit dem standardisierten Gesamtscore des Mathetests für die einzelnen von den Teilnehmern bisher erreichten Abschlüsse in Abbildung 30 abgetragen.

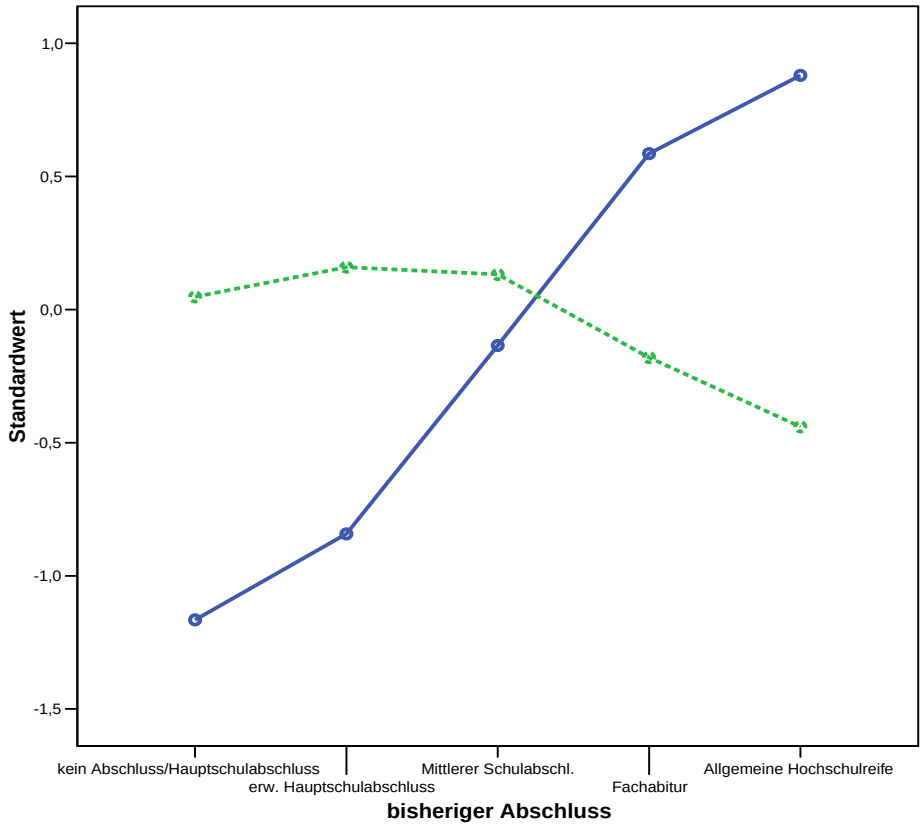


Abbildung 30 Tilt-Maß (gestrichelte Linie; größer Null: tilt Richtung mathematische Literalität) und Standardisierter Gesamtscore, getrennt nach bisher erreichtem Abschluss, *N* = 1554. Standardfehler sind aufgrund der großen Stichprobe irreführend und wurden daher nicht abgetragen.

Die Grafik zeigt zunächst ein deutliches Ansteigen des Gesamtscores über die einzelnen Abschlüsse hinweg. Bis einschließlich zum mittleren Schulabschluss zeigen sich deutliche Änderungen hinsichtlich des Gesamtlevels, jedoch gibt es kaum eine Änderung im tilt. Diese Änderung tritt erst für Probanden mit Fachabitur und allgemeiner Hochschulreife auf. Hier zeigt sich, dass sich der tilt klar in Richtung komplexes Rechnen verschiebt. Letztlich steigen die Leistungen der Probanden sowohl in mathematischer Literalität, als auch in komplexem Rechnen über die Abschlüsse hinweg an, jedoch stärker für komplexes Rechnen, was den beobachteten tilt erklärt. D.h. mit den höheren Schulabschlüssen geht ein regelrechter Schub im abstrakteren Denken einher, wie es für die Skala komplexes Rechnen typisch ist. Diese Erläuterungen zeigen, dass sowohl im Bereich der Gruppen-, als auch der Einzelfalldiagnostik die Bedeutung von Profilunterschieden nicht vernachlässigt werden sollte, da sie ein zusätzliches Klassifikationsmerkmal darstellt (Wittmann, 2004).

9.6 Multidimensional Random Coefficient Multinomial Logit Model

Das Multidimensional Random Coefficient Multinomial Logit (MRCML) Modell stellt die allgemeinste Variante in der Familie von Items-Response Modellen dar (Rost, 2004, S. 266). Ein wesentlicher Vorteil dieses Modells im Vergleich zu dem angewendetem NOHARM (McDonald, 1999) sind einmal inferenzstatistisch vergleichbare Fit-Indizes konkurrierender Modelle sowie die Möglichkeit ein Item gleichzeitig mehreren Traits zuzuordnen (Wu et al., 2007). Erst dadurch wird überhaupt die Prüfung eines Schmid-Leiman Modells möglich (im Gegensatz zu NOHARM). Darüber hinaus wurde das Modell auch bei allen PISA-Studien angewendet, wenngleich in den technischen Berichten über die genauen Programm-Einstellungen keine Informationen zu finden sind (OECD, 2005; OECD, 2009). Die Anwendung des MRCML-Modells findet sich in diesem Teil der Arbeit, da die Modelltestung in der Praxis Einschränkungen unterliegt (vgl. Tate, 2003) und da nach Information des Autors noch nie versucht wurde ein (beliebiges) Schmid-Leiman-Modell mit einer MRCML-Software zu realisieren.

Da ein Verständnis der angewendeten Methode für den Leser ohne vorab Erläuterung des Grundprinzips schwierig ist, wird demnach zunächst dieses Modell erklärt. Im folgenden werden die Elemente eines Vektors oder einer Matrix mit demselben Buchstaben, jedoch mit unterschiedlichen Indizes beschrieben. Im Fall eines Vektors wird ein Index verwendet und im Fall einer Matrix zwei. Ziel ist es die Lesbarkeit zu verbessern.

9.6.1 Das Rasch-Modell als Spezialfall

Das Rasch-Modell in seiner einfachsten Form stellt einen Spezialfall des MRCML-Modells dar, es wird häufig wie folgt geschrieben (Rost, 2004, S. 119):

$$P(X_{vi}=1)=\frac{\exp(\theta_v-\xi_i)}{1+\exp(\theta_v-\xi_i)} \tag{11}$$

Hier stellt θ die Personenfähigkeit dar, ξ_i hingegen die Schwierigkeit des Items i und die Gleichung in obiger Form beschreibt die (angenommene) Wahrscheinlichkeit der Antwort 1 (korrekt) auf Item i von einer Person mit Fähigkeit θ_v.

Es lässt sich jedoch auch als

$$P(X_{vi}=1)=\frac{\exp(\theta_v+\xi_i)}{1+\exp(\theta_v+\xi_i)} \tag{12}$$

schreiben, wenn man zuvor ξ_i durch $-\xi_i$ ersetzt, ohne die Modelleigenschaften zu verändern. Eine NOHARM-Lösung mit nur einem Faktor, die sowohl für die Endform als auch für die Vorform geprüft wurde (siehe 4.4.2.3 respektive 8.2.2), entspricht diesem Modell. Die Modellgleichung des MRCML-Modells (Adams et al., 1997, S. 3) enthält ebenfalls die Personenfähigkeit und die Itemschwierigkeit. Die Modellgleichung lautet:

$$P(X_{ik}=1;\mathrm{A},\mathrm{B},\xi|\theta)=\frac{\exp(\mathrm{b_{ik}}\theta+\mathrm{a'_{ik}}\xi)}{\sum_{k=1}^{K_i}\exp(\mathrm{b_{ik}}\theta+\mathrm{a'_{ik}}\xi)} \tag{13}$$

Diese Gleichung wird der Rasch-Modell Gleichung noch ähnlicher wenn man bedenkt, dass die 0te Antwortkategorie als Referenzkategorie angesehen wird, wodurch die „1 +“ der Rasch-Modell Gleichung wegfällt (Adams & Wu, 2007, S. 58). Dadurch, dass eine bedingte Wahrscheinlichkeit angegeben wird, muss diese Gleichung keinen Personenindex mehr enthalten und ist quasi für alle Personen mit den Fähigkeiten θ gültig.

Da das Modell für mehrere Dimensionen (mehrere Traits) ausgelegt ist und die Items mehr als zwei Kategorien (0 und 1) enthalten können stellen nun θ und ξ Vektoren und nicht Skalare, wie in der Rasch-Modell Gleichung oben, dar. θ enthält alle Traits die angenommen werden, d.h. im Falle des Mathetests 4 Traits, also $\theta_1, \theta_2, \theta_3$ und θ_4. Würde man Spearmans G-Theorie voraussetzen wäre θ schlicht ein Skalar, eine Zahl. $\xi,$ enthält einfach alle Itemparameter (Schwierigkeiten), bei drei dichotomen Items ξ_1, ξ_2 und ξ_3. Inhaltlich stellt $P(X_{ik}=1;\mathrm{A},\mathrm{B},\xi|\theta)$ nun die angenommene Wahrscheinlichkeit für die Antwort „1“ in Kategorie k des Items i (bei Rasch ist k = 0 bis 1) dar – unter der Bedingung einer bestimmten Ausprägung auf den latenten Dimensionen, die sich im Trait-Vektor θ befinden. Doch was bedeuten die beiden Matrizen A und B? Die A-Matrix ist vor allem dann interessant, wenn es mehr als zwei Antwortkategorien gibt, z.B. beim Partial-Credit-Modell (Masters, 1982). Dies ist jedoch für den vorliegenden Fall uninteressant. Bei allen in dieser Arbeit vorkommenden Modellen stellt die A-Matrix eine Einheitsmatrix dar.

In dem verwendeten Beispiel, also

$$A=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \tag{14}$$

wobei die erste Kategorie jeweils in der Diagonale der Matrix zu finden ist. In unserem Beispiel wird also a_{ik}' zu a_{i1}', weshalb wir die Bedeutung der unterschiedlichen Kategorien ab hier missachten (die 0te Kategorie wird wie erwähnt gleich 0 gesetzt). Dies heißt, dass in unserem Beispiel in jedem Fall der Itemparameter ξ_i aus ξ schlicht mit eins Mal genommen wird. Hiermit wären die A-Matrix und der ξ -Vektor erläutert, der folgende Abschnitt widmet sich der verbliebenen B-Matrix.

9.6.2 Within und between Item-Multidimensionalität

Die B-Matrix kann analog zu einer Ladungsmatrix verstanden werden. Sie bestimmt, welches der i Items welchem der Traits zugeordnet wird.

Hierbei unterscheiden Adams et al. (1997) zwischen – wie sie es nennen – *between Item multidimensionality* und *within-item multidimensionality*, was Abbildung 31 verdeutlichen soll.

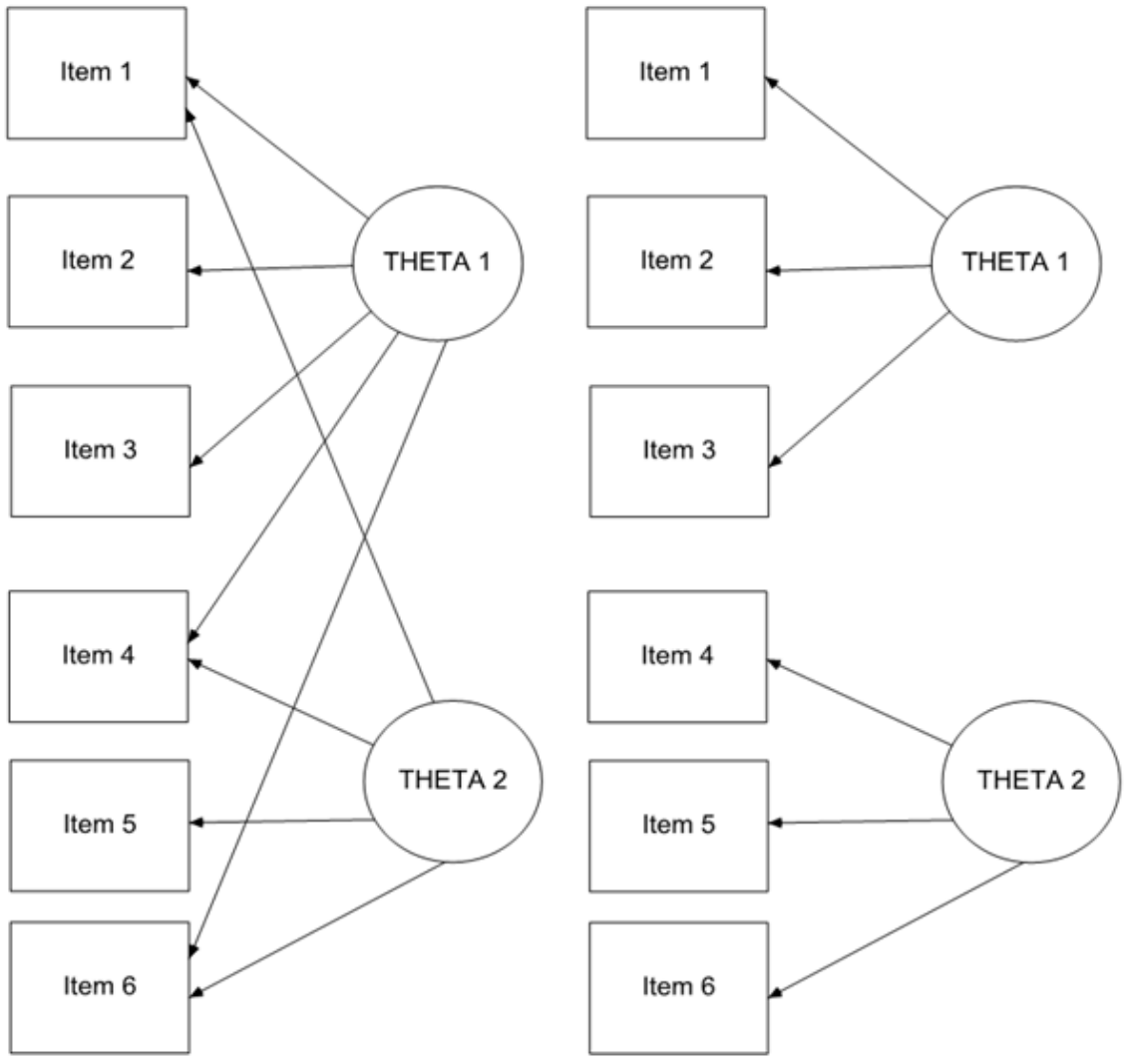


Abbildung 31 Verdeutlichung des Prinzips der *within-item* Multidimensionalität (linke Seite) und *between item* Multidimensionalität (rechts), angelehnt an Adams et al. (1997, S. 9).

Um ein Item i stets nur einem Trait (in der Grafik Theta 1 oder Theta 2) zuzuordnen muss demnach b_i, das eine Zeile aus der Matrix B darstellt, die die Ladung eines Items auf den Traits beschreibt, immer nur eine 1 enthalten. Für das rechte Beispiel in Abbildung 31 sieht B deswegen wie folgt aus:

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \tag{15}$$

Der Theta Vektor enthielte in Anlehnung an obige Abbildung nur zwei Elemente, nämlich THETA 1 und THETA 2, also:

$$\theta = \begin{bmatrix} \text{THETA 1} \\ \text{THETA 2} \end{bmatrix} \tag{16}$$

Für das linke Beispiel mit Mehrfachladungen müsste B hingegen wie folgt aussehen

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \tag{17}$$

D.h. an diesem Beispiel, für das Item 1, das den beiden Traits THETA 1 und THETA 2 zugeordnet ist, dass der entsprechende Vektor b_i, aus B nur die Elemente [1 1] enthält. Multipliziert mit dem θ-Vektor gehen für dieses Item also beide Traits, THETA 1 und THETA 2 in die Modellgleichung ein.

Somit wären alle Elemente der Modellgleichung erläutert und die extreme Vielseitigkeit dieses Modells ist demnach deutlich geworden. Da die Komplexität der Schätzalgorithmen ebenso wie die Erläuterung der Gleichung auf Ebene der Antwort einer einzelnen Person hier keinen Mehrwert versprechen wird für noch detailliertere Informationen auf Adams und Wu (2007) verwiesen.

9.6.3 Modelltests

Im MRCML-Modell stellt, wie aus obigen Ausführungen hervorgeht, der Test eines Multidimensionalen Modells lediglich einen Spezialfall dar. Der Schätzalgorithmus des Programms benötigt so genannte *nodes*, die dazu dienen das Fähigkeitskontinuum in mehrere Abschnitte aufzuteilen (Wu et al., 2007). Ein Problem stellt jedoch dar, dass die Anzahl der *nodes* exponentiell ansteigt. Das heißt für zwei Dimensionen 15^2, für drei Dimensionen 15^3 usw. Im Manual heißt es; ca. 5000 *nodes* wären in Bezug auf die Rechenzeit noch gut zu bewältigen (Wu et al., 2007). Dieses Dilemma stellt sich nun wie folgt dar: Bei Verwendung der empfohlenen 15 *nodes* pro Dimension macht dies für 3 Dimensionen 3375 *nodes*, was noch zu bewältigen wäre. Für die angestrebten 4 Dimensionen hingegen benötigen wir 50625 nodes. Tate (2003) beschreibt in einem Überblicksartikel, indem er verschiedene Varianten zur Prüfung der N-Dimensionalität eines Tests darstellt, seine Erlebnisse mit Conquest:

"Although the one-factor solution only required 8 minutes in computing time, the two-factor solution required about an hour. This computing time was too great to allow inclusion of this procedure in the comparisons to be discussed later" (S. 167). Die Autoren der Conquest-Software empfehlen ab 3 Dimensionen die nodes mit einem integrierten Monte-Carlo-Verfahren schätzen zu lassen (Wu et al., 2007), was die Rechenzeit enorm verkürzt und hier angewendet wurde.

Zunächst werden die Ergebnisse von 3- und 4-Faktormodellen korrelierter Dimensionen dargestellt. Anschließend wird versucht einen SL-Modell Ansatz mit diesem (IRT) Testmodell zu verwirklichen.

9.6.3.1 Conquest: 3- und 4 Faktormodelle

Um den Fit eines Modells in Relation zu einem konkurrierenden Modell zu vergleichen gibt die Conquest Software so genannte *Deviance* Werte aus, die die mit -2 multiplizierte Modell Log-Likelihood darstellen (Brandt, 2003; Wu et al., 2007). Für das 3-und 4 Faktor Modell ergaben sich Werte von $Deviance_{3Faktor}$= 111183 und $Deviance_{4Faktor}$= 110687. Der Unterschied zwischen diesen beiden *Deviance* Werten ist χ^2-verteilt (Wu et al., 2007) und kann daher auf Signifikanz getestet werden. Im 4-Faktor-Modell müssen drei Parameter zusätzlich geschätzt werden, da es drei zusätzliche Korrelationen zwischen laenten Variablen gibt, weshalb $df_{chi} = 3$ und $\chi^2_{Deviance} = 496.4$ was einen hochsignifikanten Unterschied zugunsten des 4-

Faktormodells ergibt. Die folgende Tabelle 43 zeigt die Korrelationen zwischen den latenten Dimensionen die durch Conquest geschätzt wurden.

Tabelle 43 Korrelationen zwischen den IRT-basierten Dimensionen nach Conquest (*N* = 1554).

	Geometrie und grafische Fkt.	Prozedurales Rechnen	Mathematischer Literalität
Prozedurales Rechnen	0,80		
Mathematische Literalität	0,68	0,74	
Komplexes Rechnen	0,80	0,89	0,77
	3-Faktorlösung		
		Prozedurales/komplexes Rechnen	
Prozedurales/komplexes Rechnen	0,83		
Mathematische Literalität	0,66	0,75	

Die Korrelationen fallen generell recht hoch aus, sind jedoch klar im Rahmen der Erwartungen und ähneln sehr jenen die mit NOHARM berechnet wurden. Hervorzuheben ist, dass die Korrelationen dennoch deutlich niedriger ausfallen als jene der 4 Inhaltsbereiche der PISA 2003 Studie, die zu Beginn der Arbeit in Abschnitt 2.2.2 vorgestellt wurden.

9.6.3.2 Conquest: 3- und 4 Faktor SL-Modelle

Nach Information des Autors und eingehender Literaturrecherche wurde keine Arbeit gefunden, in der ein Schmid-Leiman-Modell (Schmid & Leiman, 1957) mit conquest berechnet wurde. Es ist möglich in diesem Programm so genannte *anchoring* Werte zu fixieren, wie z.B. die Korrelationen zwischen den Dimensionen. Darüber hinaus kann durch das within-item Multidimensionalitätsprinzip – im Gegensatz zu NOHARM – zeitgleich ein G-Faktor postuliert werden. Da ein solches Modell im IRT-Kontext mit conquest bisher nicht erprobt wurde, schien es geboten, eine Anfrage bei einem der

Programmautoren zu stellen. Prof. Ray Adams bestätigte, dass prinzipiell kein Problem mit der Realisierung eines Schmid-Leiman Modells bestünde, jedoch Monte-Carlo Studien sinnvoll wären, um das Verhalten des Programms genauer zu untersuchen. Letztlich ergaben sich bei der Berechnung des 3-Faktor-SL-Modells eine Deviance von 111669 und im Falle des 4-Faktor-SL-Modells von 111492. Bei diesen Modellen, die nicht als nested models betrachtet werden können (siehe auch Abschnitt 9.2) kann nur grob geschlussfolgert werden, dass die 4-Faktorlösung in einer geringeren Abweichung resultiert als die 3-Faktor Lösung. Ob diese Modelle besser fitten als die nicht hierachischen conquest Modelle kann daher nicht beantwortet werden. Es scheint jedoch noch vertretbar anzugeben, dass die deviance Werte der SL-Lösungen nicht weit von den vorherigen Lösungen korrelierter Faktoren entfernt sind, was beide Varianten als prinzipiell tauglich ausweist.

9.7 Strukturelle Trennbarkeit der Taxonomiestufen

In Abschnitt 8.4, bei der Besprechung zu den Ergebnissen der Lehrerbefragung, wurde die Überprüfung der statistischen Trennbarkeit der Skalen auf Basis der kognitiven Prozesse in Anlehnung an Anderson und Krathwohl (2001) bewusst ausgespart. Das Vorhaben scheint gewagt, da überhaupt nur eine ausreichende Raterreliabilität durch die relativ hohe Anzahl von Ratern gewährleistet war. Darüber hinaus ist es unangemessen ein mittleres Rating für jedes Item zu berechnen (in Abschnitt 8.4.3.5 wurde lediglich die mittlere Häufigkeit eines Ratings berechnet). Daher bleibt an dieser Stelle eigentlich nur der Modus (Hays, 1994) über die 17 Rater, um sich für eine Kategorie pro Item zu entscheiden. Aus all diesen Gründen findet sich diese Analyse im Abschnitt 9 dieser Arbeit, der auch einen Ausblick geben soll, was in Zukunft vielleicht noch genauer erforscht werden könnte. Die folgende Tabelle 43 zeigt die Zuordnung aller Items zu den ersten 4 Taxonomiestufen. Die Stufen 5 und 6 fehlen, da sie nie den Modus darstellten.

Tabelle 44 Alle Items, die gemäß dem Modus der Kategoriezuordnung durch 17 Rater den ersten 4 Stufen der Lernzieltaxonomie zugeordnet wurden.

Taxonomiestufe							
erinnern		verstehen		anwenden		Analysieren	
Item	Skala	Item	Skala	Item	Skala	Item	Skala
A1	GEO	A6a	GEO	A2	GEO	A25b	LIT
A5	GEO	A6b	GEO	A3	GEO	A25c	LIT
A12a	GEO	A6c	GEO	A4	GEO	A25d	LIT
A12b	GEO	A10a	GEO	A7	GEO	A26c	LIT
A18	PROZ	A11a	GEO	A8	GEO	A27a	LIT
A19	PROZ	A11b	GEO	A9	GEO	A27d	LIT
		A11c	GEO	A10b	GEO	A27e	LIT
		A11d	GEO	A10c	GEO	A28	LIT
		A15d	PROZ	a13a	PROZ	A29	LIT
		A16a	PROZ	a13b	PROZ	A30a	LIT
		A16b	PROZ	a13c	PROZ	A30b	LIT
		A16c	PROZ	A14a	PROZ		
		A22a	PROZ	A14b	PROZ		
		A26a	LIT	A14c	PROZ		
		A26b	LIT	A15a	PROZ		
		A33	KOMPL	A15b	PROZ		
				A15c	PROZ		
				A17a	PROZ		
				A17b	PROZ		
				a20a	PROZ		
				a20b	PROZ		
				A21a	PROZ		
				A21b	PROZ		
				A22b	PROZ		
				A22c	PROZ		
				A23a	PROZ		
				A23b	PROZ		
				A23c	PROZ		
				A24a	PROZ		
				A24b	PROZ		
				A24c	PROZ		
				A24d	PROZ		
				A27b	LIT		
				A27c	LIT		
				A31a	KOMPL		
				A31b	KOMPL		

Tabelle 44 Fortsetzung.

Taxonomiestufe							
erinnern		verstehen		anwenden		Analysieren	
Item	Skala	Item	Skala	Item	Skala	Item	Skala
				A31c	KOMPL		
				A31d	KOMPL		
				a32a	KOMPL		
				a32b	KOMPL		
				A34a	KOMPL		
				A34b	KOMPL		
				A34c	KOMPL		
				A35	KOMPL		

Anmerkung. KOMPL = komplexes Rechnen, PROZ = prozedurales Rechnen, LIT = mathematische Literalität, GEO = Geometrie und grafische Fkt.

Basierend auf den Daten in Tabelle 43 wurden, analog zu dem Vorgehen in Abschnitt 8.3, 38 Parcels zusammengestellt, die damit jeweils hinsichtlich Schwierigkeit ausbalanciert wurden. Diese Struktur wurde nun in MPLUS unter Verwendung der WLSMV und ML Schätzmethode (analog zu Abschnitt 8.3.2.4) auf Passung mit den Normstichprobendaten geprüft. Es ergaben sich ein CFI von $CFI_{ML} = 0{,}79$ ($CFI_{WLSMV} = 0{,}80$) und ein RMSEA von $RMSEA_{ML} = 0{,}074$ ($RMSEA_{WLSMV} = 0{,}096$). Diese Werte entsprechen etwa dem Fit der 4 ursprünglich postulierten 3- und 4-Faktormodelle der Inhaltsfacetten in Abschnitt 8.3.2 und sind unzureichend, weshalb sich ein genauerer Modellvergleich erübrigt. Die geringste Korrelation der latenten Taxonomiestufen untereinander findet sich für Analysieren und Erinnern mit $r = 0{,}60$ ($r_{WLSMV} = 0{,}56$), die höchste zwischen anwenden und verstehen mit $r = 0{,}87$ ($r_{WLSMV} = 0{,}81$).

Eine seit Kropp und Stoker (1966) in der Literatur kontrovers diskutierte Frage (Hill & McGraw, 1981) ist der hierarchische Aufbau der Taxonomiestruktur (vgl. Abschnitt 3.2.1.1), der - geht man nur von den ersten 4 hier betrachteten Stufen aus - eine Struktur gemäß Abbildung 32 aufweisen sollte.

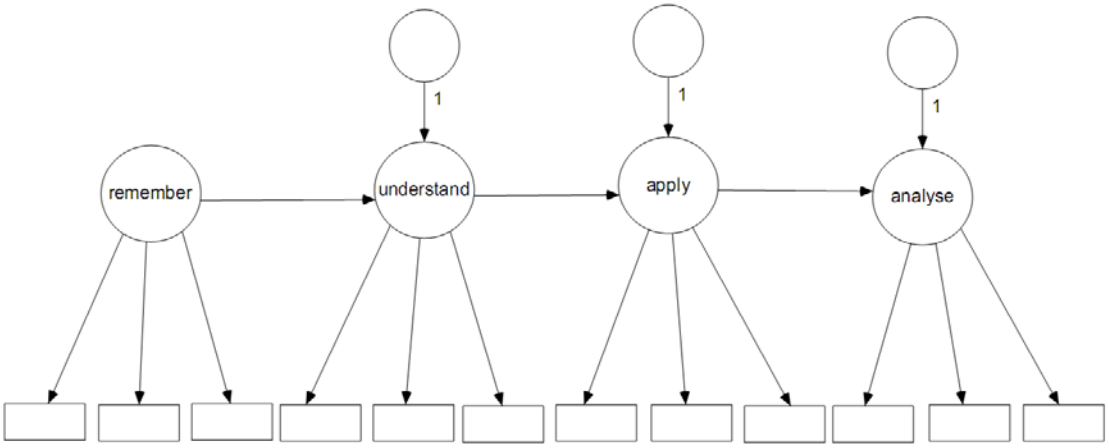


Abbildung 32 Schematische Darstellung des hierarchischen Aufbaus der ersten 4 Taxonomiestufen.

Auch wenn Anderson und Krathwohl (2001) keine strikte Hierarchie mehr annehmen, schien es, insbesondere da gleichrangige korrelierte Faktoren kein befriedigendes Ergebnis erbrachten, sinnvoll diese Struktur zu prüfen. Es ergab sich jedoch nur ein CFI von 0,79 (CFI_{WLSMV} = 0,79) und ein RMSEA von 0,097 ($RMSEA_{WLSMV}$ = 0,074). Darüber hinaus tauchten in der ML-Lösung Pfade größer 1 zwischen den Stufen auf und in der WLSMV-Lösung erreichten alle Pfade zwischen den Stufen den Wert 1. Demnach ist zumindest für diesen Mathematiktest, mit der Einschätzung durch Lehrerratings keinesfalls eine hierarchische Struktur nachweisbar. Ein möglicher Grund hierfür stellt eine zu starke Konfundierung von Mathetestskala und Stufenlevel dar, die in obigem Modell nicht bedacht wird. Bereits anhand Tabelle 43 scheint es so, als wäre zumindest die Skala mathematische Literalität für die Stufe analysieren deutlich überrepräsentiert, wobei die Tabelle nur den Modus enthält. Dem Aspekt einer genaueren Prüfung dieser Fragestellung widmet sich der folgende Abschnitt.

9.8 Zusammenhang von Taxonomielevel und Skalenzugehörigkeit

Es wurde bereits gezeigt, dass die Rater über alle Items hinweg die Stufenzuordnungen sehr unterschiedlich häufig vergaben. Nun soll geprüft werden, ob sich die (relativen) Häufigkeiten mit der Items einer Stufe zugeordnet werden für die vier Skalen des Mathetests unterscheiden. Die Rautenform des integrativen Modells, das in Abschnitt 3.3 vorgestellt wurde impliziert einen Zusammenhang der Inhaltsdimension (die 4 Skalen) und der Dimension kognitiver Prozesse (die 6 Taxonomiestufen).

Gerade weil für die vorangegangene Modelltestung der Erfolg ausblieb, bietet sich die Analyse einer weiterführenden Fragestellung an und zwar, ob für jede der 4 Mathetestskalen in gleichem Ausmaß Ratings für die 6 Taxonomiestufen gegeben wurden. Hierfür wurde die folgende Tabelle 45 erstellt, welche für die ersten 4 Stufen

der Taxonomie Antworthäufigkeiten nach Skalen zusammenfasst und den bei Unabhängigkeit der Zuordnungen erwarteten Häufigkeiten gegenüberstellt.

Tabelle 45 Häufigkeiten mit denen von den 17 Ratern Items der 4 Mathetestskalen den ersten 4 Taxonomiestufen zugeordnet wurden.

	erinnern	verstehen	anwenden	analysieren	Summe
Geometrie und grafische Fkt.	54 (42)	111 (98)	109 (143)	57 (56)	331
Prozedurales Rechnen	88 (65)	146 (138)	250 (222)	29 (88)	513
Mathematische Literalität	8 (29)	45 (62)	68 (99)	108 (39)	229
Komplexes Rechnen	9 (23)	36 (49)	116 (78)	20 (31)	181
Summe	159	338	543	214	1254

Anmerkung. In Klammern erwartete Häufigkeit bei Unabhängigkeit.

Da insgesamt, über alle 77 Items und 17 Rater, nur 39 mal die Stufe Evaluieren und 2 mal Kreieren gewählt wurden, sind sie nicht in der Tabelle aufgeführt. Der Grund liegt darin, dass dadurch sehr geringe Zellhäufigkeiten entstehen, die bei χ^2-Tests zu Problemen führen können (Hays, 1994; Siegel, 1956). Ein χ^2-Test zur Prüfung der Unabhängigkeit ergibt einen hochsignifikanten Wert von $\chi^2 = 253$ (df = 9, p < 0,00), d.h. die Zuordnungen sind eindeutig nicht voneinander unabhängig.

Interessant ist an dieser Stelle zu prüfen, wodurch diese Abhängigkeit zutage tritt. Wie Agresti (2007) feststellt sind die Roh-Residuen zwischen erwarteten und beobachteten Werten ungeeignet, da sie in diesem Fall von der unterschiedlichen Anzahl von Items der verschiedenen Skalen verzerrt werden. Ein besseres Maß stellt das standardisierte Residuum von beobachteten (O_{ik}) und erwarteten (E_{ik}) Häufigkeiten nach der Formel $\mathrm{Res} = \frac{O_{ik} - E_{ik}}{\sqrt{E_{ik}(1-p_{i.})(1-p_{.j})}}$ dar, wobei i = 1 bis 4 für die Taxonomistufen und k = 1 bis 4 für die Mathetestskalen stehen (Agresti, 2007, S. 38).

Tabelle 46 Standardisierte Residuen bei Annahme von Unabhängigkeit der Zuordnung Taxonomiestufe x Skalenzugehörigkeit.

	erinnern	verstehen	anwenden	analysieren
Geometrie und grafische Fkt.	2,3	3,1*	-4,4*	0,1
Prozedurales Rechnen	4,0*	1,0	3,2*	-8,9*
Mathematische Literalität	-4,6*	-2,8	-4,6*	13,4*
Komplexes Rechnen	-3,4*	-2,3	6,1*	-2,3

Anmerkung. Werte |RES|≥3 stellen signifikante Abweichungen dar, gekennzeichnet mit*.

Wie bereits anhand des χ^2-Wertes zu erwarten war, zeigen sich zahlreiche signifikante Abweichungen, wobei Agresti (2007) empfiehlt Abweichungen größer |3| als bedeutsam anzusehen. In Bezug auf die Taxonomiestufen fällt auf, dass analysieren bei der Skala prozedurales Rechnen unerwartet selten und bei der Skala mathematische Literalität unerwartet häufig auftaucht, was – post hoc betrachtet – durchaus zu den Skalenbeschreibungen in Abschnitt 3.1.6 passt. Direkt umgekehrt, jedoch nicht in gleichem Ausmaß, verhält es sich für die Stufe erinnern.

Bei Betrachtung in Bezug auf die Skalen fällt auf, dass sich die beste gleichmäßige Verteilung der Taxonomieratings über alle Stufen für Geometrie und grafische Funktionen ergibt, da die Residuen alle recht moderat ausfallen.

10 Gesamtdiskussion und Ausblick

In diesem Abschnitt geht es darum, den Wissenstand zur Psychometrie der Mathematik am Ende der Sekundarstufe I zu bewerten und einen Ausblick zu wagen.

Am Ende dieser Arbeit scheint definitiv gesichert, dass Mathematik wohl mehr als eindimensional ist. Dies konnte sowohl anhand eines studentischen Tests (vgl. Abschnitt 4.4), als auch anhand der Endform bestätigt werden (8.1.1). Auch wenn sie nicht im direkten Fokus der Arbeit standen konnten die Analysen zur Profildiagnostik (9.5) diese Erkenntnis untermauern. Auf der anderen Seite entsprach ein Modell der Inhaltsdimension, das 4 korrelierte Skalen (oder Facetten) umfasst vom Fit her überhaupt nicht den Erwartungen (vgl. Abschnitt 8.3.2). Darüber hinaus wies das Modell hohe Korrelationen zwischen den Dimensionen auf, die für einen G-Faktor sprechen. Eine logische Schlussfolgerung stellt das daraufhin aufgestellte Schmid-Leiman-Modell dar. Leider handelt es sich dabei um ein erst nach den anderen Analysen geprüftes Modell und kein vorab theoretisch spezifiziertes Modell.

Ein weiterer kritischer Punkt, der dem aufmerksamen Leser sicher aufgefallen ist, besteht in der partiellen Vermischung von kognitivem Prozess und Inhalten bei den Skalen prozedurales- und komplexes Rechnen. Streng genommen sind diese Skalen beide – gemäß Skalendefinition (Abschnitt 3.1.6) – der numerischen Intelligenzkomponente zuzuordnen. Ihre Trennung resultiert aus der Annahme, verschiedene Prozesse müssten angewendet werden. Dieses Problem entschärft sich jedoch, wenn man sich an Abschnitt 9.8 zurückerinnert. Anhand der standardisierten Residuen konnte gezeigt werden, dass die Abhängigkeit von einer Einordnung zu kognitiven Prozessen (gemäß Taxonomie) von der Skalenzugehörigkeit keineswegs nur die Skalen prozedurales- und komplexes Rechnen betraf.

Die Formulierung zur Bedeutung der Taxonomiestufen in dem integrativen Modell (Abschnitt 3.3) zu Beginn der Arbeit war bewusst sehr vorsichtig. So hieß es dort, dass ein zusätzliches Ordnungsmerkmal geschaffen werden soll. Schließlich gelang auch eine ausreichend reliable Einordnung der Items in die Kategorien und es zeigte sich, dass der Test vor allem jene Bereiche durch viele Items abdeckt, die Lehrer für wichtig betrachten, doch zeigte sich ebenso, dass bestimmte Taxonomielevel bei manchen Skalen unter- (z.B. Analysieren bei prozeduralem Rechnen) bzw. überrepräsentiert (z.B. analysieren bei mathematischer Literalität) sind. Sicherlich hätte man versuchen können, die Taxonomiestufen auch bereits während der Testkonstruktion stärker zu beachten. Dies wäre jedoch nur mit deutlich weniger als 17 Lehrern möglich gewesen, aus rein praktischen Gründen. Ferner hätte dies den eigenen Literaturrecherchen widersprochen, die eher gegen eine Verwendung in der Testentwicklung sprachen (Blumberg, et al., 1982; Cizek et al., 1995). Darüber hinaus spricht die eher mittelmäßige Reliabilität der Lehrerurteile dafür, dass auch ein solcher Ansatz die Schwierigkeit des empirischen Nachweises der Taxonomiestufen nicht gelöst hätte. Ob ein Ausbalancieren der verschiedenen Stufen über die 4 Bereiche mathematische Literalität, prozedurales-, komplexes Rechnen sowie Geometrie und grafische Fkt. geholfen hätte die Skalen besser nachzuweisen, ist nicht einfach zu beantworten. Hier wird die Meinung vertreten, dass dies theoretisch möglich wäre, jedoch in der Praxis durch einen Versuch, z.B. Aufgaben zu *analysieren* über alle 4 Skalen hinweg explizit zu konstruieren, die Inhaltsfacetten noch unreiner geworden wären und man am Ende die 4 Skalen überhaupt nicht mehr nachweisen könnte. Da jedoch die Skalen hier als noch wichtiger – verglichen mit den Stufen - angesehen werden scheint dieser Weg nicht gangbar. Der wichtigste Beitrag der Taxonomiestufen in dieser Arbeit liegt wohl

letztlich darin, dass er Lehrerwünsche mit dem abgleicht, was der Test aus Lehrersicht erfasst.

Diese Arbeit soll nicht ohne die Frage nach der Angemessenheit komplexer statistischer Methoden abgeschlossen werden. Von einem der Psychologie eher fachfremden, dem Ökonomen und Nobelpreisträger Ronald Coase, stammt zum übertriebenem Einsatz von Datenanalysen der Satz: „If you torture the data long enough it will confess“ (Tullock, 2001 S. 205). Demnach muss auch die kritische Frage gestellt werden, ob in dieser Arbeit zu viele, zu komplexe Methoden angewandt wurden. Einerseits wurde gezeigt, wie in praktischer Forschung Schwierigkeitsfaktoren entstehen können, wenn die Faktorenanalyse auf dichotome Items angewandt wird (Abschnitt 8.2.1), andererseits ist durchaus diskutabel, ob DIMTEST/DETECT, NOHARM, Conquest und vor allem HCA/CCPROX genügend zusätzliche Erkenntnis geliefert haben, um den doch deutlich höheren Analyseaufwand zu rechtfertigen. Dennoch, sicher wäre es nicht gelungen, ein Modell ähnlich der postulierten 4-faktoriellen SL-Struktur zu entdecken, ohne moderne Strukturgleichungsmodelle nutzen zu können (auch wenn es theoretisch möglich wäre). Letztlich diente die Anwendung vieler unterschiedlicher Methoden auch dazu, die Befunde möglichst gut abzusichern.

Ein Aspekt der in dieser Arbeit lediglich gestreift wurde, ist die Frage nach der Konstrukt- oder besser Messungsinvarianz des Mathematikmodells über verschiedene Populationen. Lässt sich die gefundene Modellstruktur für verschiedene Populationen (z.b. Haupt- vs. Realschüler) bestätigen? Falls dies der Fall ist, könnten in einem weiteren Schritt Unterschiede in den Mittelwerten latenter Variablen geprüft werden (vgl. Brunner et al., 2007; Little, 1997). Ein Problem hierbei stellt jedoch der - auch für das SL-Modell - nach wie vor verbesserungswürdige Fit dar. Normalerweise sollte ein guter Fit Vorraussetzung für derartige Analysen sein (Kline, 2005). Die Annahme metrischer Invarianz (Little, 1997) würde z.B. im Falle des 4 Faktor-SL-Modells beinhalten, dass einige Geometrie und grafische Fkt.-Parcels in beiden Gruppen Nullladungen (vgl Abschnitt 9.2) aufweisen sollen, was eine fragliche Modellannahme darstellt.

Ein völlig anderer, ebenfalls sehr interessanter Ansatz bestünde in der Anwendung der MTMM-Technik (Campbell & Fiske, 1959) auf einen Test, der Mathematikfähigkeiten erfassen soll. Das Vorgehen, einen Trait mit mehreren Methoden und mehrere Traits mit einer Methode zu erfassen, stellt eine Herausforderung an jedes Konstrukt dar. Campbell und Fiske (1959, S. 100) gingen sogar so weit zu fordern, dass: „before one

can test the relationships between a specific trait and other traits, one must have some confidence in one's measures of that trait." Hierfür einen MTMM-Ansatz zu fordern ist, vermutlich eine etwas zu strenge Sichtweise, doch bleibt die Frage interessant, ob eine Messung von Mathematikfähigkeiten auch ohne Fragebögen möglich ist und sinnvolle Ergebnisse erreicht werden können.

Ein sehr interessanter Aspekt für zukünftige Ansätze besteht darüber hinaus darin, den entwickelten Test auf eine Computer-Variante zu übertragen. Es wäre z.B. interessant, wenn für jede der Aufgaben Informationen vorlägen, wie lange sie eine Person bearbeitet hat. Schrecken z.B. manche Personen vor Aufgaben der Skala komplexes Rechnen derart zurück, dass sie nach wenigen Sekunden bereits zur nächsten Aufgabe übergehen? Es gibt noch viele Bereiche im Rahmen der Psychometrie der Mathematik am Ende der Sekundarstufe I zu erforschen. Diese Arbeit stellt einen Beitrag dar, weitere werden folgen.

11 Literatur

Abad, F. J., Colom, R., Juan-Espinosa, M. & García (2002). Intelligence differentiation in adult samples. *Intelligence, 31*, 157-166.

Abswoude, A., Ark, L. A. & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures undes nonparametric IRT models. *Applied Psychological Measurement*, 28, 3-24.

Ackerman, P. L. (1996). A theory of adult intellectual development: process, personality, interests, and knowledge. *Intelligence*, 22, 229-259.

Ackerman, P. L. (2002). Gender differences in intelligence and knowledge: How should we look at achievement score differences? *Issues in Education: Contributions from Educational Psychology*, 8, 21-29.

Adams, R. J., Wilson, M. & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23

Adams, R. J. & Wu, M. (2007). The mixed-coefficients multinomial logit model: A generalized form of the rasch model. In: M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution rasch models* (pp. 57-77). Springer: Berlin.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd Ed.). New Jersey: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: N. Petrov and F. Csadki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp. 267-281). Budapest: Akademiai Kiado.

Amelang, M. & Zielinski, W. (2001). *Diagnostik und Intervention*. Berlin: Springer.

Amthauer, R. (1953). *I-S-T. Intelligenz-Struktur-Test* (2. Aufl.). Göttingen: Hogrefe.

Amthauer, R. (1973). *Intelligenz-Struktur-Test 70 (I-S-T 70)*. Göttingen: Hogrefe.

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (1999). *Intelligenz-Struktur-Test 2000 (I-S-T 2000)*. Göttingen: Hogrefe.

Anderson, L. W. (1999). *Rethinking Bloom's Taxonomy: Implications for Testing and Assessment*. University of South Carolina.

Anger, H., Mertesdorf, Wegner, R. & Wülfing, G. (1980). *VKI Verbaler Kurzintelligenztest.* Göttingen: Hogrefe.

Aster, M., Neubauer, A. & Horn R. (2006). *Wechsler Intelligenztest für Erwachsene (WIE). Deutschsprachige Bearbeitung und Adaptation des WAIS-III von David Wechsler*. Frankfurt: Harcourt Test Services.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Ayala, R. J. (2008). *The theory and practice of items response theory*. New York: Guilford Press.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2006). *Multivariate Analysemethoden* (11. Aufl.). Berlin: Springer.

Baden-Württemberg (2004a). *Bildungsstandards für Mathematik - Werkrealschule - Klassen 9,10*. Abruf März, 03, 2009, unter http://www.bildung-staerkt-menschen.de/ service/downloads/Bildungsplaene

Baden-Württemberg (2004b). *Bildungsstandards für Mathematik - Hauptschule - Klassen 6, 9*. Abruf März, 03, 2009, unter http://www.bildung-staerkt-menschen.de/ service/downloads/Bildungsplaene

Baden-Württemberg (2004c). *Bildungsstandards für Mathematik - Realschule - Klassen 6, 8, 10*. Abruf März, 03, 2009, unter http://www.bildung-staerkt-menschen.de/ service/downloads/Bildungsplaene

Baden-Württemberg (2004d). *Bildungsstandards für Mathematik - Gymnasium-Klassen 6, 8, 10*. Abruf März, 03, 2009, unter http://www.bildung-staerkt-menschen.de/ service/downloads/Bildungsplaene

Balser, H., Ringsdorf, O. & Traxler, A. (1986). *Berufsbezogener Rechentest.* Weinheim: Beltz.

Bandalos, D. L. & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling New developments and techniques* (pp. 269–296). Mahwah: Lawrence Erlbaum.

Beauducel, A. & Herzberg, P.Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least square estimation in confirmatory factor analysis. *Structural Equation Modeling, 13,* 186-203

Beauducel, A. & Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin model of intelligence structure. *European Journal of Psychological Assessment, 18*, 97-112.

Beauducel, A. & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75.

Bender, P. (2005). Neue Anmerkungen zu alten und neuen PISA-Ergebnissen und – Interpretationen. In: G. Graumann (Hrsg.), *Beiträge zum Mathematikunterricht 2005* (S. 73-77). Hildesheim: Franzbecker.

Benson, J. & Fleishman, J. A. (1994). The robustness of maximum likelihood and distribution-free estimators to non-normality in confirmatory factor analysis. *Quality and Quantity, 28*, 117-136.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Quantitative Methods in Psychology, 2*, 238-246.

Bentler, P. M. (2003). *EQS 6 structural equations program manual*. Enico, CA: Multivariate Software, Inc.

Bentler, P. M. & Yuan, K-H. (1999). Structural equation modeling with small samples: test statistics. *Multivariate Behavioral Research, 34*, 181-197.

Bloom, B. S. (1976). *Human characteristics and school learning.* New York: McGraw-Hill.

Bloom, B. S. (1994). Reflections on the development and use of the taxonomy. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's Taxonomy. A Forty-Year Retrospective* (pp. 1-9). Chicago: Chicago Press.

Bloom, B. S., Englehart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain*. New York: McKay.

Blumberg, P., Alschuler, M. D. & Rezmovic, V. (1982). Should taxonomic levels be considered in developing examinations? *Educational and Psychological Measuerment, 42*, 1-7.

Bodin, A. (2007). What does PISA really assess? What does it not? A French view. In S. T. Hopman, G. Brinek & M. Retzl (Hrsg.), *PISA zufolge PISA – PISA according to PISA* (S. 21-57). Wien: Lit.

Bollen, K. A. and Long, J. S.(1993). *Testing Structural Equation Models*. Newbury Park, CA: Sage.

Bond, T. G. & Fox, C. M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences* (2nd Ed.). Mahwah: LEA.

Borstel, S. (2008). Schlecht in Mathe. Viele Jugendliche taugen nicht für Lehrstellen. *Die Welt Online*. Abruf Mai 01, 2008, unter http://www.welt.de/wirtschaft/article1989056/Viele_Jugendliche_taugen_nicht_fr_Lehrstellen.html.

Brandt, S. (2003). Estimation of a Rasch model including subdimensions. *IEA monograph series: Issues and methodologies in large-scale assessments*. Abruf Mai 05 ,2009, unter http://www.ierinstitute.org/IERI_Monograph_Volume_01_Chapter_3.pdf

Bremm, M. H. & Kühn, R. (1992). *Rechentest RT 9+*. Weinheim: Beltz.

Brocke, B. & Beauducel, A. (2001). Intelligenz als Konstrukt. In E. Stern & J. Guthke (Eds.), *Perspektiven der Intelligenzforschung. Ein Lehrbuch für Fortgeschrittene* (S. 13-42). Lengerich: Pabst Science Publisher.

Brocke, B., Beauducel, A. & Tasche, K.G. (1998). Der Intelligenz-Struktur-Test: Analysen zur theoretischen Grundlage und technischen Güte. *Diagnostica, 44,* 84-99.

Brunner, M., Krauss, S. & Kunter, M. (2007).Gender differences in mathematics: does the story need to be rewritten. *Intelligence, 25*, 1-19.

Brunswik, E. (1952). *The conceptual framework of psychology*. In International Encyclopedia of Unified Science (Vol. 1). Chicago: University of Chicago Press.

Bundesministerium für Bildung und Forschung. (2008). *Wissenschaftsjahr 2008. Mathematik: Alles, was zählt.* Abruf Juli 20, 2009, unter http://www.jahr-dermathematik.de/coremedia/generator/wj2008/de/b__Downloads/06__Presse/Imagebrosch_C3_BCre.pdf

Burt, C. L. & Howard, M. (1956). The multifactorial theory of inheritance and its application to intelligence. *British Journal of Statistical Psychology, 9*, 95–131.

Campbell, D. T. & Fiske, D W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105

Campbell, E. Q. & Kerckhoff, A. C. (1957). A critique of the concept: Universe of attributes. *The public opinion quarterly, 21*, 295-303.

Carroll, J. B. (1993). *Human Cognitive Abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Casé, L. R., Neer, R. & Lopetegui, S. (2003). Raven's progressive Matrices Test: scale construction and verification of "flynn effect". *Orientación y Sociedad*, 3, 1-11.

Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12*, 205-214.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York: Elsevier Science Pub. Co.

Cattell, R. B. (1998). Where is intelligence? Some answers form triadic theory. In J. J. McArdle (Ed.), *Human cognitive abilities in theory and practice* (pp. 29-38). Erlbaum: Mahwah.

Cattell, R. B., & Weiß, R. H. (1971). *Grundintelligenztest Skala 3 (CFT 3)*. Braunschweig: Westermann.

Chalmers, A. F. (2007). *Wege der Wissenschaft* (6. Aufl.). Springer: Berlin.

Champlain, A. D. & Gessaroli, M. E. (1996, April). *Assessing the Dimensionality of Item Response Matrices with small sample sizes and short Test lenghts*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J. & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*, 462-494.

Chernyshenko, O. S. & Stark, S. (2001). Investigating the hierarchical factor structure of the fifth-edition of the 16PF: An application of the Schmid-Leiman orthogonalization procedure. *Educational and psychological Measurement, 61*, 290-302.

Choi, S-Y. (1986). Application of Component Display Theory in Designing and Developing CALI. *CALICO Journal, 3*, 40-45.

Cizek, G. J., Webb, L. C. & Kalohn, J. C. (1995). The use of cognitive taxonomies in licensure and certification test development. *Evaluation & the Health Professions*, *18*, 77-91.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Hillsdale: Erlbaum.

Collins, L. M., Norman, C., McCormick, D. J. & Zatkin, J. L. (1986). Factor recovery in binary data sets: A simulation. *Multivariate behavioral research*, *21*, 377-391.

Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of applied psychology, 78*, 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Crone-Todd, D. E., Pear, J. J. & Read, C. N. (2000). Operational definitions for higher-order thinking objectives at the post-secondary level. *Academic Exchange Quarterly*, *4*, 99-106.

Davenport, E. C. & El-Sanhurry , N. A. (1991). Phi/phimax: Review and synthesis. *Educational and Psychological Measurement, 51*, 821-828.

Deary, I. J. (2000). *Looking down on human intelligence*. Oxford: University press.

Deary, I. J., Egan, V., Gibson, G. J., Austin, E. J., Brand, C. R. & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence, 23*, 105-132.

Debener, S. (2003). State-Trait-Angstinventar (STAI). In J. Hoyer & J. Margraf (Hrsg.), *Angstdiagnostik* (S. 161-163). Berlin: Springer.

Detterman, D. K. & Daniel, M. H. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low-IQ groups. *Intelligence, 13,* 349-359.

Deutsche Industrie und Handelskammer. (2006). Impulse für mehr Ausbildung. Abruf Mai 06, 2008, unter http://www.dihk.de/download.php?dload=http://www.dihk.de/inhalt/download/ausbildungsimpulse.pdf

Dlugosch, J., Englmaier, C., Götz, F-J. & Widl, J. (2006). *Mathematik 10 Wahlpflichtgruppe I*. Braunschweig: Westermann.

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *Journal of Experimental Education, 72*, 181-193.

Ehmke, T., Leiß, D., Blum, W. & Prenzel, M. (2006). Entwicklung von Testverfahren für die Bildungsstandards Mathematik. *Unterrichtswissenschaft*, *34*, 220-238.

Embretson, S. E. & Reise, S. P. (2000). *Item response Theory for Psychologists*. Mahwah: Lawrence Erlbaum.

Engel-Schermelleh, K. & Werner, C. (2008). Methoden der Reliabilitätsbestimmung. In: H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 113-133). Berlin: Springer.

Eysenck, M. W. & Keane, M. T. (2005). *Cognitive psychology: a Student's Handbook* (5th Ed.). Hove, UK: Psychology press.

Fairbrother, R. (1975). The reliability of teachers judgments of the abilities being tested by multiple choice items. *Educational research, 17*, 202-210.

Falmagne, J-C. (2005). Mathematical psychology – A perspective. *Journal of Mathematical Psychology, 49*, 436-439.

Feltes, T. & Paysen, M. (2005). *Nationale Bildungsstandards. Von der Bildungs- zur Leistungspolitik*. Hamburg: VSA.

Finney, S. J. & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R.O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269-314). Greenwich, CT: Information Age.

Fisch, E., Hylla, E. & Süllwold, F. (1965). *Rechentest RT 8+. Schulleistungstest für 8. und höhere Klassen*. Weinheim: Beltz.

Fogarty, G. J. & Stankov, L. (1995). Challenging the Law of Diminishing Returns. *Intelligence, 21*, 157–174.

Folin, O., Demis, W. & Smillie, W. G. (1914). Some observations on emotional glycosuria in man. *Journal of Biological Chemistry, 17*, 519-520.

Formann, A. K. & Piswanger, K. (1979). *WMT- Wiener Matrizentest. Ein Rasch-skalierter sprachfreier Intelligenztest*. Weinheim: PVU.

Fraser, C. & McDonald, R. P. (1988). NOHARM: least squares item factor analysis. *Multivariate Behavior Research*, *23*, 267-269.

Fraser, C. & McDonald, R. P. (2003). *NOHARM. A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Abruf Dezember 08, 2007, unter http://people.niagaracollege.ca/cfraser/download/

Fremer, J., Lohman, D. F. & Wittmann, W. W. (2002). *Evaluation of SWeSAT. The Swedish national aptitude test: A 25-year testing program. Current status and future development.* Stockholm: National Agency for Higher Education.

Frey, H. (1973). *Intelligenz und mathematische Leistung*. Freiburg: Herder.

Frings, C. (2002). Testrezensionen: VKI (Verbaler Kurzintelligenztest). In E. Brähler, H. Holling, D. Leutner & F. Petermann (Hrsg.), *Brickenkamp Handbuch psychologischer und pädagogischer Tests* (3. Aufl.) (S. 238-239). Göttingen: Hogrefe.

Gagné, R. M. (1984). Learning outcomes and their effects. *American Psychologist, 39*, 377-385.

Gärtner-Harnach, V. (1972). *Angst und Leistung*. Weinheim: PVU.

Gebert, A. (1977). Jäger's Phi (G) als Item-Interkorrelationsmaß für Faktorenanalysen. *Psychologische Beiträge*, *19*, 336-339.

Gierl, M. J. & Wang, C. (2005). Identifying content and cognitive Dimensions on the SAT. *College Board Research Report*, *11*, 1-31.

Golenia, J. & Neubert, K. (2007). *Mathematik 9M Bayern Hauptschule*. Braunschweig: Westermann.

Gorsuch, R. L. (1983). *Factor analysis* (2nd Ed.). Hillsdale, NJ: LEA.

Green, S. B. (1981). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7*, 139-147.

Green, S. B., Lissitz, R. W. & Mulaik, S. A. (1977). Limitations of Coefficient Alpha as an index of test unidimensionality. *Educational and psychological measurement*, 37, 827.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Revue*, *9*, 139-150.

Guttman, L. A. (1950). The Basis of Scalogram Analysis. In S.A. Stouffer, L. A. Gutmann, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and production* (pp. 60-90). Princeton, NJ: Princeton University Press.

Guttman, L. A. (1957). Empirical verification of the Radex structure of mental abilities and Personality Traits. *Educational and Psychological measurement*, *17*, 391-407.

Hall, R. J., Snell, A. F. & Singer, F. M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods, 2*, 233–256.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: SAGE.

Hamburger, F., Horstkemper, M. Melzer, W. & Tillmann, K-J. (2000). *Lehrpläne im Schulalltag. Eine empirische Studie zur Akzeptanz und Wirkung von Lehrplänen in der Sekundarstufe I.* Opladen: Leske & Budrich.

Hancock, G. R. & Mueller, R. O. (2006). *SEM A second course*. Greenwich, CT. Information Age.

Hartman, P. & Reuter, M. (2005). Spearman's law of diminishing returns tested with two methods. *Intelligence, 34*, 47-62.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, *19*, 49-78.

Hattie, J. (1985). Methodology Review: assessing unidimensionality of Tests and Items. *Applied psychological measurement*, *9*, 139-164.

Hattie, J., Krakwoski, K., Roger, H. & Swaminathan, H. (1996). An assessment of stout's index of essential unidimensionality. *Applied psychological Measurement*, *20*, 1-14.

Hays, W. L. (1994). *Statistics* (5th Ed.). Fort Worth, FL: Harcourt.

Hearnshaw, L. S. (1979). *Cyril Burt, Psychologist*. Cornell: University Press.

Hembree, R. (1988). Correlates, causes, and treatment of test anxiety. *Review of Educational Research*, *58*, 47-77.

Hill, P. W. & McGraw, B. (1981). Testing the Simplex assumption underlying Bloom's taxonomy. *American Educational Research Journal*, *18*, 93-101.

Hofe, R., Michael, K., Blum, W. & Pekrun, R. (2005). Zur Entwicklung mathematischer Grundbildung in der Sekundarstufe I – theoretische, empirische und diagnostische Aspekte. In M. Hasselhorn, W. Schneider & H. Marx (Hrsg.), *Diagnostik von Mathematikleistungen. Jahrbuch der pädagogisch-psychologischen Diagnostik, N. F. Band 4* (S. 263-292). Göttingen: Hogrefe.

Holling, H., Preckel, F. & Vock, M. (2004). *Intelligenzdiagnostik*. Göttingen: Hogrefe.

Hopmann, S. (2000). Lehrplan des Abendlandes – Abschied von seiner Geschichte? Grundlinien der Entwicklung von Lehrplan und Lehrplanarbeit seit 1800. In R. W. Keck & C. Ritzi (Hrsg.), G*eschichte und Gegenwart des Lehrplans*. Hohengehren: Schneider.

Hopman, S. T., Brinek, G. & Retzl, M. (2007). *PISA zufolge PISA – PISA according to PISA.* Wien: Lit.

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology*, *57*, 253-270.

Horn, J. & Noll, J. (1994). A System for understanding cognitive capabilities: A Theory and the evidence on which it is based. In D. K. Detterman (Ed.), *Current Topics in Human Intelligence* (pp. 151-205). Norwood: Alex Publishing.

Horn, W. (1983). *Leistungsprüfsystem (LPS)* (2. Aufl.). Göttingen: Hogrefe.

Horst, P. (1971). *Messung und Vorhersage*. Weinheim: Beltz.

Hu, L-T. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hu, L-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling, 6*, 1-55.

Hussy, W. (1998). *Denken und Problemlösen* (2. Aufl.). Stuttgart: Kohlhammer.

Hülsheger, U. R., Maier, G. W., Stumpp, T. & Muck, P. M. (2006). Vergleich kriteriumsbezogener Validitäten verschiedener Intelligenztests zur Vorhersage von Ausbildungserfolg in Deutschland: Ergebnisse einer Metaanalyse. *Zeitschrift für Personalpsychologie, 5,* 145-162.

Ibrahimovic, N. & Bulheller, S. (2005*). Mathematiktest. Grundkenntnisse für Ausbildung und Beruf*. Frankfurt: Harcourt.

Ingenkamp, K. (1962). D*ie deutschen Schulleistungstests*. Weinheim: Beltz.

Ingenkamp, K. (1964). *Psychologische Tests für die Hand des Lehrers*. Weinheim: Beltz.

Institut für Schulqualität und Bildungsforschung in München (2004). *Lehrplan für die bayerische Hauptschule, Kapitel III-Teil II Jahrgangsstufe M9.* Abruf Dezember 01, 2008, unter http://www.isb.bayern.de/isb/download.aspx?DownloadFileID=33db40989fabac183eeb0bf50c28c6d8

Institut für Schulqualität und Bildungsforschung in München (2005). *KMK-Bildungsstandards. Konsequenzen für die Arbeit an bayerischen Schulen*. Abruf Dezember 01, 2008, unter http://www.isb.bayern.de/isb/download.aspx?DownloadFileID=507c5c4c9dd580b1c53f22b10a1f3406

International Association for the Evaluation of Educational Achievement. (2000). *TIMSS 1999 International Mathematics report.* Boston: TIMSS & PIRLS International Study Center.

International Association for the Evaluation of Educational Achievement. (2004a). *TIMSS 2003 Technical Report.* Boston: TIMSS & PIRLS International Study Center.

International Association for the Evaluation of Educational Achievement. (2004b). *TIMSS 2003 international mathematics report*. Boston: TIMSS & PIRLS International Study Center.

International Association for the Evaluation of Educational Achievement. (2005). *TIMSS IEA's TIMSS 2003 International report on achievement in the mathematics cognitive domains*. Boston: TIMSS & PIRLS International Study Center.

International Association for the Evaluation of Educational Achievement. (2008). *TIMSS 2007 International mathematics report*. Boston: TIMSS & PIRLS International Study Center.

Jablonka, E. (2005). Mathematical literacy. Die Verflüchtigung eines ambitionierten Testkonstrukts in bedeutungslose PISA Punkte. In T. Jahnke. und W. Meyerhöfer (Hrsg.), *Pisa & Co Kritik eines Programms* (S. 247-280). Berlin: Franzbecker.

Jacobs, C. & Petermann, F. (2007). Testbesprechungen: Wechsler Intelligenztest für Erwachsene. Zeitschrift für Psychiatrie. *Psychologie und Psychotherapie*, *55*, 205-210.

Jaeggi, S. M., Buschkuehl, M., Jonides, K.& Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. In E. E. Smith (Ed.), *Procedings of the National Academy of Sciences of the United States of America*, *105*, 1-5.

Jäger, A-O. (1967). *Dimensionen der Intelligenz*. Göttingen: Hogrefe.

Jäger, A-O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, *28*, 195-225.

Jäger, A. O. & Althoff, K. (1983). *Der WILDE-Intelligenz-Test.* Göttingen: Hogrefe.

Jäger, A. O., Süß, H-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur- Test. BIS-Test.* Göttingen: Hogrefe.

Jäger, R. S. (1997). *WILDE-Intelligenz-Test (WIT)*. Zeitschrift für Differentielle und *Diagnostische Psychologie*, *18*, 62-63

Jankisz, E. & Moosbrugger, H. (2008). Item-response-theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 28-71). Berlin: Springer.

Jasper, F. (2007). *Modellkontrolle, Konstruktvalidierung und Weiterentwicklung einer rasch-homogenen Skala auf Basis von Bongard-Problemen*. Unveröffentlichte Diplomarbeit, Universität Mannheim.

Jasper, F. & Wagener, D. (in Druck). *Start-M: Mathematik. Testbatterie für Berufseinsteiger.* Göttingen: Hogrefe.

Jensen, A. R. (1998). *The* g *factor: The science of mental ability*. Westport, CT: Praeger.

Johnson-Laird, P. N. (1980). Mental Models in cognitive science. *Cognitive science*, *4*, 71-115.

Julie, C. (2006). Mathematical Literacy: Myths, further inclusions and exclusions. *Pythagoras*, *64*, 62-69.

Jung, C., Kempf, M. & Seggewiß, B. (2007). *Bericht über die Entwicklung und Verbesserung eines Mathematiktests für Auszubildende 2007. Erstellt im Rahmen des experimentellen Praktikums 2007 unter der Leitung von Dr. Wagener*. Unveröffentlichter Bericht aus dem Empiriepraktikum des psychologischen Instituts der Universität Mannheim.

Kaiser, G. & Schwarz, I. (2003). Mathematische Literalität unter einer kulturell-sprachlichen Perspektive. *Zeitschrift für Erziehungswissenschaft*, *6*, 356-376.

Kersting, M., Althoff, K. & Jäger, A. O. (2008). *Wilde-Intelligenz-Test 2.* Göttingen: Hogrefe.

Kishton, J. M. & Widaman, K. F.(1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757–765.

Klieme, E. Avenarius, H. Blum, W., Döbrich, P. G. ,Prenzel, H-M. Reiss, K. Riquarts, Rost, K-J.,Tenorth, H-E., & Vollmer, H-J. (2003). *Zur Entwicklung nationaler Bildungsstandards*. Bonn: BMBF.

Kline, P. (2000). *A Psychometrics Primer*. London: Free Association Books.

Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd Ed.). New York: Guilford press.

Krathwohl, D. R. (1994). Reflections on the Taxonomy: Its' past, present, and future. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's Taxonomy. A Forty-Year Retrospective* (pp. 64-82). Chicago: Chicago Press.

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into Practice*, *41*, 212-218.

Krathwohl, D. R., Bloom, B. S. & Masia, B. B. (1964). *Taxonomy of educational objectives, Book 2: Affective domain*. New York: Longman.

Kraus, J. (2005). *Der PISA-Schwindel.* Wien: Signum.

Kreitzer, A. E. & Madaus, G. F. (1994). Empirical investigations of the hierarchical structure of the taxonomy. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's Taxonomy. A Forty-Year Retrospective* (pp. 64-82). Chicago: Chicago Press.

Krohne, H-W. & Hock, M. (2007). *Psychologische Diagnostik*. Stuttgart: Kohlhammer.

Kropp, R. P. & Stocker, H. W. (1966). *The construction and validation of tests of the cognitive processes as described in the taxonomy of educational objectives* (cooperative research projekt No. 2117). Flordia State University, Institute of Human Leaning.

Kubinger, K. D. (Hrsg.). (1988). *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen.* München: PVU.

Kubinger, K. D. (2000). Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. *Psychologische Rundschau, 51*, 33-34

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, *45*, 106-110.

Kultusministerkonferenz. (2004a). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss (Jahrgangsstufe 10)*. München: Luchterhand.

Kultusministerkonferenz. (2004b, April). Dokumentation der Fachtagung der Kultusministerkonferenz „Implementation der Bildungsstandards“ am 2.4.2004 im Berliner Landesinstitut für Schule und Medien. Abruf Februar 07, 2008, unter http://www.kmk.org/index.php?id=1584&type=123

Kultusministerkonferenz. (2005a). *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss (Jahrgangsstufe 9)*. München: Luchterhand.

Kultusministerkonferenz. (2005b). *Bildungsstandards der Kultusministerkonferenz. Erläuterung zu Konzeption und Entwicklung*. Münchnen: Luchterhand.

Lance, C. E., Butts, M. & Michels, L. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*, 202-220.

Laux, L., Glanzmann, P., Schaffner, P. & Spielberger, C. D. (1981). *Das State-Trait-Angstinventar*. Göttingen: Hogrefe.

Leibniz-Institut für die Pädagogik der Naturwissenschaften. (1998). *Testaufgaben Mathematik TIMSS 7./8. Klasse (Population 2).*Berlin: Max-Planck-Institut für Bildungsforschung.

Leibniz-Institut für die Pädagogik der Naturwissenschaften. (2000). *TIMSS/III-Deutschland - der Abschlussbericht.* Abruf April 05, 2008 unter http://www.timss.mpg.de/TIMSS_im_Ueberblick/TIMSSIII-Broschuere.pdf

Lienert, G. A. & Hofer, M. (1972). *Mathematiktest für Abiturienten und Studienanfänger.* Göttingen: Hogrefe.

Lienert, G. & Raatz, U. (1994). *Testaufbau und Analyse* (5. Aufl.). Weinheim: PVU.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenzstrukturtest. I-S-T 2000 R* (2. Aufl.). Göttingen: Hogrefe.

Lipscomb, J. W. (1985). Is Bloom's taxonomy better than intuitive judgment for classifying test questions. *Education*, *106*, 102–107.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173.

Loehlin, J. C. (2004). *Latent variable models* (4th Ed.). Mahwah: LEA.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lorenz, M. & Rohrschneider, U. (2009). *Erfolgreiche Personalauswahl.* Springer: Berlin.

Lubinski, D., Webb, R. M., Morelock, M. & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year Follow-Up of the Profoundly Gifted. *Journal of Applied Psychology, 86*, 718-729.

MacCallum, R. C., Brown, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130-149.

Marden, J., Roussos, L. A. & Stout, W. F. (1998). Simulation study of the effectiveness of using new proximity measures with hierarchical cluster analysis to detect dimensionality. *Journal of Educational Measurement, 35,* 1-30.

Marsh, H. W., Hau, K-T. & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.

Masters, G. N. (1982). A rasch-model for partial credit scoring, *Psychometrika, 47*, 149-174.

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika, 27,* 397-415.

McDonald, R. P. (1967). A comparision of four methods of constructing factor scores. *Psychometrika, 32*, 381-401.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 110-117.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258–271). New York: Springer.

McDonald, R. P. (1999). *Test Theory. A unified treatment*. Mahwah: LEA.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin, 43*, 289-374.

Meara, K., Robin, F. & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral research, 35*, 229-259.

Merrill, M. D. (1983). Components Display Theory. In C. M. Reigeluth (Ed.), *Instructional-Design Theories and Models: An Overview of their Current Status* (pp. 279-333). Hillsdale: London.

Merrill, M. D. (1999). Instructional Transaction Theory (ITT): Instructional design based on knowledge objects. In C. M. Reigeluth (Ed.), *Instructional-Design Theories and Models: A New Paradigm of instructional theory* (pp. 397-425). Hillsdale: London.

Moosbrugger, H. (2008). Item-response-theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 215-261). Berlin: Springer.

Moosbrugger, H. & Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7-27). Berlin: Springer.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. London: CRC Press.

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-243). Newbury Park, CA: Sage.

Muthén, B., du Toit, S. H-C. & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L. K. & Muthén, B. O. (2007). *Mplus User's Guide* (5th Ed.). Los Angeles, CA: Muthén & Muthén.

Muthny, F.A. (1997). Testrezension zu State-trait-Angstinventar (STAI). *Zeitschrift für Differentielle und Diagnostische Psychologie*, *18*, 72-73

Nunnally, J. *C. &* Bernstein, I. H. (*1994*). *Psychometric theory* (3rd Ed.). New York: McGraw-Hill.

Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items-comparison of different approaches. *Journal of Educational Measurement, 31*, 1-18.

Nandakumar, R. & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 93-106). Thousand Oak, CA: SAGE.

Nandakumar, R. & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics, 18*, 41-68.

Nasser, F. & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement, 63*, 729-757.

Näsström, G. & Henriksson, W (2008). Alignment of Standards and Assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology*, *6*, 667-690.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J. et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Neumann, P. (2006). DIHK: Viele Schulabgänger sind unqualifiziert. *Die Welt Online*. Abruf Mai 05, 2008, unter http://www.welt.de/print-welt/article224430/DIHK_Viele_Schulabgaenger_sind_unqualifiziert.html

Nichols, D. P. (1998). Chosing an intraclass correlation coefficient. *SPSS Keywords, 67,* Abruf Juli 23, 2009 unter http://www.ats.ucla.edu/stat/spss/library/whichicc.htm

Niedersächsisches Kultusministerium. (2006a). *Kerncurriculum für die Hauptschule. Schuljahrgänge 5-10*. Hannover: Unidruck.

Niedersächsisches Kultusministerium. (2006b). *Kerncurriculum für die Realschule. Schuljahrgänge 5-10*. Hannover: Unidruck.

Niedersächsisches Kultusministerium .(2006c). *Kerncurriculum für das Gymnasium. Schuljahrgänge 5-10*. Hannover: Unidruck.

Nordrhein-Westfalen: Ministerium für Schule, Jugend und Kinder des Landes NRW. (2004a). *Kernlehrplan für die Hauptschule in Nordrhein-Westfalen.* Frechen: Ritterbach.

Nordrhein-Westfalen: Ministerium für Schule, Jugend und Kinder des Landes NRW. (2004b). *Kernlehrplan für das die Realschule in Nordrhein-Westfalen*. Frechen: Ritterbach.

Nordrhein-Westfalen: Ministerium für Schule, Jugend und Kinder des Landes NRW. (2007). *Kernlehrplan für das Gymnasium - Sekundarstufe I (G8) in Nordrhein-Westfalen*. Frechen: Ritterbach.

OECD. (2001). *PISA 2000 Zusammenfassung zentraler Befunde*. Abruf März 08, 2008, unter http://www.mpib-berlin.mpg.de/pisa/ergebnisse.pdf

OECD. (2001). *PISA 2000 Beispielaufgaben aus dem Mathematiktest*. Abruf März 03, 2009, unter http://www.mpib-berlin.mpg.de/pisa/beispielaufgaben_mathematik.pdf

OECD. (2003). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Abruf Januar 09, 2008, unter http://www.oecd.org/dataoecd/46/14/33694881.pdf

OECD. (2005). *PISA 2003 technical report*. Abruf Mai 15, 2008, unter http://www.oecd.org/dataoecd/49/60/35188570.pdf

OECD. (2007). *Science competencies for tomorrow's world Executive Summary*. Abruf Januar 03, 2008, unter http://www.oecd.org/dataoecd/15/13/39725224.pdf

OECD. (2009). *PISA 2006 technical report*. Waschington: OECD Publishing.

Oettinger, G. H. (2008). *Regierungserklärung zur Qualitätsoffensive Bildung von Ministerpräsident Günther H. Oettinger vor dem Landtag von Baden-Württemberg am 23. Juli 2008*. Abruf März 03, 2009, unter http://www.sm.baden-wuerttemberg.de/fm7/ 1899/ 080723_Regierungserklaerung_Oettinger_Bildungsoffensive.pdf

Olsen, R. V. (2005). *Achievement Tests from an Item perspective.* Dissertation, Universität Oslo, Department of Teacher Education and School Development.

Orth, U. (2006). *Kurzbericht über die Entwicklung eines Mathematiktests für Auszubildende.* Unveröffentlichter Bericht aus dem Empiriepraktikum des psychologischen Instituts der Universität Mannheim.

Payk, B. (2009). *Deutsche Schulpolitik nach dem PISA-Schock: Wie die Bundesländer auf die Legitimationskrise des Schulsystems reagieren*. Hamburg: Kovac.

Pearson , K. & Herron, D. (1913). On theories of association. *Biometrika, 9*, 159-315.

Petermann, F. & Petermann, U. (2008). *Hamburg-Wechsler-Intelligenztest für Kinder IV (HAWIK-IV)* (2. Aufl.). Göttingen: Hogrefe.

Pisa-Konsortium Austria. (2009). *Mathematik-Kompetenz. Sammlung aller bei PISA freigegebenen Aufgaben der Haupttests 2000, 2003 und 2006*. Abruf Juli 01, 2009, unter http://www.bifie.at/sites/default/files/items/PISA-Mathematik.pdf

Pisa-Konsortium Deutschland. (2003). *PISA 2003: Ergebnisse des zweiten Ländervergleichs. Zusammenfassung*. Abruf März 09, 2008, unter http://pisa.ipn.uni-kiel.de/PISA2003_E_Zusammenfassung.pdf

Pisa-Konsortium Deutschland. (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster: Waxmann.

Prenzel, M., Walter, O. & Frey, A. (2007). PISA misst Kompetenzen. Eine Replik auf Rindermann (2006). Was messen internationale Schulleistungsstudien? *Psychologische Rundschau, 58*, 128-136.

Raatz, U. (1980). Kritische Bemerkungen zur Anwendung von multiple-choice-Aufgaben in Mathematiktests. *Lernzielorientierter Unterricht*, *1*, 25-30.

Raykov, T. & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd Ed.). Mahwah: LEA.

Reid, W. A. (1998). Erasmus, Gates and the end of curriculum. *Journal of Curriculum Studies, 30*, 499-501.

Reigeluth, C. M. & Carr-Chellman, A. A. (2009). Situational principles of instruction. In C. M. Reigeluth & A. Carr-Chellman (Eds*.), Instructional-Design Theories and Models, Volume III: Building a Common Knowledge Base* (pp. 57-68). New York: Routledge.

Reigeluth, C. & Moore, J. (1999). Cognitive Education and the Cognitive Domain. In: C. M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (pp. 51-67). Hillsdale: New York.

Rijsdijk, F. V., Vernon, P. A. & Boomsma, D. I. (2002). Application of hierarchical genetic models to raven and wais subtests: A dutch twin study. *Behavior Genetics, 32*, 199-210.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? *Psychologische Rundschau*, 57, 69-86.

Rogers, W. M. & Schmitt, N. (2004). Parameter recovery and model fit using multidimensional composites: A comparison of four empirical parceling algorithms. *Multivariate Behavioral Research, 39*, 379-412.

Rost, J. (1999) Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau, 50,* 140-156.

Rost, J. (2004a). *Testtheorie und Testkonstruktion* (2. Aufl.). Bern: Hans Huber.

Rost, J. (2004b). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogische Psychologie*, *50*, 662-678.

Rost, J. (2007). Zur Messung von Kompetenzen einer Bildung für nachhaltige Entwicklung. In I. Bormann & G. Haan (Hrsg.), *Kompetenzen der Bildung für nachhaltige Entwicklung* (S. 61-73). Wiesbaden: VS Verlag.

Roussos, L. A. & Ozbek, O. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement, 43*, 215 – 243.

Rust, R. T., Lee, C. & Valente, E. (1995). Comparing covariance structure models: A general methodology. *International Journal of Research in Marketing, 12*, 279-291.

Saklofse, D. H., Yan, Z., Zhu, J. & Austin, E. J. (2008). Spearman's law of diminishing returns in normative samples for the WISC-IV and WAIS-III. *Journal of Individual Differences, 29*, 57-69.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity, 24,* 367-386.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: sage.

Schmid, J. & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Schmidt-Atzert, L. (2002). Rezension zum Intelligenz-Struktur-Test 2000 R. *Zeitschrift für Personalpsychologie*, *1,* 50-56.

Schmidt-Atzert, L., Deter, B. & Jaeckel, S. (2004). Prädiktion von Ausbildungserfolg: Allgemeine Intelligenz (g) oder spezifische kognitive Fähigkeiten? *Zeitschrift für Personalpsychologie*, *3*, 147-158.

Schnotz, W. & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*, 141-156.

Schuler, H., Hell, B., Trapmann, S., Schaar, H. & Boramir, I. (2007). Die Nutzung psychologischen Verfahren der externen Personalauswahl in deutschen

Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie*, *6*, 60-70.

Schumacker, R. E. & Lomax, R. G. (2004). *A beginner's guide to structural Equation modeling* (2nd Ed.). Mahwa: LEA.

Schweizer, K. (2006). *Leistung und Leistungsdiagnostik*. Heidelberg: Springer.

Schweizerischer Verband für Berufsberatung. (2006).Wechsler Intelligenztest für Erwachsene (WIE). *Deutschsprachige Bearbeitung und Adaption des WAIS-III von David Wechsler.* Abruf Februar 04, 2009, unter: http://www.testraum.ch/Serie%209/def_WIE.pdf.

Seddon, G. M. (1978). The properties of Bloom's taxonomy of educational objectives for the cognitive domain. *Review of educational research*, *48*, 303-323.

Seipp, B. (1990). *Angst und Leistung in Schule und Hochschule. Eine Meta-Analyse.* Frankfurt: Lang.

Seraphine, A. E. (2000). The performance of dimtest when latent trait and item difficulty distributions differ. *Applied psychological Measurement, 24,* 82-94.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428

Siegel, S. (1956). *Nonparametric Statistics for the behavioral sciences.* New York: McGraw-Hill.

Solman, R. & Rosen, G. (1986). Bloom's six cognitive levels represent two levels of performance. *Educational psychology, 6*, 243-263.

Solso, R. L. MacLin, M. K. & MacLin, O. H. (2005). *Cognitive Psychology* (7th Ed.). Boston: Pearson.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Spearman, C. (1927). *The abilities of man*. MacMillan: New York.

Spielberger, C. D., Gorsuch, R. L. & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consultin Psychologists Press.

Steer, R. A. (2009). Amount of general factor saturation in the Beck Anxiety Inventory responses of outpatients with anxiety disorders. *Journal of Psychopathological Behavior assessment, 31*, 112-118.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251.

Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all (Commentary). In E. E. Smith (Ed.), *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6791-6792.

Sternberg, R. J. & Powell, J. S. (1982). Metatheory of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 975-1027). Cambridge: University press.

Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing Research. *Journal of Marketing research*, *18*, 51-62.

Stookey, J. A. & Baer, M. A. (1976). A critique of Guttman scaling: With special attention to its application to the study of collegial bodies. *Quality and Quantity*, *19*, 251-260.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589-617.

Süß, H. M. (2001). Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich. In E. Stern & J. Guthke (Hrsg.), *Perspektiven der Intelligenzforschung* (S. 109-136). Lengerich: Pabst.

Süß, H-M. (2003) Intelligenztheorien. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der psychologischen Diagnostik* (S. 217-224). Weinheim: PVU.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

Tewes, U., Rossmann, P. & Urs, S. (1999). *Wechsler Intelligence Scale for Childern Hamburg-Wechsler Intelligenztest für Kinder* (3. Aufl.). Bern: Huber.

Thompson, B. & Vidal-Brown, S. A. (2001, Februar). *Principle components versus principle axis factors: when will we ever learn?* Annual meeting of the southwest educational research association. New Orleans.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: Chicago Press.

Thurstone, L. L. (1944). Second-Order Factors, *Psychometrika*, *9*, 71-100.

Thurstone, L. L. (1952). L. L. Thurstone. In E. G. Boring, H. S. Langfeld, H. Werner, & R. M. Yerkes (Eds.), *A history of psychology in autobiography, Vol. IV*, (pp. 295-321). Worcester, MA: Clark University Press.

Trochim, W. & Donnelly, D. (2006). *The research methods knowledge base* (3rd. Ed.). Mason: Atomic Dog.

Tullock, G. A (2001). Comment on Daniel Klein's: A plea to economists who favor liberty. *Eastern Economic Journal, 27*, 203-207.

Überla, K. (1977). *Faktorenanalyse* (2. Aufl.). Springer: Berlin.

Velicer, W. F. & Jackson, D. N. (1990). Component analysis versus common factor analysis. Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1-28.

Vernon, P. E. (1979). *Intelligence: Heredity and environment*. San Francisco: W. H. Freeman & Company.

Wacker, A. (2008). *Bildungsstandards als Steuerungsinstrumente der Bildungsplanung. Eine empirische Studie zur Realschule in Baden-Württemberg*. Bad Hilbrunn: Julius Klinkhardt.

Wagener, D. (2008). *Start-C: Computerwissen. Testbatterie für Berufseinsteiger*. Göttingen: Hogrefe.

Wechsler, D. (1961). *Die Messung der Intelligenz Erwachsener* (2. Aufl.). Bern: Hans Huber.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In: F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen.* Weinheim: Beltz.

Wilson, J. W. (1970). Evaluation of Learning in Secondary School Mathematics. In B. S. Bloom, J. T. Hastings & G. F. Madaus (Hrsg.), *Handbook of formative and summative evaluation of student learning* (S. 643-697). New York: McGraw Hill.

Witte, E. H. & Caspar, F. M. (1976). Zur Identifizierbarkeit von Schwierigkeitsfaktoren. *Diagnostica, 22*, 126-138.

Wittmann, W. W. (1985). *Evaluationsforschung*. Berlin: Springer.

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd Ed.) (pp. 505-560). New York: Plenum.

Wittmann, W. W. (2004). Group differences in intelligence and related measures. In O. Wilhelm & W. R. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 223-241). Thousand Oaks: Sage.

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393-409.

Wittmann, W. W. & Süß, H-M. (1997, Juli). Challenging g-Mania in intelligence research: Answers not given, due to questions not asked. In R. D. Robert & P.

Kyllonen (Chair), *New directions in ability research*. Symposium der International Society for the Study of Individual Differences, Aarhus, Dänemark.

Witz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Wolff, H-G. & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-leiman solution. Syntax codes for SPSS and SAS. *Behavior Research Methods, 37*, 48-58.

Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the rasch model. *Applied Psychological Measurement, 6*, 83-91.

Worthington, R. L. & Whittaker, T. A. (2006). Scale development research. *The Counseling Psychologist, 34*, 806-838.

Wu, M. (2009). *A critical comparison of the contents of PISA and TIMSS mathematics assessments*. Abruf Juli 05, 2009, unter https://edsurveys.rti.org/PISA/documents/WuA_Critical_Comparison_of_the_Contents_of_PISA_and_TIMSS_psg_WU_06.1.pdf

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *Acer Conquest Version 2.0.* Melbourne: ACER.

Wuttke, J. (2007). Uncertainities and Bias in Pisa. In S. T. Hopman, G. Brinek & M. Retzl (Hrsg.), *PISA zufolge PISA – PISA according to PISA*. (S. 241-264). Wien: Lit.

Zeidner, M. (1998). *Test anxiety*. Berlin: Springer.

Zhang, J. & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.

12 Anhang

12.1 Reanalyse des Expra-Tests

12.1.1 Klassische Kennwerte aller Items

Tabelle 47 Klassische Kennwerte aller Items des Expra-Tests, N = 182.

Item	Mittelwert	$r_{it\_korrigiert}$	Cronbach's α nach Ausschluss
A8_D	0,29	0,57	0,87
A8_C	0,37	0,54	0,87
A4_B	0,65	0,53	0,87
A2_E	0,58	0,49	0,87
A3_E	0,28	0,48	0,87
A3_C	0,30	0,47	0,87
A9_D	0,45	0,45	0,87
A4_D	0,47	0,44	0,87
A8_A	0,64	0,42	0,87
A8_B	0,64	0,42	0,87
A9_F	0,21	0,41	0,87
A9_E	0,44	0,41	0,87
A7_C	0,26	0,40	0,87
A9_H	0,16	0,40	0,87
A3_D	0,36	0,40	0,87
A7_B	0,20	0,40	0,87
A4_E	0,41	0,39	0,87
A2_C	0,64	0,38	0,87
A4_C	0,57	0,37	0,87
A5_C	0,76	0,35	0,87
A9_I	0,06	0,34	0,87
A5_D	0,15	0,34	0,87
A9_K	0,05	0,34	0,87
A2_F	0,19	0,33	0,87
A4_A	0,69	0,32	0,87
A7_A	0,76	0,31	0,87
A6_A	0,75	0,31	0,87
A2_A	0,78	0,31	0,87
A9_A	0,80	0,30	0,87
A9_C	0,34	0,30	0,87
A6_B	0,81	0,30	0,87
A5_B	0,89	0,30	0,87
A2_B	0,79	0,29	0,87
A2_H	0,07	0,26	0,87
A2_G	0,15	0,26	0,87
A3_B	0,56	0,26	0,87
A5_A	0,95	0,25	0,87
A1_B	0,47	0,24	0,87
A1_D	0,12	0,23	0,87

Item	Mittelwert	$r_{it\_korrigiert}$	Cronbach's α nach Ausschluss
A1_C	0,25	0,23	0,87
A9_B	0,48	0,23	0,87
A6_C	0,66	0,23	0,87
A2_D	0,47	0,21	0,87
A2_I	0,06	0,21	0,87
A3_A	0,92	0,19	0,87
A5_E	0,03	0,19	0,87
A6_D	0,21	0,18	0,87
A9_G	0,12	0,10	0,87
A1_A	0,69	0,07	0,88
A7_D	0,02	0,03	0,87
A9_J	0,10	0,00	0,87

12.1.2 NOHARM Lösung 3-Faktoren, explorativ

Varimax Rotated Factor Loadings

	1	2	3
1	0.488	-0.050	0.092
2	0.487	0.117	-0.033
3	0.451	0.054	0.200
4	0.177	0.048	0.668
5	0.334	-0.221	0.696
6	0.081	0.221	0.669
7	-0.014	0.051	0.475
8	0.075	0.416	0.722
9	0.297	0.050	0.508
10	0.050	0.259	0.446
11	0.129	0.391	0.377
12	-0.029	0.340	0.490
13	0.026	0.492	0.240
14	0.389	0.094	0.106
15	0.489	0.316	0.307
16	0.288	0.210	0.456
17	0.620	0.396	0.147
18	0.109	0.442	0.272
19	0.404	0.589	0.258
20	0.396	0.352	0.156
21	0.095	0.655	0.276
22	0.071	0.669	0.192
23	0.694	0.126	0.276
24	0.639	0.275	0.012
25	0.476	0.244	0.124
26	0.295	0.147	0.552
27	0.401	0.066	0.578
28	0.300	0.045	0.450
29	0.509	0.155	0.152

30	0.498	-0.013	0.006
31	0.041	0.250	0.193
32	-0.031	0.563	0.289
33	0.244	0.365	0.443
34	0.222	0.306	0.522
35	-0.031	0.130	0.108
36	0.374	0.811	-0.225
37	0.349	0.899	-0.265
38	0.333	0.790	0.081
39	0.421	0.748	0.200
40	0.392	0.151	0.214
41	0.010	0.245	0.311
42	0.207	0.166	0.390
43	0.460	0.279	0.272
44	0.689	0.089	0.203
45	0.453	0.273	0.346
46	0.477	-0.257	-0.035
47	0.557	0.053	0.427
48	0.658	0.475	0.141
49	0.377	0.485	0.413

Promax (oblique) Rotated Factor Loadngs

	1	2	3
1	0.506	-0.096	0.028
2	0.505	0.153	-0.160
3	0.429	-0.034	0.152
4	0.044	-0.278	0.796
5	0.245	-0.589	0.849
6	-0.084	-0.086	0.783
7	-0.119	-0.181	0.590
8	-0.130	0.105	0.815
9	0.203	-0.194	0.571
10	-0.077	0.067	0.503
11	4.9e-4	0.250	0.376
12	-0.181	0.134	0.558
13	-0.095	0.429	0.206
14	0.377	0.056	0.039
15	0.408	0.205	0.231
16	0.181	0.010	0.477
17	0.568	0.376	-0.012
18	-0.006	0.359	0.239
19	0.287	0.533	0.136
20	0.334	0.320	0.053
21	-0.054	0.593	0.208
22	-0.064	0.650	0.104
23	0.660	0.011	0.186
24	0.633	0.309	-0.162

25	0.443	0.217	0.016
26	0.178	-0.108	0.608
27	0.296	-0.209	0.635
28	0.220	-0.170	0.497
29	0.485	0.103	0.062
30	0.528	-0.011	-0.089
31	-0.033	0.184	0.188
32	-0.176	0.483	0.266
33	0.113	0.189	0.441
34	0.083	0.084	0.556
35	-0.075	0.091	0.118
36	0.320	1.020	-0.505
37	0.289	1.139	-0.568
38	0.218	0.845	-0.110
39	0.293	0.739	0.031
40	0.349	0.066	0.164
41	-0.090	0.118	0.344
42	0.115	-0.007	0.418
43	0.390	0.182	0.199
44	0.675	0.007	0.102
45	0.368	0.137	0.295
46	0.551	-0.263	-0.091
47	0.495	-0.147	0.417
48	0.597	0.467	-0.040
49	0.243	0.339	0.354

Factor Correlations

	1	2	3
1	1.000		
2	0.256	1.000	
3	0.395	0.573	1.000

Sum of squares of residuals (lower off-diagonals) = 0.1451449

Root mean square of residuals (lower off-diagonals) = 0.0111096

Tanaka index of goodness of fit = 0.9252611

12.2 Itembennenungen in allen Testformen

Tabelle 48 Für jede Aufgabe ist abgetragen in welchem Test sie auftaucht und wie sie dort heißt.

Nummerierung Endform	Benennung Endform	Expra	Vorform	Skalen-zuordnung
1	A1		B32	GEO
2	A2	A9_D	B22	GEO
3	A3	A9_B	B17	GEO
4	A4	A9_A	B13	GEO
5	A5		A19	GEO
6	A6a		B29a	GEO
7	A6b		B29b	GEO
8	A6c		B29c	GEO
9	A7	A9_F	B15	GEO
10	A8		B21	GEO
11	A9	A9_H	B19	GEO
12	A10a		B28a	GEO
13	A10b		B28b	GEO
14	A10c		B28c	GEO
15	A11a		B27a	GEO
16	A11b		B27b	GEO
17	A11c		B27c	GEO
18	A11d		B27d	GEO
19	A12a		B25a	GEO
20	A12b		B25b	GEO
21	A13a	A3_A	A3A	PROZ
22	A13b	A3_B	A3B	PROZ
23	A13c	A3_E	A3E	PROZ
24	A14a	A2_A	A2A	PROZ
25	A14b	A2_D	A2D	PROZ
26	A14c	A2_E	A2E	PROZ
27	A15a		A6a	PROZ
28	A15b		A6b	PROZ
29	A15c		A6c	PROZ
30	A15d		A6d	PROZ
31	A16a		A5b	PROZ
32	A16b		A5c	PROZ
33	A16c		A5d	PROZ
34	A17a[1]			PROZ
35	A17b[1]			PROZ
36	A18		A20	PROZ
37	A19		A22	PROZ
38	A20a		A10b	PROZ

Nummerierung Endform	Benennung Endform	Expra	Vorform	Skalen-zuordnung
39	A20b		A10c	PROZ
40	A21a		A8a	PROZ
41	A21b		A8c	PROZ
42	A22a	A4_A	A4A	PROZ
43	A22b	x	A4F	PROZ
44	A22c	x	A4G	PROZ
45	A23a		A13b	PROZ
46	A23b		A13c	PROZ
47	A23c		A13d	PROZ
48	A24a		A12a	PROZ
49	A24b		A12b	PROZ
50	A24c		A12c	PROZ
51	A24d		A12d	PROZ
52	A25B		B5B	LIT
53	A25C		B5C	LIT
54	A25D		B5D	LIT
55	A26a		B10a	LIT
56	A26b		B10b	LIT
57	A26c		B10c	LIT
58	A27A	A6_A	B2A	LIT
59	A27B	A8_A	B4A	LIT
60	A27C	A8_C	B4C	LIT
61	A27D	A6_B	B2B	LIT
62	A27E	A8_D	B4D	LIT
63	A28		B7	LIT
64	A29		B12	LIT
65	A30a		A27a	LIT
66	A30b		A27b	LIT
67	A31a	A5_C	A9C	KOMPL
68	A31b	A5_E	A9E	KOMPL
69	A31c	x	A9F	KOMPL
70	A31d	x	A9G	KOMPL
71	A32a		A11a	KOMPL
72	A32b		A11b	KOMPL
73	A33		A18	KOMPL
74	A34a		A17a	KOMPL
75	A34b		A17b	KOMPL
76	A34c		A17c	KOMPL
77	A35		A16c	KOMPL
	x		A10a	PROZ
	x		A13a	PROZ

Nummerierung Endform	Benennung Endform	Expra	Vorform	Skalen-zuordnung
	x		A13e	PROZ
	x		A14a	PROZ
	x		A14b	PROZ
	x		A14c	PROZ
	x		A14d	PROZ
	x		A15	PROZ
	x		A16a	KOMPL
	x		A16b	KOMPL
	x		A1a	PROZ
	x		A1b	PROZ
	x		A1c	PROZ
	x		A1d	PROZ
	x		A21a	GEO
	x		A21b	GEO
	x		A21c	GEO
	x		A23[2]	PROZ
	x	A9_G	A24	GEO
	x	A9_K	A25	GEO
	x		A26a	LIT
	x		A26b	LIT
	x		A28	LIT
	x		A29	GEO
	x	A2_B	A2B	PROZ
	x	A2_C	A2C	PROZ
	x	A2_F	A2F	PROZ
	x	A2_G	A2G	PROZ
	x	A2_H	A2H	PROZ
	x	A2_I	A2I	PROZ
	x	x	A2J	PROZ
	x		A30	GEO
	x		A31	GEO
	x	A3_C	A3C	PROZ
	x	A3_D	A3D	PROZ
	x	A4_B	A4B	PROZ
	x	A4_C	A4C	PROZ
	x	A4_D	A4D	PROZ
	x	A4_E	A4E	PROZ
	x		A5a	PROZ
	x		A7a	PROZ
	x		A7b	PROZ
	x		A7c	PROZ

Nummerierung Endform	Benennung Endform	Expra	Vorform	Skalen-zuordnung
	x		A7d	PROZ
	x		A8b	PROZ
	x	A5_A	A9A	PROZ
	x	A5_B	A9B	PROZ
	x	A5_D	A9D	PROZ
	x		B10d	LIT
	x		B11a	GEO
	x		B11b	GEO
	x	A9_C	B14	GEO
	x		B16	GEO
	x	A9_E	B18	GEO
	x	A1_B	B1A	LIT
	x	A1_C	B1B	LIT
	x	A1_D	B1C	LIT
	x	A9_I	B20	GEO
	x		B23a	GEO
	x		B23b	GEO
	x		B23c	GEO
	x		B24a	GEO
	x		B24b	GEO
	x		B26b	PROZ
	x		B26c	PROZ
	x		B26d	PROZ
	x	A6_C	B2C	LIT
	x		B30	GEO
	x		B31	GEO
	x		B33	GEO
	x		B34a	GEO
	x		B34b	GEO
	x		B35	PROZ
	x	A7_A	B3A	LIT
	x	A7_B	B3B	LIT
	x	A7_C	B3C	LIT
	x	A8_B	B4B	LIT
	x		B6A	GEO
	x		B6B	GEO
	x		B8a	LIT
	x		B8b	LIT
	x		B8c	LIT
	x		B9a	LIT
	x		B9b	LIT

Nummerierung Endform	Benennung Endform	Expra	Vorform	Skalen-zuordnung
	x	A1_A	x	LIT
	x	A6_D	x	LIT
	x	A7_D	x	LIT
	x	A9_J	x	GEO

Anmerkung. [1]nur in der Endform enthalten. [2] Item fehlerhaft, entfernt.

12.3 SPSS-Skript zum Vergleich abhängiger Korrelationen

```
* Das Programm erwartet die Korrelation von A mit C, von B mit C und von A mit B.
* Logik des Verfahrens: Steiger (1980) Tests for comparing Elements of a Correlation Matrix, psych bulletin, 2, 245-251.

**********************************************************************************************.
* Die Werte unter Begin Data und die Stichprobengröße müssen angepasst werden.
**********************************************************************************************.
DATA LIST free
/ Rab Rac Rbc .
BEGIN DATA.
0,41 0,52 0,48
END DATA.

* Stichprobengröße.
COMPUTE n=100.

* Fischers Z-Transformation der Korrelationen hier.
 compute Zab=0.5*ln((1+Rab)/(1-Rab)).
 compute Zac=0.5*ln((1+Rac)/(1-Rac)).


COMPUTE ra=(Rab+Rac)/2.

COMPUTE CV=((  1/  ((1-ra**2)**2)    )*(( Rbc*(1-2*ra**2) )   -
(ra**2*0.5*   (1-2*ra**2-Rbc**2)))).

COMPUTE Z=((sqrt(n-3))*(Zab-Zac))/(sqrt(2-2*CV)).
```

```
EXECUTE.

SUMMARIZE
  /TABLES=Z
  /FORMAT=LIST NOCASENUM TOTAL
  /TITLE='Z-Verteilte Statistik zum Vergleich der Korrelationen'
  /MISSING=VARIABLE
  /CELLS=COUNT .
```

12.4 Ladungen einer dreifaktoriellen MPLUS-ML Lösung der Endform

Tabelle 49 Dreifaktorielle Lösung korrelierter Faktoren N = 1554.

Parcel Nummer	Geometrie und grafische Fkt.	Mathematische Literalität	Prozedurales/ komplexes Rechnen
1	0,16	0,40	0,29
2	0,18	0,43	0,23
3	0,23	0,51	0,29
4	0,32	0,36	0,23
5	0,33	0,49	0,25
6	0,57	0,38	0,50
7	0,70	0,46	0,33
8	0,71		0,53
9	0,50		0,46
10	0,30		0,53
11			0,43
12			0,39
13			0,43
14			0,50
15			0,48
16			0,58
17			0,49
18			0,55
19			0,56
20			0,59

Anmerkung. Schätzmethode=ML. Die Zusammensatzung der Parcels entspricht jener der 4-Faktor Lösung. Die Parcels 15-20 des prozedurales/komplexes Rechnen-Faktors entsprechen den KOMPL1 bis KOMPL5 Parcels. Alle Ladungen sind hochsignifikant von Null unterschiedlich.

12.5 4-Faktorielle SL-Lösung der Endform mit WLSMV-Schätzung

Tabelle 50 4-Faktorielle Schmid-Leiman-Lösung der Endform mit WLSMV-Schätzung.

Parcel Nummer	G-Faktor	Mathematische Literalität	Prozedurales Rechnen	Komplexes Rechnen	Geometrie und Grafische Fkt.
1	0,71	0,53	-0,20	0,19	0,15
2	0,65	0,54	-0,07	0,28	0,17
3	0,60	0,68	-0,23	0,43	0,33
4	0,71	0,31	0,49	0,35	0,30
5	0,61	0,54	0,45	0,40	0,32
6	0,61	0,35	0,31		0,62
7	0,57	0,44	0,24		0,78
8	0,58		0,31		0,80
9	0,70		0,32		0,32
10	0,66		0,33		0,02
11	0,83		0,36		
12	0,66		0,05		
13	0,75		0,11		
14	0,41		0,42		
15	0,40		0,22		
16	0,77				
17	0,49				
18	0,76				
19	0,67				
20	0,71				
21	0,60				
22	0,55				
23	0,63				
24	0,66				
25	0,65				
26	0,53				
27	0,61				
28	0,63				
29	0,49				
30	0,56				
31	0,52				
32	0,60				
33	0,72				
34	0,71				
35	0,76				
36	0,74				
37	0,77				

Anmerkung. Zusammensetzung der Parcel siehe Tabelle 32. Schätzmethode: WLSMV. Varianz aller Faktoren=1. Alle Koeffizienten sind hochsignifikant ($p < 0,01$).

12.6 Kennwerte für die Skalen der Vorform A, vor jeglicher Itemselektion

12.6.1 Geometrie und grafische Funktionen

Tabelle 51 Klassische Kennwerte vor Itemselektion, Form A (N = 73).

Item	p	r_{it}	Cronbach's α nach Ausschluss
A21a	,60	,390	,613
A21b	,21	,510	,591
A21c	,33	,394	,612
A24	,47	,380	,615
A25	,12	,414	,617
A29	,53	,266	,644
A30	,27	,378	,617
A31	,62	,021	,698
A19	,70	,331	,627

Anmerkung. Cronbach's α aller Items: 0,655

12.6.2 Komplexes Rechnen

Tabelle 52 Klassische Kennwerte vor Itemselektion, Form A (N = 73).

Item	p	r_{it}	Cronbach's α nach Ausschluss
A9c	,90	,237	,784
A9e	,22	,402	,772
A9f	,32	,478	,764
A9g	,25	,585	,754
a11a	,70	,286	,783
a11b	,48	,576	,753
A18	,63	,453	,767
A17a	,52	,569	,754
A17b	,60	,439	,768
A17c	,12	,409	,772
A16c	,52	,535	,758
A16a	,86	,167	,789
A16b	,77	,157	,793

Anmerkung. Cronbach's α aller Items: 0,785

12.6.3 Mathematische Literalität

Tabelle 53 Klassische Kennwerte vor Itemselektion, Form A (N = 73).

Item	p	r_{it}	Cronbach's α nach Ausschluss
A27a	,29	,667	,592
A27b	,21	,611	,624
A26a	,32	,429	,697
a26b	,14	,362	,715
A28	,30	,367	,722

Anmerkung. Cronbach's α aller Items: 0,721.

12.6.4 Prozedurales Rechnen

Tabelle 54 Klassische Kennwerte vor Itemselektion, Form A (N = 73).

Item	p	r_{it}	Cronbach's α nach Ausschluss
a3a	,93	,387	,916
a3b	,68	,298	,917
a3e	,52	,414	,916
A2a	,95	,134	,917
A2d	,86	,328	,916
A2e	,85	,481	,915
A6a	,64	,584	,914
A6b	,70	,538	,915
A6c	,73	,504	,915
A6d	,53	,654	,913
a5b	,62	,520	,915
a5c	,30	,556	,915
a5d	,79	,421	,916
A20	,63	,429	,916
A22	,55	,381	,916
a10b	,59	,530	,915
A10c	,42	,514	,915
A8a	,95	,354	,917
A8c	,51	,647	,914
A4a	,95	,290	,917
A4f	,47	,613	,914
A4g	,45	,628	,914
A13b	,47	,440	,916
A13c	,19	,586	,915
A13d	,45	,428	,916
A12a	,40	,509	,915
A12b	,22	,676	,914
A12c	,33	,549	,915

Item	p	r_{it}	Cronbach's α nach Ausschluss
A12d	,12	,601	,915
A10a	,86	,208	,917
A13a	,60	,224	,918
A13e	,11	,485	,916
A14a	,78	,312	,917
A14b	,75	,097	,918
A14c	,88	,249	,917
A14d	,27	,389	,916
A15	,36	,474	,915
A1a	,79	,246	,917
A1b	,71	,400	,916
A1c	,63	,423	,916
A1d	,99	,135	,917
A2b	,93	,272	,917
A2c	,88	,317	,917
A2f	,44	,288	,917
A2g	,55	,265	,917
A2h	,19	,476	,915
A2i	,16	,467	,915
A2j	,22	,290	,917
a3c	,59	,064	,919
a3d	,60	,323	,917
A4b	,96	,126	,917
a4c	,93	,101	,918
A4d	,81	,107	,918
A4e	,85	,272	,917
a5a	,90	,164	,917
A7a	,93	,033	,918
A7b	,53	,242	,917
A7c	,93	,111	,918
A7d	,30	,267	,917
A8b	,45	,354	,916
A9a	,96	,146	,917
A9b	,93	-,085	,919
A9d	,44	,468	,915

Anmerkung. Cronbach's α aller Items: 0,917

12.7 Kennwerte der Skalen der Vorform B, vor jeglicher Itemselektion

12.7.1 Geometrie und Grafische Funktionen

Tabelle 55 Klassische Kennwerte vor Itemselektion, Form B (N = 76).

Item	P	r_{it}	Cronbach's α nach Ausschluss
B32	,51	,803	,967
B22	,49	,853	,967
B17	,50	,815	,967
B13	,53	,746	,968
B29a	,39	,783	,968
B29b	,38	,757	,968
B29c	,36	,750	,968
B15	,36	,704	,968
B21	,29	,698	,968
B19	,30	,695	,968
B28a	,26	,681	,968
B28b	,25	,664	,968
B28c	,25	,664	,968
B27a	,18	,602	,968
B27b	,24	,670	,968
B27c	,14	,554	,969
B27d	,12	,526	,969
B25a	,13	,520	,969
B25b	,12	,512	,969
B11a	,58	,795	,967
B11b	,57	,684	,968
B14	,38	,619	,968
B16	,20	,548	,969
B18	,30	,532	,969
B20	,24	,636	,968
B23a	,57	,780	,968
B23b	,38	,744	,968
B23c	,45	,818	,967
B24a	,43	,742	,968
B24b	,37	,751	,968
B30	,39	,676	,968
B31	,49	,737	,968
B33	,17	,556	,969
B34a	,29	,580	,969
B34b	,29	,546	,969
B6a	,08	,068	,970
B6b	,29	,527	,969

Anmerkung. Cronbach's α aller Items: 0,969

12.7.2 Mathematische Literalität

Tabelle 56 Klassische Kennwerte vor Itemselektion, Form B (N = 76).

Item	p	r_{it}	Cronbach's α nach Ausschluss
B5b	,54	,880	,974
B5c	,51	,844	,974
B5d	,46	,761	,975
B10a	,63	,925	,974
B10b	,62	,901	,974
B10c	,54	,898	,974
B2a	,57	,885	,974
B4a	,55	,900	,974
B4c	,53	,856	,974
B2b	,39	,698	,975
B4d	,43	,749	,975
B7b	,30	,608	,976
B12	,12	,373	,977
B10d	,22	,518	,976
B1a	,26	,544	,976
B1b	,11	,348	,977
B1c	,05	,249	,977
B2c	,54	,831	,974
B3a	,61	,886	,974
B3b	,49	,767	,975
B3c	,50	,806	,975
B4b	,49	,772	,975
B8a	,63	,925	,974
B8b	,63	,925	,974
B8c	,45	,769	,975
B9a	,47	,814	,975
B9b	,49	,827	,974

Anmerkung. Cronbach's α aller Items: 0,98

12.7.3 Prozedurales Rechnen

Tabelle 57 Klassische Kennwerte vor Itemselektion, Form B (N = 76).

Item	p	r_{it}	Cronbach's α nach Ausschluss
B26b	,32	,626	,778
B26c	,17	,781	,714
B26d	,16	,751	,730
B35	,46	,478	,860

Anmerkung. Cronbach's α aller Items: 0,817