Methoden der Psychologischen Diagnostik

Klaus D. Kubinger

Weil die meisten mathematisch-statistischen Methoden, die im Rahmen der Psychologischen Diagnostik von Bedeutung sind, in anderen Kapiteln dieses Handbuchs beschrieben werden, befaßt sich der vorliegende Beitrag nur mit zwei Themen: *„adaptives Testen"* und *„diagnostische Entscheidungstheorie"*. Zur Komplettierung sei allerdings auf die anderen Methoden und ihre Verwendungszwecke hingewiesen: Das sind zunächst alle experimentellen Methoden (vgl. Bredenkamp, in diesem Band) innerhalb der Grundlagenforschung der Psychologischen Diagnostik, ferner die Repräsentativerhebung (vgl. Schäffer, in diesem Band) zur *Normierung* eines Tests, Zusammenhangsmaße (vgl. Hager, in diesem Band) zur Bestimmung von *Reliabilität* und *Validität*, die Faktorenanalyse (vgl. Schönemann & Borg, in diesem Band) zur Konstruktvalidierung eines Tests, die *Latent-Trait-Modelle* der sogenannten „probabilistischen Testtheorie" (vgl. Roskam, in diesem Band, aber auch im folgenden) zur *Skalierung* von Tests sowie spezielle Methoden innerhalb der sogenannten „klassischen Testtheorie" (vgl. Stumpf, in diesem Band), insbesondere der *„Multitrait-Multimethod"*-Ansatz (MTMM). – Zur Definition der Gütekriterien *Normierung* und *Skalierung* siehe Kubinger (1995).

1 Adaptives Testen

Der Begriff „adaptives Testen" ist in der aktuellen Literatur unabdingbar an die *probabilistische Testtheorie* (*Latent-Trait-Theorie* oder *Item-Response-Theorie*, IRT) gebunden. Ausgangspunkt ist die Kritik am konventionellen Vorgehen psychologischer Testungen, nämlich jeder Testperson (Tp) dieselben Items in ein und derselben Reihenfolge vorzugeben. Dabei fallen leistungsfähigen Personen einige Items regelmäßig zu leicht, zumindest sehr leicht, leistungsschwachen dagegen einige andere zu schwer, zumindest sehr schwer. Solche Items sind für die jeweils betroffene Personengruppe insofern nicht *informativ*, als der Ausgang ihrer Bearbeitung von vornherein bekannt, jedenfalls höchstwahrscheinlich ist. Ob sie tatsächlich vorgegeben werden oder nicht, ist weitgehend unerheblich. Zur Erhöhung der „Testökonomie" schiene es daher von Vorteil, nur „informative" Items vorzugeben.

Einerseits die Bestimmung der so verstandenen „Information" eines Items, andererseits der faire Vergleich von Testleistungen zwischen zwei oder mehreren Personen, die verschiedene Items bearbeiteten, macht die Anwendung der *probabilistischen Testtheorie* notwendig. Zwar gibt es auch für manche Tests, die nach der *klassischen Testtheorie* konzipiert sind, personenspezifische, leistungsabhängige Itemauswahlstrategien; der angesprochene faire Leistungsvergleich ist dabei allerdings nicht möglich.

https://doi.org/10.25521/HQM41

Die „Theorie der *Maximum-Likelihood*-Schätzung" nach R. A. Fisher (vgl. z.B. Lord & Novick, 1968) ermöglicht die problemspezifische Berechnung dieser „Information" (*information in the sample*) nach folgender Formel:

$$I(i,v) = \frac{[p'(+|i,v)]^2}{p(+|i,v) \cdot p(-|i,v)} \quad . \tag{1}$$

Die Information bestimmt sich also als Funktion der Wahrscheinlichkeit (p), daß Tp v Item i löst (+) bzw. nicht löst (−). Der Term $p'(+|i,v)$ im Zähler ist hierbei die erste Ableitung der „*Item-Response*-Funktion" und beschreibt den Anstieg der Lösungswahrscheinlichkeit von Item i in Abhängigkeit von der Fähigkeit einer Tp (*item characteristic curve*, s.u.), und zwar an der durch die Fähigkeit von Tp v definierten Stelle. Verallgemeinerungen, die über die Kategorisierung in „richtig" und „falsch" hinausgehen, werden erst später angedeutet.

An Modellen, die Aussagen über die fraglichen Wahrscheinlichkeiten machen, kommen vor allem diejenigen von *Rasch* und *Birnbaum* in Frage. Das *dichotome logistische Testmodell von Rasch* setzt die Wahrscheinlichkeit für die Lösung eines Items – unter der Voraussetzung, daß die Fähigkeit der Tp v durch den eindimensionalen Personenparameter ξ_v, die Schwierigkeit des Items i durch den eindimensionalen Itemparameter σ_i beschreibbar ist – wie folgt an:

$$p(+|\xi_v, \sigma_i) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}}. \tag{2}$$

Die Angabe der Wahrscheinlichkeit für eine Nicht-Lösung als Gegenwahrscheinlichkeit kann hier wie im folgenden entfallen. Das *zwei-parametrige logistische Modell von Birnbaum* (2-PL-Modell) sieht zusätzlich zum Schwierigkeitsparameter σ_i einen zweiten Itemparameter vor, den Diskriminationsparameter α_i. Dieser trägt dem möglichen Umstand Rechnung, daß nicht alle Items zwischen Personen mit verschiedenen Fähigkeiten gleich gut diskriminieren, so daß die Unterschiede zwischen den Lösungswahrscheinlichkeiten zweier bestimmter Personen für jeweils zwei Items mit demselben Schwierigkeitsparameter nicht notwendigerweise gleich sein müssen (vgl. dazu die beiden Items 1 und 3 für die beiden Personen v und w in Abbildung 1). Die fragliche Wahrscheinlichkeit im 2-PL-Modell[1] beträgt nun:

$$p(+|\xi_v; \sigma_i, \alpha_i) = \frac{e^{\alpha_i(\xi_v - \sigma_i)}}{1 + e^{\alpha_i(\xi_v - \sigma_i)}}. \tag{3}$$

Offensichtlich stimmt im Fall $\alpha_i = 1$ (für alle i) das 2-PL-Modell mit dem *dichotomen logistischen Modell von Rasch* (1-PL-Modell, weil „ein-itemparametrig") überein. Im sogenannten *drei-parametrigen logistischen Modell von Birnbaum* (3-PL-Modell) wird noch ein dritter Itemparameter, der Rateparameter β_i, berücksichtigt; er soll im Fall von *Multiple-Choice*-Antwortformaten dem Umstand Rechnung tragen, daß Lösungen auch durch Raten zustandekommen können, infolgedessen sich die Lösungswahrscheinlichkeit gegenüber dem 2-PL-Modell erhöht:

$$p(+|\xi_v; \sigma_i, \alpha_i, \beta_i) = \frac{\beta_i + e^{\alpha_i(\xi_v - \sigma_i)}}{1 + e^{\alpha_i(\xi_v - \sigma_i)}}. \tag{4}$$

[1]Um die logistische Funktion an eine Ogive und damit das Modell an das sogenannte „*normal ogive model*" anzugleichen, welches auf eine Arbeit von Lawley aus dem Jahre 1943 zurückgeht, findet sich in der Literatur oft α_i durch $1.7\,\alpha_i$ ersetzt (vgl. z.B. Hulin, Drasgow & Parsons, 1983).

Sind alle $\beta_i = 0$, vereinfacht sich das Modell zum 2-PL-Modell, sind andererseits alle $\alpha_i = 1$, ergibt sich ein *Rasch-Modell mit Rateparameter.*

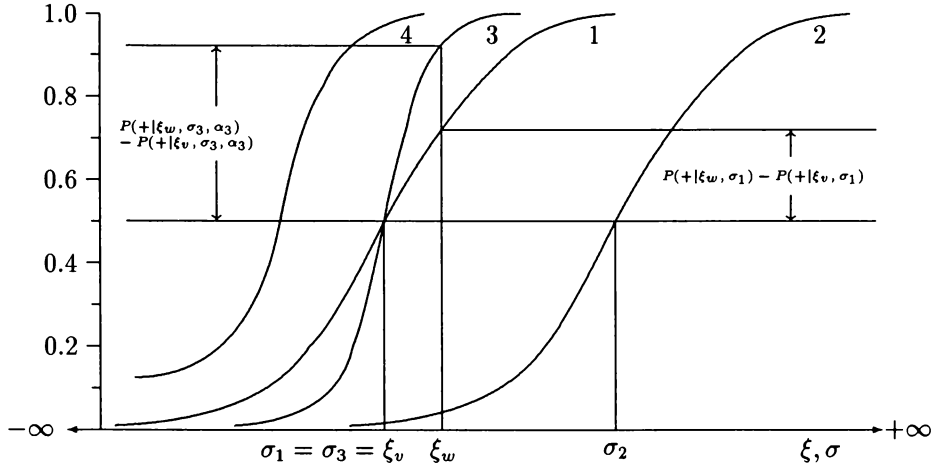

ABBILDUNG 1. **Gegenüberstellung der Modelle von *Rasch* und *Birnbaum.***

Abbildung 1 veranschaulicht die zu den genannten Modellen gehörenden *item characteristic curves* (ICCs). Diese Kurven geben für ein Item mit einem bestimmten Schwierigkeitsparameter σ_i jeweils die Lösungswahrscheinlichkeit in Abhängigkeit vom Personenparameter ξ_v an. Mit dem 1-PL-Modell sind nur die Items (Kurven) 1 und 2 konform; sie unterscheiden sich lediglich hinsichtlich der Schwierigkeitsparameter σ_1 und σ_2. Mit dem 2-PL-Modell ist zusätzlich das Item 3 verträglich, welches wegen $\alpha_3 > \alpha_1 = \alpha_2$ trotz $\sigma_1 = \sigma_3$ besser zwischen ξ_v und ξ_w diskriminiert als Item 1. Schließlich ist mit dem 3-PL-Modell sogar noch das Item 4 verträglich, das im Gegensatz zu allen übrigen Items einen Rateparameter $\beta_4 > 0$ aufweist.

Je nachdem, welches dieser Modelle zur Anwendung[2] gelangt, kann das durch Gleichung (1) definierte Informationsmaß aus empirischen Daten geschätzt und das jeweils informativste Item ausgewählt werden. Im einfachsten Fall, dem 1-PL-Modell, beläuft sich diese „Information" anschaulich und plausibel auf $I(i,v) = p(+|i,v) \cdot p(-|i,v)$, mit dem Maximum bei der Lösungswahrscheinlichkeit von .50: Dasjenige Item ist am informativsten, für welches die Chancen der Lösung durch Tp v 50:50 stehen; bei Lösungswahrscheinlichkeiten nahe eins bzw. null ist auch die „Information" des Items nahezu null.

Das informativste Item zu finden, setzt in jedem Fall voraus, daß sowohl die Itemparameter aus entsprechend großen Voruntersuchungen (sog. Kalibrierungs-Stichproben) bekannt sind als auch eine vorläufige Schätzung für den unbekannten Personen-

[2]Weil es sich bei Modellen stets um Annahmen handelt, sollte immer dasjenige Modell zur Anwendung gelangen, welches sich anhand empirischer Daten auch als gültig herausgestellt hat. Dabei ist aber nur die Modellgültigkeit des 1-PL-Modells eigentlich prüfbar: Zwar sind für das 2- und das 3-PL-Modell Modellanpassungstests (*goodness-of-fit tests*) möglich, doch lassen sich für sie keinerlei Modellimplikationen prüfen (genaueres z.B. bei Kubinger, 1989). Aus diesem Grund, verbunden mit der Tatsache, daß die Parameterschätzung im 2- und 3-PL-Modell wesentlich größere Probleme mit sich bringt, ist adaptives Testen unter Verwendung des 1-PL-Modells vorzuziehen.

parameter ξ_v vorliegt. Eine erste solche Schätzung ergibt sich entweder, indem von einer durchschnittlichen Fähigkeit der Tp v ausgegangen wird, oder aufgrund von Vorinformationen über einen ungefähr zu erwartenden Parameter. Um beim Fehlen jeder Vorinformation möglichst bald zu einer empirischen Schätzung zu gelangen, empfiehlt es sich, als erstes Item ein mittelschwieriges vorzugeben und danach – je nachdem ob dieses gelöst oder nicht gelöst wurde – das schwierigste oder das leichteste. Nur ausnahmsweise wird es nach dieser Strategie mehr als zweier Items für eine erste Schätzung bedürfen. Damit ist es aber bereits möglich, die Testleistungen von Personen, denen verschiedene Items vorgegeben wurden, zu vergleichen. Durch die Vorgabe weiterer Items verbessert sich jedesmal die Schätzung.

Auf Grund des modellierten Zusammenhangs von Personenparameter und Itemparametern kann zum Beispiel für den allgemeinsten Fall, das 3-PL-Modell, ξ_v als *Maximum-Likelihood*-Schätzung $\hat{\xi}_v$ aus folgender (*Likelihood*-) Funktion bestimmt werden:

$$L_v = \prod_{j=f_1(v)}^{f_{k_v}(v)} \left(\frac{\beta_j + e^{\alpha_j(\hat{\xi}_v - \sigma_j)}}{1 + e^{\alpha_j(\hat{\xi}_v - \sigma_j)}} \right)^{x_{vj}} \cdot \left(\frac{1 - \beta_j}{1 + e^{\alpha_j(\hat{\xi}_v - \sigma_j)}} \right)^{1 - x_{vj}} , \tag{5}$$

wobei $x_{vj} = 1$ im Fall der Lösung (d.h. +) und $x_{vj} = 0$ im Fall der Nicht-Lösung (d.h. −) gilt; $f_1(v), f_2(v), \ldots, f_{k_v}(v)$ geben die Nummern derjenigen k_v Items an, welche Tp v infolge der adaptiven Auswahl vorgegeben wurden. Bekanntlich sind *Maximum-Likelihood*-Schätzungen asymptotisch normalverteilt, mit der (Fehler-) Varianz als dem Kehrwert der sogenannten „Informationsfunktion" (vgl. z.B. Kendall & Stuart, 1979). Letztere ist im gegebenen Fall identisch mit der über die vorgegebenen Items summierten „Information", d.h. $I(v) = \sum_i I(i, v)$. Für das 1-PL-Modell resultiert beispielsweise der Standardschätzfehler (*standard error of estimation*):

$$S(\xi_v) = \sqrt{\frac{1}{I(v)}} = \left[\sum_{j=f_1(v)}^{f_{k_v}(v)} \frac{e^{\xi_v - \sigma_j}}{1 + e^{\xi_v - \sigma_j}} \cdot \frac{1}{1 + e^{\xi_v - \sigma_j}} \right]^{-\frac{1}{2}} \tag{6}$$

Er entspricht dem Standard*meß*fehler der *klassischen Testtheorie*. Daraus ist unmittelbar ersichtlich, daß bei ideal informativen Items der Schätzfehler in bezug auf ξ_v minimal wird.

Gegenüber der optimalen Strategie, nämlich für jede Tp zu jedem Zeitpunkt der Testvorgabe das unter den vorhandenen Items maximal informative zu wählen (sog. „*tailored testing*"), kann jede davon abweichende Variante adaptiven Testens nur suboptimal sein; trotzdem hat das sogenannte „*branched testing*", also eine von vornherein festgelegte, festverzweigte Itemauswahlstrategie, Attraktivität: Sie rührt aus dem Bestreben, der beim *tailored testing* obligaten computerunterstützten Testvorgabe eine Alternative gegenüberzustellen. Während eine (verbesserte) Schätzung des gesuchten Personenparameters nach jedem einzelnen Item und vor allem die Auswahl des jeweils informativsten Items ohne die Verwendung von Computern unrealistisch ist, kann das *branched testing* auch bei Papier-Bleistift-Vorgabe angewendet werden. Dabei sind einzelne Items zu mehreren Itemgruppen zusammengefaßt, so daß sich die adaptive Testvorgabe darauf beschränkt, nach jeder Item*gruppe* leistungsabhängig zu verzweigen. Da hierbei wesentlich weniger Entscheidungsschritte notwendig sind als beim *tailored testing*, ist eine solche Vorgabe im Einzelversuch

vom Testleiter durchaus bewältigbar. Abbildung 2 veranschaulicht ein beispielhaftes Verzweigungsschema beim *branched testing*: Je nach Testleistung pro Itemgruppe, und zwar dreifach abgestuft, kann zur nächsten Itemgruppe weitergegangen werden, *ohne* zuvor den gesuchten Personenparameter zu schätzen; auch am Ende der Testung muß dieser nicht extra geschätzt werden, weil die Anzahl aller realisierbaren Itemkombinationen und Testleistungen gegenüber dem *tailored testing* extrem reduziert ist und daher sämtliche Parameterschätzungen im vorhinein bestimmt und tabellarisch aufbereitet werden können.

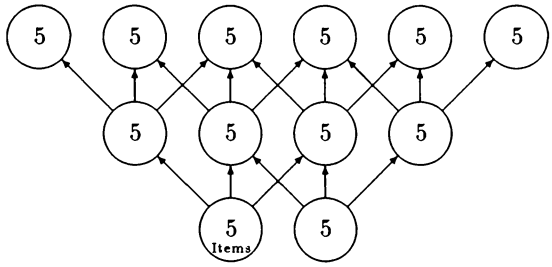

ABBILDUNG 2. **Beispiel eines Verzweigungsschemas beim *branched testing* (2 Startgruppen, 3 Schritte, dreigeteilte Verzweigung, 5 Items pro Itemgruppe).**

In Abbildung 2 sind von links nach rechts insgesamt sechs verschiedene Niveaustufen der Items abgetragen. Es ist vorgesehen, daß die Tp auf Grund von Vorinformationen bei der leichteren oder der schwereren Startgruppe beginnt (Schwierigkeitsniveau 3 und 4); löst sie höchstens ein Item, wird sie in der Folge zu einer leichteren Itemgruppe verwiesen, löst sie mindestens vier, zu einer schwierigeren, und löst sie zwei bis drei (das sind ungefähr 50%), bleibt sie beim gegebenen Schwierigkeitsniveau der Items, usw.

Theoretische Arbeiten inklusive Simulationsstudien zur Praktikabilität und Effizienz des adaptiven Testens lassen folgendes erkennen:

- Die Vorgabe beim *tailored testing* kann dann beendet werden, wenn die Schätzungen ein und desselben Personenparameters nach zwei aufeinanderfolgenden Items nicht um mehr als um einen bestimmten, vorher festgelegten und diagnostisch vertretbaren Betrag differieren.

- Dieses Kriterium wird bei einem zu kleinen Itempool kaum erfüllbar sein, weil die dafür nötigen Items nicht (mehr) existieren. Nach bisherigen Erfahrungen reichen jedoch 60 bis 70 Items aus, wobei nach ca. 15 Items hinreichende Genauigkeit der Parameterschätzungen erreicht wird (vgl. am besten Wild, 1986).

- Praktisch keinen Einfluß auf die endgültige Schätzung des gesuchten Personenparameters hat die Wahl des bzw. der Startitems.

- Dementsprechend besteht beim *branched testing* für den Zuwachs an Schätzgenauigkeit als Folge einer größeren Anzahl verwendeter Startgruppen bald eine (natürliche) Grenze.

- Im Vergleich zu einer größeren Anzahl von Verzweigungsschritten ist laut Kubinger und Wild (1989) in bezug auf den Schätzfehler die größere Anzahl von Verzweigungsmöglichkeiten wichtiger. Ebenso ist im Vergleich zu mehr Items pro Schritt eine größere Anzahl von Verzweigungsschritten wichtiger.

Zur Illustration diene ein willkürlich gewählter Itempool bestehend aus 53 Items, die sich empirisch als dem 1-PL-Modell konform herausgestellt haben; die auf Summe null normierten Schwierigkeitsparameter weisen eine Standardabweichung von 3.5 auf und liegen, annähernd eingipfelig und symmetrisch („*peaked-*") verteilt, im Intervall $-7.0 \leq \sigma_i \leq 6.0$. Für diesen Itempool wird der resultierende Standardschätzfehler bei verschiedenen Testvorgaben in Abbildung 3 gegenübergestellt.

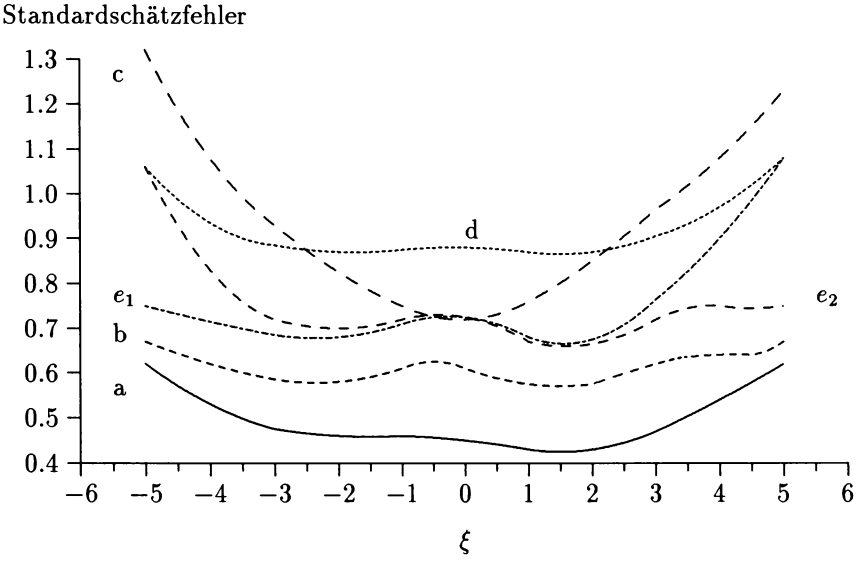

ABBILDUNG 3. Der Standardschätzfehler bei verschiedenen Testvorgaben (a: konventionelle Vorgabe aller 53 Items; b: *tailored testing* mit jeweils 15 Items; c: konventionelle Vorgabe von ausgewählten 15 Items, deren Schwierigkeitsparameter sich über das Intervall *peaked* verteilen; d: konventionelle Vorgabe von ausgewählten 15 Items, deren Schwierigkeitsparameter sich über das Intervall gleich verteilen; e: *branched testing* mit jeweils 15 Items gemäß Verzweigungsschema aus Abbildung 2; die durchschnittliche Streuung der Schwierigkeitsparameter pro Itemgruppe ist 1.6 – wegen der zwei Startgruppen ergeben sich hier zwei Kurven, e_1 und e_2.)

Für die drei konventionellen Vorgaben bestimmt sich der Standardschätzfehler nach der oben gegebenen Formel (6), wobei sich diese noch vereinfacht, weil jeweils dieselben Items in Betracht kommen. Für das *branched testing* muß dagegen der Standardschätzfehler nach dieser Formel zunächst pro Personenparameterwert ξ (und pro Startgruppe) für jede der nach dem Verzweigungsschema möglichen Itemkombinationen berechnet werden. Im zweiten Schritt werden die derart erhaltenen *möglichen* Standardschätzfehler mit der jeweiligen Wahrscheinlichkeit gewichtet, daß es beim gegebenen ξ gerade zu der betreffenden und nicht zu irgend einer anderen Itemkombination kommt (vgl. die genaue Formel bei Kubinger & Wild, 1989). Schließlich

werden die mit ihren Wahrscheinlichkeiten gewichteten möglichen Standardschätzfehler aufsummiert, so daß man als Ergebnis (eigentlich) den Erwartungswert des Schätzfehlers erhält. Für das *tailored testing* ist die Berücksichtigung aller möglichen Itemkombinationen zu aufwendig, so daß hier pro Personenparameterwert ξ 300 sogenannte „*simulees*" (fiktive Testpersonen innerhalb eines Simulationsverfahrens) herangezogen wurden, aus deren Testergebnissen über obige Formel der (*mittlere*) Standardschätzfehler bestimmt wurde.

An publizierten adaptiven Tests gibt es bis jetzt: Die Intelligenz-Testbatterie AID (*Adaptives Intelligenz Diagnostikum*; Kubinger & Wurst, 1985) mit neun *branched tests* (für Papier-Bleistift-Vorgabe); acht davon liegt das 1-PL-Modell zugrunde, einem eine Verallgemeinerung des 1-PL-Modells für dreikategorielle Antwortbewertungen („richtig und schnell", „richtig, aber langsam" und „falsch"); den BBT (*Begriffs-Bildungs-Test*; Kubinger, Fischer & Schuhfried, 1993), ein computergestützter *branched test*, und schließlich die *Syllogismen* (Srp, 1994), ein computergestützter *tailored test*.

2 Diagnostische Entscheidungstheorie

Daß sich psychologisches Diagnostizieren nicht ausschließlich an der herkömmlich bestimmten Validität eines Tests orientieren kann, wird allein aus den berühmten *Taylor-Russel-Tafeln* offensichtlich: Auch Entscheidungen, die auf wenig validen Tests beruhen, sind zufälligen Entscheidungen dann eindeutig überlegen, wenn die Selektionsquote niedrig ist, vor allem bei mittlerer Grundrate geeigneter Kandidaten (vgl. z.B. Kubinger, 1995). Es geht also auch um die Konsequenzen einer Diagnose.

Ausgangspunkt jeder diagnostischen Entscheidungstheorie ist dabei eine oft zitierte Arbeit von Cronbach und Gleser (1965); daraus wird klar, daß es Unterschiede macht, ob die Diagnose im Interesse der Testperson oder einer Institution liegt, ob und wieviele Interventionsalternativen es gibt, ob ein oder mehrere Testergebnisse zur Diskussion stehen u.ä.

Im einfachsten Fall handelt es sich um Alternativentscheidungen, die sachlich richtig oder falsch sein können. Fehler sind dabei, genauso wie beim Hypothesenprüfen innerhalb der *Neyman-Pearson*-Statistik (vgl. Willmes, in diesem Band), auf zweierlei Art möglich: Entweder ein Faktum nicht (positiv) zu befunden oder trotz Fehlens des Faktums es doch (positiv) zu befunden. Läge in diesem Fall die Diagnose allein im Interesse der Testperson und bestünde nur die Wahl zwischen Intervention (ohne „Nebenwirkungen") und keiner Intervention, bräuchte man für die Entscheidung eigentlich gar kein Testergebnis; gibt es jedoch (auch) institutionelle Interessen, so müssen Effizienzbetrachtungen miteinbezogen werden.

Das auf der folgenden Seite in Tabelle 1 dargestellte inhaltliche Beispiel soll dies illustrieren. Vordergründig stellt sich in diesem Beispiel die Frage, ob eine globale Trefferrate von $.13 + .51 = .64$ lohnt.

Das Beispiel zeigt folgendes:

- Im Interesse einer Tp würde es (u.U.) liegen, entweder weitere Untersuchungen zur besseren Absicherung der Entscheidung anzustellen oder die möglichen Therapieprogramme in jedem Fall einzusetzen.

TABELLE 1. Ein inhaltliches Beispiel mit den (geschätzten) Wahrscheinlichkeiten richtiger und falscher positiver sowie richtiger und falscher negativer Diagnosen (empirische Daten nach Kubinger, 1984).

		Testdiagnose	
		positiv	*negativ*
Tatsächlicher Zustand	cerebral geschädigt	.13	.07
	nicht cerebral geschädigt	.29	.51

- Geht es dem institutionellen Interesse dagegen tatsächlich (nur) um die globale Trefferrate, so wäre diese bei gegebener Grundrate cerebral geschädigter Kinder von $.13 + .07 = .20$ gegenüber der testbedingten mit .64 allerdings leicht zu erhöhen. Dies gelingt im Beispiel entweder durch eine zufällige Entscheidung oder durch eine spieltheoretisch optimierte Entscheidung. Im ersten Fall würde die Wahrscheinlichkeit dafür, daß das jeweils in Betracht gezogene Kind sowohl tatsächlich cerebral geschädigt ist als auch rein zufällig als solches bezeichnet wird, $.20 \cdot .20 = .04$ betragen, die Wahrscheinlichkeit, daß das Kind sowohl tatsächlich nicht cerebral geschädigt ist als auch dem Zufall nach als nicht cerebral geschädigt bezeichnet wird, $.80 \cdot .80 = .64$, was in der Summe $.68 > .64$ ergibt. Im zweiten Fall bräuchte man bloß jedesmal „nicht cerebral geschädigt" entscheiden, um auf eine Trefferrate von $.80 > .64$ zu kommen. In einem anderen Beispiel mit geringerer Grundrate würde, so gesehen, ein Test fast keine Chance haben, sich zu bewähren.

- Offensichtlich haben die beiden Treffermöglichkeiten (institutionell) differentielle Bedeutung bzw. liefern differentiellen *Nutzen*. So mag bei bestimmten Rahmenbedingungen eine hohe „Spezifität", d.h. eine hohe Wahrscheinlichkeit negativer Diagnosen bei tatsächlich negativem Zustand (hier: $.51/(.29+.51) = .64$), relevant sein, eine hohe „Sensitivität", d.h. eine hohe Wahrscheinlichkeit positiver Diagnosen bei tatsächlich positivem Zustand (hier: $.13/(.13 + .07) = .65$), jedoch weitgehend irrelevant.

- In der Regel haben auch die beiden Fehlermöglichkeiten differentielle Bedeutung bzw. wirken dem angeführten Nutzen quasi als „Schaden" unterschiedlich entgegen. Spätestens damit wird der Entscheidung eine Nutzenfunktion zugrunde gelegt, die jedem der vier nach Tabelle 1 möglichen Ergebnisse einen Nutzenwert (bzw. Schadenwert) zuordnet. Dabei braucht „Nutzen" (bzw. dessen Gegenteil) nicht auf monetäre Konsequenzen beschränkt zu sein.

Kubinger (1984) hat versucht, an Hand einer größeren Stichprobe von Psychologen solche Nutzenfunktionen empirisch zu bestimmen, und zwar mittels der Paar-

vergleichs-Skalierungsmethode unter Anwendung einer Verallgemeinerung des BTL-(*Bradley-Terry-Luce-*) Modells. Damit waren eindimensionale Nutzenfunktionswerte (u_1 bis u_4) gewährleistet. Es stellte sich erstens heraus, daß es zumindest zwei ziemlich gegensätzliche, typische Nutzenfunktionen gibt, wobei der dementsprechende Gesamtnutzen

$$U = u_1(.13) - u_2(.07) - u_3(.29) + u_4(.51) \tag{7}$$

einmal positiv, das andere Mal negativ resultierte.

Cronbach und Gleser (1965) definieren den Gesamtnutzen einer Entscheidung allgemeiner, indem sie mehrkategorielle Diagnosen $j = 1, 2, \ldots$ sowie mehr als einen Test $h = 1, 2, \ldots$ zulassen, ferner Verteilungsannahmen zur Testwertvariablen X_h treffen und sowohl die Nutzenfunktion als auch die Kosten der Testung pro Testwert K_x als bekannt annehmen. Für eine Testwertvariable mit Realisationen x (bzw. allgemein: $\mathbf{x}$ als Vektor der Realisationen aller Testwertvariablen) ergibt sich:

$$U = \sum_x \phi_x \sum_j f_{jx} \sum_v u_v p_{vjx} - \sum_x \phi_x K_x. \tag{8}$$

Hier bezeichnet ϕ_x die Wahrscheinlichkeit des Testwerts x, f_{jx} die Wahrscheinlichkeit der Diagnose j im Fall des Testwerts x, v eine Ausprägung der „Validitätsvariablen", in bezug auf welche diagnostiziert wird, u_v den Nutzen von v und p_{vjx} die Wahrscheinlichkeit von v im Fall der Diagnose j beim Testwert x.

Neue Ansätze innerhalb der *klassischen* Testtheorie beziehen unmittelbar nutzentheoretische Strategien für differentialdiagnostische Fragestellungen ein, indem sie die fraglichen Trennwerte für die verschiedenen Diagnosen so bestimmen, daß der insgesamt zu erwartende Nutzen optimiert wird (vgl. z.B. Mellenbergh & van der Linden, 1981). Immer wird dabei aber die Nutzenfunktion als bekannt vorausgesetzt.

3 Weiterführende Literatur

Zusätzlich zu den bereits zitierten Arbeiten kann das Buch von Wainer (1990) und der Artikel von Hornke (1993) zum adaptiven Testen empfohlen werden. Einen Einstieg in die diagnostische Entscheidungstheorie bietet das Buch von Wottawa und Hossiep (1987).

Literaturverzeichnis

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Hornke, L. F. (1993). Mögliche Einspareffekte beim computergestützten Testen. *Diagnostica, 39*, 109–119.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item response theory.* Homewood: Dow Jones-Irwin.

Kendall, M. G. & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2). London: Griffin.

Kubinger, K. D. (1984). Nutzentheoretische Beurteilung differential-diagnostischer Entscheidungen. *Diagnostica, 30*, 249–266.

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen* (S. 19–83). Weinheim: Beltz.

Kubinger, K. D. (1995). *Einführung in die Psychologische Diagnostik.* Weinheim: Psychologie Verlags Union.

Kubinger, K. D., Fischer, D. & Schuhfried, G. (1993). *Begriffs-Bildungs-Test (BBT). Software und Manual.* Mödling: Schuhfried.

Kubinger, K. D. & Wild, B. (1989). Die Optimierung der Meßgenauigkeit beim „branched"-adaptiven Testen. In K. D. Kubinger (Hrsg.), *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen* (S. 187–218). Weinheim: Beltz.

Kubinger, K. D. & Wurst, E. (1985). *Adaptives Intelligenz Diagnostikum (AID).* Weinheim: Beltz.

Lord, F. M. & Novick, M. R. (Eds.) (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Mellenbergh, G. J. & van der Linden, W. J. (1981). The linear utility model for optimal selection. *Psychometrika, 46,* 283–293.

Srp, G. (1994). *Syllogismen. Test: Software und Manual.* Frankfurt: Swets.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale: Erlbaum.

Wild, B. (1986). *Der Einsatz adaptiver Teststrategien in der Fähigkeitsmessung* (unveröffentlichte Dissertationsschrift). Wien: Institut für Psychologie der Universität Wien.

Wottawa, H. & Hossiep, R. (1987). *Grundlagen der psychologischen Diagnostik.* Göttingen: Hogrefe.