Arch. Psychol. 134, 197–218 (1981/82)

Psychodiagnostik auf Ordinalskalenniveau: Meßtheoretische Grundlagen, Modelltest und Parameterschätzung

Hans Irtel* und Franz Schmalhofer**

* Institut für Psychologie, Universität Regensburg, 8400 Regensburg.
** Department of Quantitative Psychology, University of Colorado, Boulder, Colo 80309, USA.
Eingegangen am 15. Oktober 1981

Zusammenfassung. Zur Konstruktion psychologischer Tests wird ein meßtheoretisch begründetes Modell entwickelt, das die Messung von Personenfähigkeit u und Aufgabenleichtigkeit v auf Ordinalskalen gestattet, so daß für die Antwortwahrscheinlichkeiten der Personen g und h bei den Aufgaben i und j gilt

$u(g) < u(h)$ g.d.w. für eine Aufgabe i $P(g,i) < P(h,i)$,
$v(i) < v(j)$ g.d.w. für eine Person g $P(g,i) < P(g,j)$.

Zur Prüfung, ob empirisch erhobene relative Häufigkeiten diese Restriktionen erfüllen, wird ein statistischer Test vorgeschlagen und an einem Beispiel durchgeführt. Außerdem wird ein Schätzverfahren dargestellt, das es erlaubt, aus relativen Häufigkeiten Antwortwahrscheinlichkeiten so zu schätzen, daß die empirischen Restriktionen des Modells erfüllt sind. Aufbauend auf diesem Schätzverfahren kann das vorgestellte Modell zur Konstruktion individualisierter Testverfahren verwendet werden.

Stichwörter: Psychodiagnostik, Ordinalskala, Ordnungs-Unabhängigkeit, Schätzung ordinaler Parameter, individualisiertes Testen.

Ordinal Psychological Testing: Measurement Theory, Statistical Test, and Parameter Estimation

Summary. A probabilistic conjoint measurement model for application in psychological testing is presented. This model yields ordinal scales u and v for subjects and items respectively such that the probabilities of a correct answer of subjects g and h to items i and j satisfy

$u(g) < u(h)$ iff for some item i $P(g,i) < P(h,i)$,
$v(i) < v(j)$ iff for some subject g $P(g,i) < P(g,j)$.

A statistical test for this model is proposed and applied to a set of empirical data. In addition, an estimation procedure is developed, which calculates the probabilities of correct responses such that the empirical restrictions of the model are satisfied. This algorithm proves useful for a tailored testing administration of the proposed model.

Key words: psychological testing, ordinal scaling, ordinal independence, estimation of ordinal parameters, tailored testing.

1. Einleitung

In der psychologischen Diagnostik haben seit den Arbeiten von Rasch (1960) und Birnbaum (1968) bedeutsame Veränderungen der theoretischen Grundlagen stattgefunden. Diese Veränderungen betreffen nicht nur die Inhalte der zugrundeliegenden Theorien, sondern davon sind auch die Kriterien betroffen, mit denen diese beurteilt werden (Rasch, 1961; Fischer, 1974). So betrachtet man heute als größten Mangel der klassischen Testtheorie, daß viele ihrer Annahmen nicht empirisch prüfbar und damit auch nicht zu widerlegen sind (Fischer, 1974). Das von Rasch (1960) entwickelte logistische Testmodell ist dagegen nicht nur in sich stringent formuliert, sondern es kann auch statistisch geprüft werden (Andersen, 1973; Hamerle & Tutz, 1980). Trotz dieser Vorzüge hat es in der praktischen Testkonstruktion bisher nur geringe Bedeutung erlangt. Dies liegt zum größten Teil an den starken Anforderungen des Rasch-Modells an die Daten, die es sehr schwierig machen, eine modellkonforme Aufgabensammlung zu finden. So ergaben die statistischen Modellkontrollen vieler Anwendungsversuche signifikante Modellabweichungen (Fricke, 1972; Spada, Fischer & Heyner, 1973; Hehl & Hehl, 1975; Herbst, 1978). In der vorliegenden Arbeit soll deshalb ein Modell vorgestellt werden, das meßtheoretisch fundiert und statistisch testbar ist, jedoch erheblich schwächere Restriktionen an die Daten stellt. Die schwächeren Annahmen bedingen ein niedriges Skalenniveau, das aber für viele Anwendungen psychologischer Tests ausreichend ist.

2. Ein ordnungs-unabhängiges psychologisches Testmodell

Die Konstruktion eines psychologischen Tests besteht in der Auswahl einer Menge $A = \{a, b, \ldots\}$ von Personen, einer Menge $B = \{i, j, \ldots\}$ von Testaufgaben und einem Zufallsexperiment, nämlich der Lösung der Aufgaben durch die Personen. Dies kann durch einen Zufallsvektor $X = [X(1,1), \ldots, X(|A|, |B|)]$ beschrieben werden, wobei A und B endlich sein sollen und mit g aus A und i aus B definiert wird

$$X(g,i) = \begin{cases} 1 & \text{falls Person g Aufgabe i richtig löst,} \\ 0 & \text{sonst.} \end{cases}$$

Die Zufallsvariablen $X(g,i)$ sollen unabhängig verteilt sein, so daß die Wahrscheinlichkeit $\mathbb{P}\{X(g,i) = 1\}$ eine Abbildung P von $A \times B$ in das reelle Intervall $[0,1]$ definiert, die jedem Paar (g,i) aus $A \times B$ genau ein $P(g,i)$ zuordnet, wobei $P(g,i) = \mathbb{P}$

$\{X(g,i)=1\}$. Wir werden die Abbildung P deshalb auch als „Lösungswahrscheinlichkeit" bezeichnen.

Definition 1. Ein Tripel $< A,B,P >$ in dem A und B endliche Mengen sind und P eine Abbildung von A x B in das geschlossene, reelle Intervall [0,1] ist, nennen wir (endliches) *probabilistisches, psychologisches Testsystem* oder kürzer *System.*

Mit einem System $< A,B,P >$ ist die Grundlage zur Entwicklung meßtheoretisch begründeter Modelle gelegt und wir können damit einige kaum verzichtbare Voraussetzungen für die praktische Anwendung formulieren. Eine dieser Voraussetzungen ist die Existenz der Abbildung P, die prinzipiell nicht überprüft werden kann, da wiederholte Beobachtungen von X in der Regel nicht möglich sind. Für die Testkonstruktion ist aber nicht nur die konstante Lösungswahrscheinlichkeit notwendig, die X (g, i) müssen auch stochastisch unabhängig sein. Dies ist für die Personen sicher unkritisch, aber die Unabhängigkeit der X (g, i) über die Testaufgaben einer Person hinweg ist ein Postulat, das zum Beispiel wegen Lerneffekten verletzt sein kann. Ist die Unabhängigkeit jedoch nicht gewährleistet, so kann das Modell nicht auf den Wahrscheinlichkeiten $\mathbb{P}\{X(g,i)=1\}$ aufgebaut werden, sondern man muß die Verbundwahrscheinlichkeit des gesamten Datenvektors einer Person verwenden. Mit beispielsweise 20 Aufgaben könnte X dann 2^{20} verschiedene Werte annehmen und eine Schätzung der Verteilungsfunktion von X wäre praktisch unmöglich. Neben diesen unverzichtbaren Forderungen wollen wir noch davon ausgehen, daß für Personen und Aufgaben jeweils ein einziger Parameter ausreicht um deren Einfluß auf die Wahrscheinlichkeit P zu beschreiben. Der Zusammenhang zwischen den Elementen h aus A, i aus B und der Abbildung P soll durch ein probabilistisches Meßmodell beschrieben werden. Diese Forderungen lassen sich, wie wir weiter unten zeigen werden, in der folgenden Bedingung zusammenfassen:

Definition 2. Ein System $< A,B,P >$ heißt *schwach ordnungsunabhängig* oder kürzer *schwach unabhängig* g.d.w. für alle g, h aus A und alle i, j aus B gilt

$$\exists i(P(g,i) < P(h,i)) \rightarrow \forall i(P(g,i) \leq P(h,i)) \quad (1\,a)$$

$$\exists g(P(g,i) < P(g,j)) \rightarrow \forall g(P(g,i) \leq P(g,j)). \quad (1\,b)$$

Die Bedingung (1 a) verlangt, daß eine Person g, die bei einer einzigen Aufgabe i eine kleinere Lösungswahrscheinlichkeit hat als die Person h, bei keiner anderen Aufgabe eine größere Lösungswahrscheinlichkeit haben darf. Bedingung (1 b) ist eine analoge Forderung für zwei Testaufgaben i und j. Damit können alle Zeilen und Spalten einer Matrix von Antwortwahrscheinlichkeiten mit der gleichen Reihenfolge von Personen und Aufgaben monoton geordnet werden. (1 a) und (1 b) stellen erheblich schwächere Restriktionen an die Daten als das Rasch-Modell, denn dieses verlangt über (1 a) und (1 b) hinaus, daß die Quotienten aus Lösungs- und Gegenwahrscheinlichkeit zweier Personen über alle Aufgaben proportional sind (vgl. Hamerle & Tutz, 1980). Die Bedingungen (1 a) und (1 b) sind jedoch stark genug, um zu garantieren, daß die Rangordnung der Antwortwahrscheinlichkeiten sowohl für die Aufgaben als auch für die Personen von einem einzigen Skalenwert bestimmt wird und der Vergleich zweier Personen unabhängig davon ist, welche speziellen Aufgaben dazu verwendet werden. Dies zeigt der folgende Satz:

Satz 1. Ein endliches System $< A, B, P >$ ist schwach unabhängig g.d.w. es eine Funktion u: A → ℝ und eine Funktion v: B → ℝ gibt, so daß für alle g, h aus A und alle i, j aus B gilt

$$u(g) < u(h) \leftrightarrow \exists i(P(g,i) < P(h,i)) \tag{2a}$$

$$v(i) < v(j) \leftrightarrow \exists g(P(g,i) < P(g,j)). \tag{2b}$$

Die Funktionen u und v sind eindeutig bis auf streng monotone Transformationen.

Der Beweis von Satz 1 ist sehr einfach und wird deshalb nur kurz skizziert. Man definiert auf A und B zwei Relationen und zeigt, daß sie schwache Ordnungen sind. Mit g, h aus A und i, j aus B gelte

$$g \leq_A h \leftrightarrow \forall i(P(g,i) \leq P(h,i))$$

$$i \leq_B j \leftrightarrow \forall g(P(g,i) \leq P(g,j)).$$

Es ist leicht zu zeigen, daß $\leq_A$ und $\leq_B$ konnex und transitiv sind. Damit sind $< A, \leq_A >$ und $< B, \leq_B >$ schwache Ordnungen und es gibt die bis auf streng monotone Transformationen eindeutigen Abbildungen u: A → TR und v: B → TR, so daß gilt

$$u(g) \leq u(h) \leftrightarrow g \leq_A h$$

$$v(i) \leq v(j) \leftrightarrow i \leq_B j$$

(vgl. Krantz, Luce, Suppes & Tversky, 1971, S. 14–15). Hieraus folgt sofort

$$u(g) < u(h) \leftrightarrow \exists i(P(g,i) < P(h,i))$$

$$v(i) < v(j) \leftrightarrow \exists g(P(g,i) < P(g,j)).$$

Die schwache Unabhängigkeit ist damit hinreichend für die Existenz von u und v. Die Notwendigkeit ergibt sich unmittelbar aus der Annahme

$$\exists i(P(g,i) < P(h,i)) \wedge \exists i(P(h,i) < P(g,i)),$$

die bei Gültigkeit von (2a) zu einer Kontradiktion führt. Entsprechendes gilt für (2b).

Die schwache Ordnungs-Unabhängigkeit wurde aus meßtheoretischer Sicht zuerst von Fishburn (1973) untersucht. Dort sind auch bereits die wesentlichen Aussagen von Satz 1 bewiesen. Daneben zeigt Fishburn, daß es in einem schwach unabhängigen System reelle Funktionen u′ auf A und v′ auf B und eine in beiden Argumenten nicht fallende Funktion F auf u′(A) x v′(B) gibt, so daß gilt

$$P(g,i) = F[u'(g), v'(i)]. \tag{3}$$

Es läßt sich zeigen, daß in einem endlichen System $< A, B, P >$ die schwache Unabhängigkeit notwendig und hinreichend ist für dieses Modell, das von Fishburn „monotone Skalierbarkeit" genannt wird. Jedoch erfüllen u′ und v′ nicht notwendigerweise (2a) und (2b). Dies folgt daraus, daß wegen (2a) u(g) = u(h) g.d.w. für alle i P(g, i) = P(h, i). Jedoch folgt wenn für alle i P(g, i) = P(h, i) nicht u′(g) = u′(h). Damit ist u′ keine Repräsentation die (2a) erfüllt (und auch keine Ordinalskala).

Verlangt man, daß die Funktion F in beiden Argumenten streng monoton steigend ist, dann wird aus (3) die „einfache Skalierbarkeit". Notwendig und hinreichend hierfür ist die „einfache Unabhängigkeit"

$$\exists i(P(g,i) \leq P(h,i)) \rightarrow \forall i(P(g,i) \leq P(h,i))$$

$$\exists g(P(g,i) \leq P(g,j)) \rightarrow \forall i(P(g,i) \leq P(g,j)).$$

Dieses Modell wurde zuerst von Krantz (1967) im Bereich der Psychophysik untersucht und danach von einigen anderen Autoren aufgenommen (Tversky & Russo, 1969; Fishburn, 1973; Doignon & Falmagne, 1974; Falmagne & Iverson, 1979; Holman, 1979). Bei Krantz, e. al., (1971) erscheint es in algebraischer Form unter dem Begriff „decomposability". Das Rasch-Modell ist ein Spezialfall von (3), wobei für F(x, y) die logistische Funktion der Summe (x + y) eingesetzt wird:

$$P(h,i) = 1/[1 + \exp(u(h) + v(i))].$$

Rasch (1961) kam bei seinen Überlegungen zur Objektivität von Messungen in der Psychologie zu dem Schluß, daß ein Modell, das der einfachen Unabhängigkeit äquivalent ist, notwendig und hinreichend sei für „unrestricted and transitive comparability of both individuals and stimuli" (Rasch, 1961, S. 332).

Satz 1 zeigt, daß auch die schwache Unabhängigkeit ausreicht, um Personen und Aufgaben unabhängig voneinander auf einer Ordinalskala zu skalieren. Die von Rasch (1961) geforderte Unabhängigkeit des Personenvergleichs von den verwendeten Aufgaben erfährt bei der schwachen Unabhängigkeit jedoch eine geringfügige Abschwächung. Die Forderung von Rasch ist, daß die Antwortwahrscheinlichkeiten zweier Personen g und h bei der selben Aufgabe i vollständig und unabhängig von den anderen Aufgaben die Ordnung der Skalenwerte der Personen bestimmen muß. Es muß also gelten

$$P(g,i) < P(h,i) \leftrightarrow u(g) < u(h).$$

In einem schwach skalierbaren System gilt dagegen nur

$$P(g,i) < P(h,i) \rightarrow u(g) < u(h)$$

$$u(g) < u(h) \rightarrow P(g,i) \leq P(h,i).$$

Damit bestimmt die Aufgabe i die Ordnung der Skalenwerte u(g) und u(h) nur dann vollständig, wenn sich P(g, i) und P(h, i) echt unterscheiden. Sind P(g, i) und P(h, i) gleich, so folgt aus der einfachen Unabhängigkeit, daß auch die Skalenwerte u(g) und u(h) gleich sein müssen. Die schwache Unabhängigkeit dagegen läßt bei P(g, i) = P(h, i) zu, daß u(g) < u(h), nämlich genau dann, wenn eine andere Aufgabe j existiert bei der P(g, j) < P(h, j). Allerdings kann es dann auch bei der schwachen Unabhängigkeit keine Aufgabe j' geben, so daß P(h, j') < P(g, j'). Die Abschwächung der Vergleichbarkeit beim Übergang von der einfachen zur schwachen Unabhängigkeit ist für die praktische Anwendung deshalb nicht sehr bedeutsam. Sie besteht im wesentlichen darin, daß Aufgaben (und Personen) mit unterschiedlicher „Trennschärfe" zugelassen werden: So kann wenn P(g, i) = P(h, i) die Aufgabe i die Personen g und h nicht trennen, während bei P(g, j) < P(h, j) die Aufgabe j zur Unterscheidung von g und h beiträgt.

Wegen der Ordnungs-Unabhängigkeit der Parameter in diesem Modell werden wir im folgenden ein System < A, B, P > das schwach ordnungs-unabhängig ist, als *U-Modell* bezeichnen.

Ein besonderes Problem aller probabilistischen Testmodelle ist die Schätzung der Wahrscheinlichkeiten P(g, i), da bei psychologischen Tests in der Regel keine wiederholten Beobachtungen möglich sind. Beim Rasch-Modell läßt sich dieses Problem lösen, da die Summe der richtigen Lösungen einer Person g eine suffiziente Statistik für den Personenparameter u(g) ist (Fischer, 1974, S. 193 ff.) und damit die gesamte Information der Daten über die Person g enthält. Personen mit der gleichen Summe richtiger Lösungen müssen den gleichen Parameter erhalten und können so als Meßwiederholungen betrachtet werden, auch wenn diese Klassifikation erst nach der Datenerhebung erfolgen kann.

Mokken (1971) untersuchte einige probabilistische Varianten der bekannten Guttman-Skala, darunter auch solche, die den hier vorgeschlagenen Modellen äquivalent sind. Im Unterschied zu unserer an der Meßtheorie orientierten Darstellung betrachtet Mokken vor allem die statistischen Probleme, die bei der Anwendung dieser Modelle auftreten. Er geht davon aus, daß die Mengen A und B Stichproben aus Personen- bzw. Aufgabenpopulationen sind und daß für diese Populationen die schwache oder die einfache Skalierbarkeit gilt. Hieraus werden dann für die Stichproben A und B einige notwendige Bedingungen abgeleitet. Insbesondere zeigt Mokken (1971, S. 122 ff.) auch, daß bei Gültigkeit der einfachen Skalierbarkeit die aus den Stichproben geschätzten Randwahrscheinlichkeiten konsistente Schätzungen der Populationsparameter sind. Damit können aus diesen Randwahrscheinlichkeiten Schätzungen für die Ordnungen von Personen und Aufgaben gewonnen werden, die unabhängig von der speziellen Verteilung der Parameter in der Population sind.

Die folgende Überlegung zeigt, daß im U-Modell der Rangplatz der Summe richtiger Lösungen die gesamte Information über die unbekannten Parameter enthält. Seien

$$t(g,*) = \sum_{i \in B} X(g,i) \text{ und } t(*,i) = \sum_{g \in A} X(g,i)$$

die Summenstatistiken der Person g und der Aufgabe i, dann gilt im U-Modell für deren Erwartungwerte unter der Bedingung, daß die X(g, i) stochastisch unabhängig sind,

$$E\{t(g,*)\} = \sum_{i \in B} P(g,i) \text{ und}$$

$$E\{t(*,i)\} = \sum_{g \in A} P(g,i).$$

Gilt nun für zwei Personen g und h aus A die Ungleichung $u(g) \leq u(h)$, so folgt hieraus für alle i aus B $P(g,i) \leq P(h,i)$ und damit auch

$$E\{t(g,*)\} \leq E\{t(h,*)\}. \tag{4}$$

Dieser Schluß gilt im U-Modell auch umgekehrt. Denn gäbe es ein i mit $P(h,i) < P(g,i)$ so müßte es wegen (4) mindestens ein j geben mit $P(g,j) < P(h,j)$, woraus ein Widerspruch zu (1 a) folgt. Da die Parameter u und v nur eindeutig bis auf streng monotone Transformationen sind, ergibt sich daraus, daß die Randsummen t(g, *) und t(*, i) die gesamte Information der Daten über die Skalenwerte u und v enthalten. Wir benützen die Randsummen, um vom System $< A, B, P >$ auf das System $< \underline{A}, \underline{B}, p >$ der Rohwert-Äquivalenzklassen überzugehen. Dabei werden die Randsummen als bekannt

vorausgesetzt. $\overline{A}$ und $\overline{B}$ sind dann die Mengen von Personen- bzw. Itemrohwertgruppen:

$$\overline{A} = \{\bar{g} \mid \bar{g} \subseteq A \wedge [g \in \bar{g} \wedge g' \in \bar{g} \leftrightarrow t(g,*) = t(g',*)]\}$$
$$\overline{B} = \{\bar{i} \mid \bar{i} \subseteq B \wedge [i \in \bar{i} \wedge i' \in \bar{i} \leftrightarrow t(*,i) = t(*,i')]\}.$$

Die $X(g,i)$ sollen für alle (g,i) aus $\bar{g} \times \bar{i}$ identisch verteilt sein. $p(\bar{g},\bar{i})$ sei die Wahrscheinlichkeit mit der eine Person der Rohwertgruppe $\bar{g}$ eine Aufgabe der Menge $\bar{i}$ richtig löst. Wir werden später die Zufallsvariablen

$$Z(\bar{g},\bar{i}) = \sum_{i \in \bar{i}} \sum_{g \in \bar{g}} X(g,i)$$

betrachten. Sie sind binomialverteilt mit den Parametern $p(\bar{g},\bar{i})$ und $n(\bar{g},\bar{i}) = |\bar{g} \times \bar{i}|$. Da durch den Kontext Verwechslungen ausgeschlossen sind, schreiben wir im folgenden jedoch statt $\bar{g}$ bzw. $\bar{i}$ einfach g bzw. i und benützen die mit $\bar{g}$ und $\bar{i}$ definierten Ausdrücke in der Notation $p(g,i)$, $Z(g,i)$ und $n(g,i)$. Bei der Testdurchführung sollen genau $m = |\overline{A}|$ verschiedene Personen- und $k = |\overline{B}|$ verschiedene Aufgabengruppen vorgekommen sein, also $g = 1, \ldots, m$ und $i = 1, \ldots, k$. Unsere weiteren Überlegungen werden sich ausschließlich mit einem ordnungs-unabhängigen System $< \overline{A}, \overline{B}, p >$ befassen.

3. Ein statistischer Test des U-Modells

Ein statistischer Test, der für das U-Modell geeignet ist, wurde von Schaafsma (1966) entwickelt (vgl. auch Schaafsma & Smid, 1966). Dabei wird von den Zufallsvariablen $Z(h,i)$, $(h = 1, \ldots, m; i = 1, \ldots, k)$ ausgegangen, deren binomiale Verteilungen durch $\mathfrak{B}[n(h,i), p(h,i)]$ beschrieben seien. Der Test prüft die Hypothese

$$H: p(1,1) = \ldots = p(1,k) = \ldots = p(h,i) = \ldots = p(m,k)$$

gegen die Alternativhypothese

K: $p(h,1) \leq \ldots \leq p(h,i) \leq \ldots \leq p(h,k)$, $(h = 1, \ldots, m)$
und mindestens eine Ungleichung gilt streng;
$p(1,i) \leq \ldots \leq p(h,i) \leq \ldots \leq p(m,i)$, $(i = 1, \ldots, k)$
und mindestens eine Ungleichung gilt streng.

Die Hypothese besteht also aus (mk-1) Gleichungen und die Alternative aus (2mk-m-k) Ungleichungen. Die Prüfgröße des Tests basiert auf einem linearen Kontrast der Zufallsvariablen $Z(h,i)$:

$$Y = \sum_{h=1}^{m} \sum_{i=1}^{k} w(h,i)\, Z(h,i), \text{ mit}$$

$$\sum_{h=1}^{m} \sum_{i=1}^{k} w(h,i) = 0.$$

Die Bestimmung der Gewichte w(h, i) erfolgt nach einem Kriterium, das von Abelson & Tukey (1963) vorgeschlagen wurde. Im folgenden schreiben wir für einen mk-dimensionalen Vektor [y(1,1) ... y(h, i) ... y(m, k)] kürzer [y(h, i)]. Der Vektor [p(h, i)] wird als Punkt in einem mk-dimensionalen Vektorraum R^{mk} aufgefaßt. Die Alternativhypothese K ist ein mk-dimensionaler Kegel in R^{mk}, der Durchschnitt von (mk - 1) Halbräumen, dessen Spitze in H liegt. Die Mantelflächen des Kegels sind Hyperebenen, bzw. Durchschnitte von Hyperebenen, die aus K entstehen, wenn einzelne oder mehrere der Ungleichungen zu Gleichungen werden. Das Verfahren von Abelson & Tukey (1963) sucht innerhalb des Kegels den Punkt [w(h, i)] so, daß eine Gerade d, die durch [w(h, i)] und die Spitze des Kegels bestimmt ist, den maximalen Winkel zwischen d und den Kanten des Kegels minimiert. Dies führt im Fall der Restriktionen K dazu, daß die in diesem Sinne optimale Gerade d* mit einem Teil der Kanten von K den gleichen Winkel ψ und mit allen anderen Kanten einen kleineren Winkel bildet. Die Prüfgröße des Tests ist dann die Projektion des Stichprobenpunktes auf d*. Dieser Test – wir nennen ihn φ – testet die Hypothese, daß [p(h, i)] in H liegt gegen die Alternative, daß [p(h, i)] nicht in H aber auf d* liegt. Die Bestimmung von d* wird mit Überlegungen zur Macht des Tests φ begründet. Der Test ist approximativ „most stringent, somewhere most powerful" (Schaafsma, 1966) bezüglich der Tests zum Niveau α, die H gegen K testen. Dies bedeutet, daß es in K Punkte gibt, zu deren Test gegen H der Test φ ein mächtigster Test ist („somewhere most powerful") und daß die maximale Differenz an Macht zu allen anderen Tests zum Niveau α von H gegen K von φ minimiert wird („most stringent"). Der Test approximiert unbekannte Parameter durch normalverteilte Schätzungen, deshalb gelten die genannten Kriterien nur für große n(h, i).

Um diesen Test anzuwenden, ist die Bestimmung der Gewichte w(h, i) notwendig. Während sich für den Fall eines allgemeinen Systems linearer Ungleichungen hierfür keine explizite Lösung angeben läßt, wurden von Abelson & Tukey (1963) und Schaafsma (1966) für einige Spezialfälle, insbesondere den der einfachen Ordnung, explizite Lösungen angegeben[1]. Die von Schaafsma (1966, S. 101 ff.) ausführlich dargestellten Überlegungen, die eine Lösung für den Sonderfall der Hypothese K mit m = 1 ergeben, können aber in einfacher Weise auf den vorliegenden Fall ausgedehnt werden. Um die daraus folgende Lösung für die Komponenten w(h, i) besser darstellen zu können, werden ausgehend von den Zufallsvariablen Z(h, i) und ihren Verteilungsparametern n(h, i) und p(h, i) folgende Ausdrücke definiert:

$$n = \sum_{h=1}^{m} \sum_{i=1}^{k} n(h,i),$$

$$Z = \sum_{h=1}^{m} \sum_{i=1}^{k} Z(h,i),$$

$$p = Z/n$$

1 Für den Fall eines allgemeinen Systems linearer Ungleichungen wurde von McClelland & Coombs (1975) ein Computerprogramm (ORDMET) entwickelt (vgl. auch Lehner & Noma, 1980), das die Abelson & Tukey-Lösung des Ungleichungssystems liefert und damit auch zum Auffinden der Gewichte herangezogen werden kann.

$$s(f,g) = \sum_{h=1}^{g} \sum_{i=1}^{f} n(h,i),$$

$$s(0,f) = s(g,0) = s(0,0) = 0.$$

Die Formel für die einzelnen Komponenten $w(h,i)$ lautet dann:

$$\begin{aligned} w(h,i) = n(h,i)^{-1} n^{1/2} \Big(& - \{s(h,i)\,[n - s(h,i)]\}^{1/2} \\ & + \{s(h-1,i)\,[n - s(h-1,i)]\}^{1/2} \\ & + \{s(h,i-1)\,[n - s(h,i-1)]\}^{1/2} \\ & - \{s(h-1,i-1)\,[n - s(h-1,i-1)]\}^{1/2} \Big). \end{aligned}$$

Betrachtet man nun für gegebenes p, das als bekannt vorausgesetzt wird, die Verteilung der Statistik

$$T' = [p(1-p)]^{-1/2} \sum_{h=1}^{m} \sum_{i=1}^{k} w(h,i)\, Z(h,i),$$

so hat diese unter H den Erwartungswert 0 und die Varianz

$$\mathrm{Var}(T') = n/(n-1) \sum_{h=1}^{m} \sum_{i=1}^{k} n(h,i)\, w(h,i)^2.$$

Demnach ist

$$T = \frac{[p(1-p)]^{-1/2} \sum_{h=1}^{m} \sum_{i=1}^{k} w(h,i)\, Z(h,i)}{\left(n/(n-1) \sum_{h=1}^{m} \sum_{i=1}^{k} n(h,i)\, w(h,i)^2\right)^{1/2}} \tag{5}$$

approximativ standardnormalverteilt und die Hypothese H wird abgelehnt, wenn $T \geq z(\alpha)$ (Schaafsma, 1966, S. 105).

Die Anwendung dieses Tests zur Prüfung des U-Modells setzt die Kenntnis der Ordnungen in K voraus. Bei einer Testkonstruktion muß diese Ordnung aber aus den Daten gewonnen werden. Würde man zur Bestimmung der Personen- und Aufgabengruppen die gesamte Datenmatrix der Werte $X(h,i)$ einfach nach den Randsummen ordnen, so wäre die Hypothesenbildung nicht unabhängig von der statistischen Überprüfung. Selbst bei Gültigkeit der Nullhypothese H wäre dann eine Ablehnung von H im Test zu erwarten, da die Fehlerstreuungen der Randsummen als „wahre" Ordnungen interpretiert würden. Dieser Fehler kann jedoch leicht durch eine Halbierung der Ausgangsstichprobe vermieden werden. Dabei wird die eine Hälfte der Daten zur Bestimmung der Personen- und Aufgabenordnungen verwendet, die zweite Hälfte wird dann nach diesen Ergebnissen geordnet und diese Ordnungen werden in der Alternativhypothese getestet.

4. Durchführung des statistischen Tests[2]

Als Beispiel zum statistischen Test des U-Modells wird die Datenmatrix von Herbst (1978) verwendet[3]. Sie entstand aus einer Menge A von 463 Studenten und aus einer Menge B von 48 Mathematikaufgaben. Die 463 x 48-Matrix der X (h, i) wird in vier Teilmatrizen aufgeteilt. Dazu werden die Personen und die Aufgaben auf jeweils zwei Teilmengen A_1, A_2 und B_1, B_2 verteilt, wodurch sich vier Teilmatrizen bilden lassen: A_1 x B_1, A_1 x B_2, A_2 x B_1 und A_2 x B_2. Die Teilmatrizen A_1 x B_1 und A_2 x B_2 werden zur Bestimmung der Ordnungen von Personen und Aufgaben verwendet und mit den Teilmatrizen $A_1 \times B_2$ und $A_2 \times B_1$ wird der statistische Test durchgeführt. Genauer bedeutet dies, daß die Rohwertgruppe, zu der eine Person aus A_1 gehört, durch die Summe der richtigen Lösungen bei den Aufgaben B_1 bestimmt wird. Jede Person aus A_1 wird dann für die Testmatrix $\overline{A}_1 \times \overline{B}_2$ der Personengruppe zugeordnet, die bezüglich der Aufgaben B_1 ihre Rohwertgruppe ist. Die Aufgaben B_2 werden in der Matrix $\overline{A}_1 \times \overline{B}_2$ nach den Randsummen der Lösungshäufigkeiten geordnet, die durch die Personen A_2 erzielt wurden. Entsprechend werden auch die Personen- und Aufgabengruppen der Datenmatrix $A_2 \times B_1$ bestimmt. Dieses Verfahren wurde bei der Datenmatrix von Herbst (1978) durchgeführt. Dabei wurde die Personenmenge A so geteilt, daß A_1 die Personen der unteren und A_2 die der oberen Rohwertgruppen enthielt. Die Aufgaben wurden ebenfalls nach den Rohwertgruppen geteilt, so daß damit auch geprüft wird, ob die Ordnungen der Lösungswahrscheinlichkeiten in den unteren und oberen Fähigkeits- bzw. Schwierigkeitsbereichen gleich sind. Da nicht alle möglichen Rohwertgruppen vertreten waren, ergaben sich für die Prüfmatrizen $\overline{A}_1 \times \overline{B}_2$ und $\overline{A}_2 \times \overline{B}_1$ die Größen 11 x 19 und 12 x 18.

Die Prüfgröße T aus Gleichung (5) kann für jede der beiden Testmatrizen einzeln berechnet werden, es besteht jedoch auch die Möglichkeit einen gemeinsamen Wert für beide Matrizen zu bestimmen. Dazu wird die Summe der Statistiken T' für beide Teilmatrizen betrachtet:

$$T'' = T'(\overline{A}_1 \times \overline{B}_2) + T'(\overline{A}_2 \times \overline{B}_1).$$

Die gleichen Überlegungen, die zur Prüfgröße T führten, ergeben in diesem Fall die Prüfgröße

$$T = T''/[Q(\overline{A}_1 \times \overline{B}_2) + Q(\overline{A}_2 \times \overline{B}_1)]^{1/2},$$

wobei $Q(\overline{A}_i \times \overline{B}_j)$ der aus $\overline{A}_i \times \overline{B}_j$ berechnete Nenner der Prüfgröße T ist. Tabelle 1 gibt die Werte der Prüfgrößen für die Teilmatrizen einzeln und gemeinsam an. In jedem Fall muß mit einem kritischen Wert z (.05) = 1.64 die Hypothese H abgelehnt werden. Neben den Prüfgrößen enthält Tabelle 1 auch die zu den jeweiligen Tests gehörenden Winkel ψ, die von Schaafsma (1966, S. 37) als Anhaltspunkte für die Effektivität des Tests vorgeschlagen wurden, da sie die Macht des Tests φ gegenüber anderen Tests der Hypothese H gegen K beeinflussen. Aus den Überlegungen von Schaafsma, auf die hier

---

2 Die Berechnungen zur Parameterschätzung und zum Modelltest wurden am Computer Laboratory for Instruction and Psychological Research, das von der University of Colorado unterstützt wird, und am Rechenzentrum der Universität Regensburg durchgeführt.

3 Wir danken Herrn Dr. K. Herbst für das Überlassen der von ihm erhobenen Daten.

nicht näher eingegangen werden soll, geht hervor, daß die Winkel ψ in unserem Fall bezogen auf die Zahl mk als klein zu bezeichnen sind und der Test für das vorliegende Problem gut geeignet ist.

Tabelle 1. Ergebnisse des statistischen Tests

Teilmatrix:	$\overline{A}_1 \times \overline{B}_2$	$\overline{A}_2 \times \overline{B}_1$	Gesamt
Prüfgröße:	8.0	16.4	15.8
Winkel ψ:	75.7	68.9	78.2

Bei der Konstruktion eines psychologischen Tests kann es manchmal erwünscht sein, die Testaufgaben einzeln auf die Erfüllung der Restriktionen K zu prüfen, um eine Selektion der Aufgaben durchzuführen. Der vorgeschlagene Test ermöglicht dies dadurch, daß die Prüfgröße T auch für jede Aufgabe einzeln berechnet werden kann (mit $k = 1$). Dazu empfiehlt es sich jedoch, die Daten der Teilmatrizen $A_1 \times B_1$ bzw. $A_2 \times B_2$ zu verwenden, da dann die endgültige Aufgabenmenge in den unabhängigen Modelltest der Matrizen $A_1 \times B_2$ und $A_2 \times B_1$ eingebracht werden kann. Tabelle 2 zeigt die Prüfgröße T für die einzelnen Aufgaben der Menge B_2. Man sieht, daß bei einem kritischen Wert von $z(.05) = 1.64$ für die Aufgaben 10 und 34 die Nullhypothese beibehalten werden kann. Die Lösungswahrscheinlichkeiten der verschiedenen Personengruppen erfüllen deshalb bei diesen Aufgaben nicht die geforderten Ordnungsrestriktionen. Sie liefern damit auch keine nennenswerte Information zur Trennung der Personen und können aus der endgültigen Aufgabenmenge weggelassen werden.

Tabelle 2. Die Werte der Prüfgrößen T für die 24 Aufgaben aus B_2. Die Aufgabennummern sind von Herbst (1978) übernommen und hier nach fallender Schwierigkeit geordnet. Mit Ausnahme von {15,39}, {14, 17, 33, 37, 46} und {11, 26} bestehen die Aufgabengruppen in B_2 aus einzelnen Aufgaben.

15	39	14	17	33	37	46	23
2.75	5.65	3.01	4.92	5.53	4.88	5.53	4.45
21	28	20	11	26	16	22	3
7.07	4.70	6.50	6.38	5.01	3.59	4.65	2.31
35	31	41	38	10	24	5	34
2.37	4.59	5.33	1.72	1.59	5.71	9.40	-.16

5. Maximum Likelihood Schätzung der Parameter des U-Modells

Eine Schätzung von Wahrscheinlichkeiten p(g, i), die (1 a, b) erfüllen, ist aus mehreren Gründen von Interesse. Sie ermöglicht, wie wir später zeigen werden, eine Beurteilung der Diskriminationsfähigkeit des Tests in Abhängigkeit vom Leistungsniveau der Personen und kann damit zur Itemselektion benützt werden. Darüber hinaus ergibt sich bei Kenntnis der p(g, i) und der Funktionen u und v die Möglichkeit, den Parameter

u(g) einer Person g mit einer begrenzten Auswahl von Testaufgaben zu schätzen. Wir werden in Abschnitt 7 kurz erläutern, wie dies zur Entwicklung eines individualisierten Testverfahrens benützt werden kann.

Berechnet man aus den erhobenen Daten die relativen Lösungshäufigkeiten q(g, i), mit der eine Personengruppe g eine Aufgabe der Gruppe i gelöst hat, so wird man im allgemeinen feststellen, daß diese Werte die Restriktionen des U-Modells nicht erfüllen. Da alle empirischen Beobachtungen Zufallseinflüssen unterliegen, werden Verletzungen von (1 a, b) selbst dann auftreten, wenn die erhobenen Daten Realisationen eines U-Systems sind. Hat man an Hand des dargestellten statistischen Tests entschieden, daß die vorliegenden Daten die Restriktionen des U-Modells erfüllen, können (1 a) und (1 b) als Nebenbedingungen im Schätzverfahren verwendet werden. Man gewährleistet dadurch, daß die geschätzten Wahrscheinlichkeiten modellkonform sind. Wir werden im folgenden die Maximum Likelihood Schätzung der Parameter p(g, i) beschreiben unter der Nebenbedingung, daß (1 a, b) erfüllt werden.

Die Schätzung der Skalenwerte u(g) und v(i) aus den Randsummen t(g, *) und t(*, i) wurde bereits dargestellt. Aus dieser Schätzung folgt

$$E\{t(g,*)\} < E\{t(h,*)\} \leftrightarrow u(g) < u(h).$$

Die gleiche Beziehung gilt analog für Skalenwerte v(i) und v(j). Für alle modellkonformen Lösungswahrscheinlichkeiten p muß deshalb gelten

$$u(g) < u(h) \rightarrow \forall i(p(g,i) \leq p(h,i)) \tag{6a}$$

$$v(i) < v(j) \rightarrow \forall g(p(g,i) \leq p(g,j)). \tag{6b}$$

Wir werden im folgenden eine ML-Schätzung beschreiben, deren Schätzwerte p (6 a, b) für beliebige Funktionen u und v erfüllen. Bestimmt man die Skalen u und v aus den Randsummen einer Datenmatrix, so können u und v immer so festgelegt werden, daß (6 a, b) und (2 a, b) gleichzeitig gelten, so daß die geschätzten Parameter das U-Modell erfüllen. Zunächst wird nun der Minimum Lower Set Algorithmus (MILSA) dargestellt, der für $\overline{A} \times \overline{B}$ eindeutige Schätzwerte p(h, i) bestimmt, die (6 a, b) erfüllen. Zur Beschreibung von MILSA benötigen wir einige Definitionen (vgl. Barlow, Bartholomew, Bremner & Brunk, 1972, S. 76):

Definition 3. Eine Teilmenge $L \subseteq \overline{A} \times \overline{B}$ ist eine *niedrige Menge* von $\overline{A} \times \overline{B}$ bezüglich (6 a, b) g.d.w. für alle g, h aus $\overline{A}$ und alle i, j aus $\overline{B}$ gilt

$$(h,j) \in L \wedge u(g) \leq u(h) \rightarrow (g,j) \in L$$

$$(h,j) \in L \wedge v(i) \leq v(j) \rightarrow (h,i) \in L.$$

Eine Teilmenge $U \subseteq \overline{A} \times \overline{B}$ ist eine *hohe Menge* von $\overline{A} \times \overline{B}$ bezüglich (6 a, b) g.d.w. für alle g, h aus $\overline{A}$ und alle i, j aus $\overline{B}$ gilt

$$(g,i) \in U \wedge u(g) \leq u(h) \rightarrow (h,i) \in U$$

$$(g,i) \in U \wedge v(i) \leq v(j) \rightarrow (g,j) \in U.$$

Eine Menge N ist eine *Niveaumenge* von $\overline{A} \times \overline{B}$ bezüglich (6 a, b) g.d.w. eine niedrige Menge L und eine hohe Menge U von $\overline{A} \times \overline{B}$ bezüglich (6 a, b) existieren, so daß $N = L \cap U$.

Nun läßt sich MILSA sehr einfach darstellen:

1. Bezeichne $\overline{A} \times \overline{B}$ als aktive Menge C.
2. Erstelle die Menge aller niedrigen Mengen L der aktiven Menge C und berechne für jede Menge L die relative Lösungshäufigkeit q(L).
3. Bestimme L*, die größte Menge L mit der kleinsten Lösungswahrscheinlichkeit q(L). q(L*) ist der Schätzwert p für die Elemente der Menge L*.
4. Nun wird eine neue aktive Menge C definiert: C: = C - L*. Falls C die leere Menge ist, dann ist jedem Element (g, i) aus $\overline{A} \times \overline{B}$ ein Schätzwert p(g, i) zugeteilt. Falls C nicht leer ist, werden die Schritte 2,3 und 4 wiederholt.

Das Resultat von MILSA ist die Kleinst Quadrate (KQ-) Schätzung der Parameter p(h, i) des U-Modells, d. h. die Summe der quadrierten Abweichungen der beobachteten Häufigkeiten von den nach (6 a, b) zu erwartenden Häufigkeiten wird durch p minimiert. Wir wollen hier den Beweis für diese Aussage, der bei Barlow e. al. (1972) nachgelesen werden kann, nur skizzieren. Aus der Theorie der Quadratischen Programmierung (Hadley, 1964) ist bekannt, daß eine eindeutige KQ-Lösung des vorliegenden Problems existiert. Falls die q(h, i) (6 a, b) bereits erfüllen, berechnet MILSA p(h, i) = q(h, i). Falls jedoch eine Ordnungsrestriktion verletzt ist, also z. B. gilt q(h. i) < q(g, i) und u(g) < u(h) und kein g' existiert, so daß u(g) < u(g') < u(h), dann ist die KQ-Schätzung gegeben durch

$$p(g,i) = p(h,i) = \frac{n(g,i)\, q(g,i) + n(h,i)\, q(h,i)}{n(g,i) + n(h,i)}.$$

Denn nimmt man an, daß die KQ-Lösung p'(g, i) ≠ p'(h, i) liefert, dann existiert ein positives e, so daß p(g, i) - e und p(h, i) + e (6 a, b) erfüllen und einen kleineren KQ-Wert liefern. Da dies zu einem Widerspruch führt, muß gelten: p(g, i) = p(h, i). Aus der Statistik ist jedoch bekannt, daß der Mittelwert die KQ-Funktion minimiert. Falls also Ordnungsverletzungen vorliegen, erhält man die KQ-Schätzwerte durch Berechnung des arithmetischen Mittels. Verallgemeinert man dieses Resultat von p(g, i) und p(h, i) auf die gesamte Matrix $\overline{A} \times \overline{B}$ so erhält man genau die Vorschrift MILSA.

Barlow et al. (1972, S. 91 ff.) konnten zeigen, daß KQ-Schätzungen der Parameter von Binomialverteilungen unter den Nebenbedingungen (6 a, b) gleichzeitig die Likelihoodfunktion maximieren. Der Algorithmus MILSA berechnet damit nicht nur die KQ-Schätzungen, sondern auch die damit identischen ML-Schätzungen der Lösungswahrscheinlichkeiten eines U-Systems. Im Gegensatz zur bedingten und unbedingten ML-Schätzung, die für das Rasch-Modell angewandt werden (Fischer, 1974), hat MILSA den Vorteil, daß er kein Iterationsverfahren benötigt. Daher sind die berechneten Schätzwerte nicht von einem vorgegebenen Konvergenzkriterium abhängig.

Für die praktische Anwendung hat MILSA jedoch den Nachteil, daß er für große Datenmatrizen (z. B. 50 x 50) sehr lange Rechenzeiten benötigt. Außerdem beansprucht MILSA dann am meisten Rechenzeit, wenn die empirischen Daten (6 a, b) bereits erfüllen. Die kürzeste Rechenzeit wird dagegen erreicht, wenn alle Ungleichungen verletzt sind. In diesem Fall bearbeitet MILSA nur eine aktive Menge, nämlich $C = \overline{A} \times \overline{B}$, während im ungünstigsten Fall – (6 a, b) ist erfüllt – $|\overline{A} \times \overline{B}|$ aktive Mengen zu bearbeiten sind. Da der Algorithmus aber auf Daten angewandt werden soll die (6 a, b) „im Prinzip" erfüllen, ist diese Beziehung zwischen Daten und erforderlicher Rechenzeit äußerst ungünstig.

Für Daten, die aus dem U-Modell stammen, wird L*, die niedrige Menge mit der kleinsten Lösungswahrscheinlichkeit, immer nur eine geringe Anzahl von Elementen enthalten. Daher kann man sich auf die Erstellung von niedrigen Mengen mit weniger als r Elementen beschränken, wenn man gleichzeitig eine Methode zur Entdeckung von größeren Mengen L* zur Verfügung hat. Wir schlagen deshalb einen adaptiven Minimum Lower Set Algorithmus (AMILSA) vor. Das Prinzip dieses Algorithmus besteht darin, die Erstellung von niedrigen Mengen zunächst auf Mengen mit weniger als r Elementen zu beschränken. Um jedoch auch in Situationen, wo die niedrige Menge mit der kleinsten Lösungshäufigkeit mehr als r Elemente enthält, den korrekten Schätzwert zu finden, wird die ausgewählte niedrige Menge zunächst in einem Puffer gespeichert. Dadurch kann eine vorläufige Entscheidung im weiteren Verlauf des Algorithmus korrigiert werden. Falls später eine niedrige Menge mit einer kleineren Lösungswahrscheinlichkeit erstellt wird, können die im Puffer gespeicherten Mengen zu einer neuen aktiven Menge vereinigt werden. Diese neue aktive Menge hängt von den Charakteristika des gegebenen Datensatzes ab und wird nur erstellt, wenn dies notwendig ist. Deshalb benötigt AMILSA für Datensätze, die von einem U-System erzeugt wurden, sehr viel kürzere Rechenzeiten als MILSA.

Der Algorithmus AMILSA:

1. Bezeichne $\overline{A} \times \overline{B}$ als aktive Menge C.
2. Erstelle die Menge aller niedrigen Mengen von C mit weniger als r Elementen, wähle daraus L*, die größte Menge mit der kleinsten Lösungshäufigkeit q(L) und speichere L* und q(L*) in einem Puffer der Länge k. Falls $q(L^*) < q(L')$, wobei L' ein Vorgänger von L* ist, vereinige die beiden Mengen und berechne ihre durchschnittliche Lösungshäufigkeit. Gleichzeitig wird diese Menge als vereinigte Menge markiert. Wiederhole den Test mit dem Vorgänger der vereinigten Menge, bis die Lösungshäufigkeiten im Puffer monoton ansteigen.
3. Falls der Puffer gefüllt ist (der Puffer enthält k Mengen), wird eine Entscheidung über einen Schätzwert getroffen. Durch diese Entscheidung, die sich im weiteren Verlauf des Algorithmus als fehlerhaft erweisen kann, wird der erste Speicherplatz im Puffer geleert. Dabei sind zwei Fälle zu unterscheiden:

3.1. Falls L_1, die erste Menge des Puffers, nicht markiert ist, wird der Menge L_1 $q(L_1)$ als Schätzwert zugewiesen. Falls $q(L_1)$ kleiner ist als ein früher zugewiesener Schätzwert, so wird die Berechnung beendet und die berechneten Werte als fehlerhaft bezeichnet. Andernfalls wird Schritt 4 ausgeführt.

3.2. Falls L_1 markiert ist, wird ein Durchgang von MILSA auf die aktive Menge $C = L_1$ angewandt. Die Lösungsmenge M von MILSA liefert einen (möglicherweise) korrekten Schätzwert q(M). Falls q(M) kleiner ist als ein vorausgegangener Wert, wird die Berechnung abgebrochen und die berechneten Werte als fehlerhaft bezeichnet. Ansonsten wird q(M) der Schätzwert für M. Falls $L_1 - M$ nicht leer ist, wird $L_1 - M$ als erste Menge im Puffer gespeichert, markiert und als L_1 bezeichnet. Danach wird die Ordnung zwischen $q(L_1)$ und den anderen Mengen im Puffer geprüft, Ordnungsverletzungen werden durch Vereinigung der entsprechenden Mengen korrigiert und anschließend wird Schritt 3 wiederholt.

4. Man setzt $C := C - L^*$ und wiederholt die Schritte 2 bis 4 so lange, bis die Berechnung in 3.1. abgebrochen wird oder bis C leer ist. In diesem Fall ist jedem Element aus $\overline{A} \times \overline{B}$ ein Schätzwert zugeteilt.

Für beliebige r und k erstellen MILSA und AMILSA entweder identische kleinste niedrige Mengen, oder AMILSA bricht die Berechnung als fehlerhaft ab. Der Puffer der als Warteschlange fungiert, adaptiert die Größe der aktiven Menge an den vorgegebenen Datensatz. Das folgende Argument gilt sowohl für die ursprünglichen aktiven Mengen, als auch für die aktiven Mengen, die im Puffer definiert werden. Wir bezeichnen die von MILSA bzw. AMILSA erstellten niedrigen Mengen mit L(M) bzw. L(A). Falls MILSA und AMILSA immer identische L* erstellen, d. h. $L(M)^* = L(A)^*$, führen beide

Algorithmen trivialerweise zum gleichen Ergebnis. Falls $L(M)^* \neq L(A)^*$, dann ist $L(M)^*$ größer als $L(A)^*$, denn AMILSA erstellt nur niedrige Mengen mit weniger als r Elementen. Wir definieren: $L(M)' := L(M)^* - [L(A)^* \cap L(M)^*]$. Somit ist $q[L(M)'] < q[L(A)^*]$. Außerdem gilt $L(M)' \subseteq C$, wobei C die vollständige aktive Menge von MILSA ist. Falls $|L(M)'| < r$, wird AMILSA $L(M)'$ als nächste niedrige Menge finden und daher eine größere aktive Menge definieren (Schritt 2) oder AMILSA bricht die Berechnung als fehlerhaft ab (Schritt 3.1.). Für $|L(M)'| > r$ erhält man das gleiche Ergebnis.

Damit eignet sich AMILSA besonders zur Parameterschätzung von großen Datenmatrizen. Dies soll am Beispiel von zwei 10 x 10 Matrizen, welche die Restriktionen des U-Modells mehr oder weniger erfüllen, demonstriert werden.

Als erste Testmatrix wurde eine 10 x 10 Einheitsmatrix $E(i,j)$ gewählt. Die Ordnungsrestriktionen erfordern, daß $e(i,j) < e(k,l)$ wenn $i < k$ oder $j < l$. In der zweiten Matrix, welche die Ordnungsrestriktionen des U-Modells noch seltener erfüllt, wurden zusätzlich alle Elemente der ersten zwei Zeilen gleich eins gesetzt. In beiden Fällen wurde eine Puffergröße von $k = 7$ gewählt. Tabelle 3 zeigt die durchschnittliche Anzahl von erstellten niedrigen Mengen pro Schätzwert, die Anzahl der im Puffer korrigierten Ordnungsverletzungen, sowie die Anzahl von Verletzungen im Endergebnis als Funktion des (Programm-) Parameters r. Für die Einheitsmatrix wurden 10 eindeutige Schätzwerte berechnet, während sich für die zweite Testmatrix nur 4 verschiedene Schätzwerte ergaben.

Aus Tabelle 3 erkennt man, daß die durchschnittliche Anzahl von erstellten niedrigen Mengen drastisch mit der Größe des Parameters r ansteigt. Gleichzeitig sieht man, daß für die zweite Testmatrix r kleiner als 10 oder größer als 15 gewählt werden muß, um ein korrektes Schätzergebnis zu erhalten, während für die „modellkonformere" Matrix 1 beliebige Parameter r zu einem korrekten Ergebnis führen. Die korrekten Lösungen für Matrix 2 mit $r < 10$ kommen dadurch zustande, daß im Puffer wegen häufiger Ordnungsverletzungen große aktive Mengen gebildet werden. Für $9 < r < 16$ entstehen dagegen nicht so viele Ordnungsverletzungen im Puffer. Da weder anfangs noch im Puffer genügend große aktive Mengen erzeugt werden, führt deshalb das Schätzverfahren in diesem Fall zu einem fehlerhaften Endergebnis.

Tabelle 3. Indikatoren für die erforderliche Rechenzeit von AMILSA als Funktion des Parameters r und zweier Datenmatrizen. Matrix 1, mit 10 eindeutigen Schätzwerten, erfüllt eine größere Anzahl von Ordnungsrestriktionen des U-Modells als Matrix 2 (4 eindeutige Schätzwerte).

r	mittlere Anzahl von erstellten niedr. Mengen pro Schätzwert		Im Puffer korrigierte Ordungs-Verletzungen		Anzahl von Verletzungen im Ergebnis	
	Mat. 1	Mat. 2	Mat. 1	Mat. 2	Mat. 1	Mat. 2
1	56	11001	90	97	0	0
3	68	11026	29	37	0	0
5	85	11036	16	21	0	0
10	172	3397	5	7	0	1
15	410	3840	2	5	0	1
20	752	12375	0	3	0	0
100	25094	56852	0	0	0	0

Wir können aus Tabelle 3 folgern, daß r um so größer gewählt werden muß, je mehr Ordnungsverletzungen in den gegebenen Daten vorliegen. Gleichzeitig soll r jedoch möglichst klein gehalten werden, um die erforderliche Rechenzeit zu verkürzen. Der Parameter k soll möglichst groß gewählt werden, damit auch Datensätze, die relativ viele Ordnungsverletzungen enthalten, zu einem korrekten Ergebnis führen.

Tabelle 3 verdeutlicht auch, warum AMILSA im Vergleich mit MILSA zu sehr viel kürzeren Rechenzeiten führt. Für Matrix 1 und 2 würde MILSA r = 100 wählen, während AMILSA bereits mit sehr viel kleineren Werten ein korrektes Ergebnis liefert. Wie Tabelle 3 zeigt, ist die Rechenzeitersparnis beträchtlich. Wir wollen nun AMILSA dazu benutzen, um die Parameter des U-Modells für die Daten von Herbst (1978) zu schätzen.

6. Durchführung des Schätzverfahrens

Die vollständige 463 x 48 Matrix der Rohwerte X(g, i) von Herbst (1978) wird für das Schätzverfahren zu einer 43 x 44 Matrix relativer Häufigkeiten mit 43 Personen- und 44 Itemrohwertgruppen zusammengefaßt (vgl. Abschnitt 2). Mit den Parametern k = 17 und r = 25 berechnet AMILSA die korrekten Schätzwerte für diese Matrix. Die geschätzten Wahrscheinlichkeiten sind in Abbildung 1 als itemcharakteristische Funktionen (ICC) dargestellt.

Abbildung 1 zeigt erwartungsgemäß, daß die verschiedenen Items nur in bestimmten Ausschnitten des Fähigkeitskontinuums gut differenzieren können, da die ICC's nur dort ausreichend steil verlaufen. Die Schätzung der modellkonformen Lösungs-

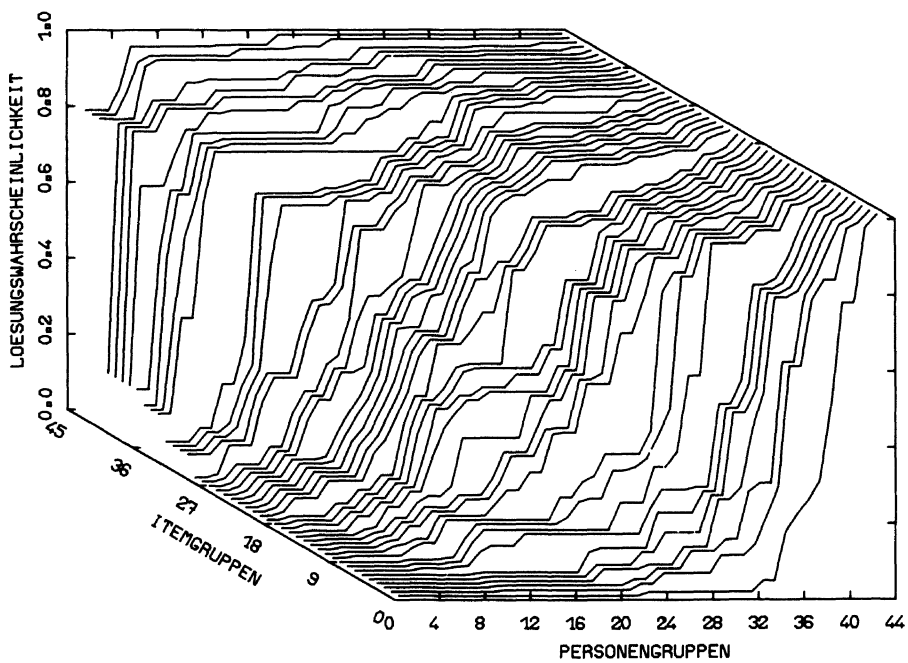

Abb. 1. Modellkonform geschätzte Lösungswahrscheinlichkeiten für 43 Personen- und 44 Item-Rohwertgruppen, dargestellt als Itemcharakteristiken.

wahrscheinlichkeiten bietet damit die Möglichkeit, die Differenzierungsfähigkeit einzelner Testaufgaben oder auch des gesamten Tests zu beurteilen. Dazu ist es hilfreich, für jedes Item eine Funktion zu definieren, die die Diskriminationsfähigkeit des Items für die verschiedenen Rohwertgruppen beschreibt. In den logistischen Testmodellen leistet dies die Informationsfunktion (Fischer, 1974), eine Funktion, die von der Ableitung der (logarithmierten) ICC abhängt. Der Informationsbeitrag eines Items ist dort groß, wo die ICC steil ansteigt. Dies ist durchaus plausibel, denn hat ein Item für zwei Rohwertgruppen die gleiche Lösungswahrscheinlichkeit, so kann es keine Information zur Trennung dieser Gruppen liefern. Unterscheiden sich die beiden Lösungswahrscheinlichkeiten jedoch, dann wird die Beantwortung dieses Items die Likelihood des Probanden, einer der beiden Rohwertgruppen anzugehören, unterscheidbar machen (vgl. Abschnitt 7, Gl. 7). Zur Beschreibung der Differenzierungsfähigkeit eines Items wird deshalb eine Diskriminationsfunktion definiert, die für jedes Item die Veränderung der ICC angibt: Für alle Rohwertgruppen g mit $u(g) > \min\{u(g) \mid g \in \overline{A}\}$ wird definiert

$$D(g \mid i) = p(g, i) - p(g^*, i),$$

wobei g* die zu g nächstniedrige Personengruppe bezeichnet.

Die Diskriminationsfunktion $D(g \mid i)$ eines Items i gibt also an, wie sich die Lösungswahrscheinlichkeit einer Rohwertgruppe g von der der nächstniedrigeren Rohwertgruppe unterscheidet. Abbildung 2 zeigt die Diskriminationsfunktionen zweier Testaufgaben, deren Maxima in ganz unterschiedlichen Bereichen der Personenfähigkeit liegen. Daraus geht hervor, daß zum Beispiel die Aufgabe 2 die Rohwertgruppen 2 bis 22 nicht unterscheiden kann. Diese Einschränkung führt dazu, daß in Fällen, in denen bereits Informationen über die Fähigkeit eines Probanden vorliegen, diese Aufgabe im Test möglicherweise überhaupt keine weitere Information liefern kann.

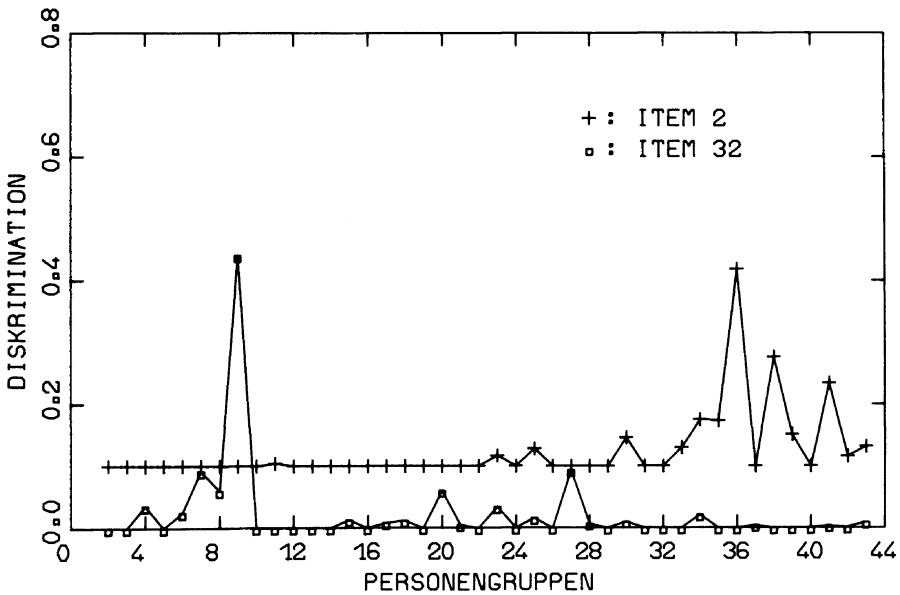

Abb. 2. Diskriminationswerte für zwei Itemgruppen, deren Maxima in unterschiedlichen Fähigkeitsbereichen liegen. Zur besseren Übersicht wurde bei Item 2 der Betrag 0.1 hinzuaddiert.

Es kann daher vorteilhaft sein, für einen Probanden individuell die am besten geeigneten Items auszuwählen, falls man aus Zeit- oder Kostengründen nicht den Gesamttest anwenden will.

7. Individualisiertes Testen mit dem U-Modell

Die individualisierte Testvorgabe stellt drei wesentliche Probleme (vgl. Lord, 1974):

1. Wie sollen die vorzulegenden Aufgaben aus der Menge aller vorhandenen Aufgaben ausgewählt werden?
2. Wie kann der Personenparameter bestimmt werden?
3. Wann wird die Testvorgabe abgebrochen?

Wir wollen hier keine vollständige Lösung dieser Probleme mit dem U-Modell darstellen, es soll jedoch gezeigt werden, daß Lösungen möglich sind. Da die Funktion p(g, i) für konstantes i die ICC des Items i darstellt (vgl. Abb. 1), können die von Lord (1974) vorgeschlagenen Methoden zur Lösung der oben genannten Probleme einer individualisierten Testvorgabe nahezu unverändert benützt werden. Die Funktionen p, u und v werden als bekannt vorausgesetzt. Soll nun der Parameter u(g) einer Person g mit einer individuellen Aufgabenauswahl bestimmt werden, dann werden dem Probanden zu Anfang Aufgaben unterschiedlicher Schwierigkeit vorgegeben bis Antwortwechsel auftreten, d. h. bis mindestens zwei verschiedene Ausprägungen von X aufgetreten sind. Dann kann durch Maximierung der Likelihood

$$L\,i(X \mid u(g), v) = \prod_{i \in B'} p(g,i)^{X(g,i)} [1 - p(g,i)]^{[1 - X(g,i)]}, \tag{7}$$

wobei $B' \subseteq B$ die Menge der bisher vorgelegten Aufgaben ist, der Parameter u(g) geschätzt werden. Die Likelihoodfunktion wird im allgemeinen kein eindeutiges Maximum besitzen, so daß für u(g) nur ein Intervall angegeben werden kann. Die nächste vorzulegende Aufgabe j wird so gewählt, daß j aus $B - B'$ zwischen den Fähigkeitswerten im geschätzten u(g)-Bereich maximal diskriminiert. Als Abbruchkriterium des Verfahrens kann neben den üblichen Kriterien, wie der Anzahl der bearbeiteten Aufgaben oder der Zahl der Antwortwechsel (vgl. Lord, 1974), zusätzlich die Größe des geschätzten Intervalls für u(g) benützt werden. Eine genaue Schätzung des Personenparameters wird jedoch nur dann möglich sein, wenn in allen Fähigkeitsbereichen viele Testaufgaben mit hoher Diskrimination zur Verfügung stehen.

Verwendet man das Rasch-Modell zum individualisierten Testen (Weichselgartner, 1981; Drösler & Lohner, 1981), so versucht man dem Probanden g ein Item i vorzulegen, welches seiner Fähigkeit am besten entspricht, d. h. $u(g) \approx v(i)$ (vgl. Lord, 1974). Dabei werden Aufgaben und Personen auf einer gemeinsamen Skala gemessen. Das U-Modell, das Personen und Aufgaben auf zwei verschiedenen Skalen mißt, zeigt jedoch, daß eine gemeinsame Skala für Personen und Aufgaben keine notwendige Voraussetzung für ein individualisiertes Testverfahren ist. Der eigentliche Grundsatz der individualisierten Testvorgabe besteht darin, dem Probanden dasjenige Item anzubieten, das zur Schätzung seiner Fähigkeit die maximale Information liefert. Je größer die Diskri-

mination eines Items ist, desto größer ist auch die Information, die durch dieses Item über einen bestimmten Probanden gewonnen werden kann (Fischer, 1974, S. 321).

Durch die Kenntnis der Schätzwerte p(g, i) ist es im U-Modell sehr einfach, die im Verlauf des Testvorgangs erhobene Information optimal auszunützen. Dies geschieht durch Maximierung der Likelihoodfunktion (7). Dabei wird für die vorgelegten Testaufgaben i, j, ... genau die Rohwertgruppe g' gesucht, deren Lösungswahrscheinlichkeiten p(g', i), p(g', j), ... die Likelihood des gefundenen Datenvektors X maximieren. Das U-Modell bietet damit für die individualisierte Testvorgabe ähnliche Möglichkeiten wie das Rasch-Modell. Der einzige wesentliche Unterschied (über das Skalenniveau hinaus) besteht darin, daß beim Rasch-Modell eine differenzierbare Modellgleichung vorliegt und dadurch eine Approximation der Varianz der Personen- und Itemparameterschätzungen möglich ist (Fischer, 1974). Hierfür kann zumindest vorerst im U-Modell noch keine befriedigende Lösung angeboten werden.

8. Diskussion

Bei der Darstellung des U-Modells wurde ein meßtheoretisches Konzept verfolgt. Dies hat vor allem den Vorteil, daß es eine klare Trennung zwischen dem Modell selbst und den bei einer Anwendung auftretenden statistischen Problemen ermöglicht. In der psychologischen Diagnostik entstehen besondere statistische Schätzprobleme dadurch, daß es einerseits notwendig ist, von probabilistischen Systemen im Sinne der Def. 1 auszugehen, aber andererseits nicht möglich ist, individuelle Lösungswahrscheinlichkeiten über relative Häufigkeiten zu schätzen. Mit dem U-Modell wurde versucht, die schwächsten hinreichenden Bedingungen zu formulieren, die ein probabilistisches Testsystem erfüllen muß, so daß gleichzeitig die Existenz von Ordinalskalen für Personen und Aufgaben und die Möglichkeit, die Skalenwerte aus den Randsummen der Datenmatrix zu schätzen, gesichert sind.

Diese beiden Forderungen, die für eine praktische Anwendung als unverzichtbar gelten können, führen damit zu bedeutsamen Restriktionen an die Daten. Die Itemcharakteristiken des U-Modells müssen zwar nicht parallel sein, wie im Rasch-Modell, sie dürfen sich jedoch nicht überschneiden, wie dies zum Beispiel im 3-parametrigen Birnbaum-Modell (Birnbaum, 1968) möglich ist. Die meßtheoretische Analyse des U-Modells demonstriert damit, daß bereits eine als Ordinalskala interpretierte Rohwertsumme statistisch prüfbare Annahmen voraussetzt.

In allen Situationen, in denen ein Psychologe ausschließlich daran interessiert ist, Probanden entsprechend ihrer Leistung zu ordnen, ist das U-Modell eine brauchbare theoretische Grundlage. Bezüglich dieser Problemstellung liefert es für den Psychologen die gleiche Information über die getesteten Personen wie das Rasch-Modell. Gleichzeitig zeigt das U-Modell jedoch eine bessere Anpassung an die Daten. Das Problem der Parameterschätzung und des Modelltests sind für das U-Modell und das Rasch-Modell ähnlich gut gelöst. Da das U-Modell, wie wir gezeigt haben, auch zum individualisierten Testen verwendet werden kann, ist es als ernsthafte Alternative zum Rasch-Modell zu betrachten.

Sowohl das Birnbaum- als auch das U-Modell stellen eine Verallgemeinerung des Rasch-Modells dar. Sie unterscheiden sich jedoch in ihren empirischen Restriktionen,

d. h. es sind Datensätze möglich, welche das U-Modell erfüllen, nicht jedoch das Modell von Birnbaum und umgekehrt. Im Birnbaum-Modell werden die empirischen Anforderungen des Rasch-Modells durch die Einführung eines dritten Parameters gelokkert. Im Gegensatz dazu werden im 2-parametrigen U-Modell die Restriktionen des Rasch-Modells dadurch abgeschwächt, daß an Stelle einer Verhältnisskala eine Ordinalskala konstruiert wird.

Somit ist das U-Modell auch als Alternative zum Birnbaum-Modell zu verstehen. Im Vergleich zum Birnbaum-Modell besitzt das U-Modell vor allem den Vorteil, daß Personen und Aufgaben unabhängig sind, was die Interpretation der Skalen erheblich erleichtert. Ein weiterer Vorteil des U-Modells liegt in dem einfachen und theoretisch zufriedenstellenden Schätzverfahren. Dagegen existiert für das Birnbaum-Modell kein Algorithmus der für einen beliebigen Datensatz eine korrekte Lösung garantiert (vgl. Wood, Wingersky & Lord, 1976). Die Arbeit von Färber & Zimmer (1980) hat dies deutlich gezeigt. Diese Autoren erhielten für das Birnbaum-Modell nur dann konvergierende Schätzwerte, wenn spezifische, durch das Rasch-Modell bestimmte Anfangswerte in das iterative Schätzprogramm von White (vgl. Färber & Zimmer, 1980) eingegeben wurden. Da das resultierende Schätzergebnis nicht durch mehrere verschiedene Anfangswerte erzeugt werden konnte, kann man nicht ausschließen, daß das berechnete Schätzergebnis nur ein lokales Maximum darstellt. Dagegen vermeiden MILSA und AMILSA die Probleme, die mit iterativen Schätzverfahren verbunden sind, indem sie die Schätzwerte algebraisch berechnen.

Besonders interessant wäre es, das U-Modell auf Datensätze anzuwenden, bei denen sich für das Rasch-Modell signifikante Modellabweichungen ergeben haben. Dadurch kann überprüft werden, ob die Modellabweichungen durch die Restriktionen der speziellen logistischen ICC's zustandegekommen sind, oder ob diese Abweichungen grundliegender sind. Eine Verletzung des U-Modells bedeutet im wesentlichen einen Verstoß gegen die Eindimensionalität der Personen- bzw. Itemskala. Es gibt dann zum Beispiel keine Möglichkeit, die Personen so zu ordnen, daß die Ordnung ihre Fähigkeit, ein beliebiges Testitem zu lösen, beschreibt. Verletzungen des Rasch-Modells bedeuten nicht notwendig Verstöße gegen die Eindimensionalität der Skalen. Sie können auch auf die spezielle Form der ICC zurückgehen. Das U-Modell bietet hier die Möglichkeit, die Skalierbarkeit eines Merkmals zu prüfen, ohne Annahmen über die spezielle Form der ICC's zu machen.

Literatur

Abelson, P. R. & Tukey, J. W., Efficient utilization of nonnumerical information in quantitative analysis: general theory and the case of simple order. Ann. Math. Statist., 1963, *34,* 1347–1369.

Anderson, E. B., A goodness of fit test for the Rasch model. Psychometrika, 1973, *38,* 123–140.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D., Statistical inference under order restrictions: the theory and application of isotonic regression. New York: Wiley & Sons, 1972.

Birnbaum, A., Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Addison-Wesley, Reading/Mass. 1968.

Doignon, J. P. & Falmagne, J. C., Difference measurement and simple scalability with restricted solvability. Journ. Math. Psych., 1974, *11,* 473–800.

Drösler, J. & Lohner, M., Verschränktes „Tailored Testing" zur Bestimmung der Kennwerte eines Tests während des praktischen Untersuchungsbetriebes. Zeitsch. exp. angew. Psychol., 1981, *28*, 80–99.
Färber, B. & Zimmer, A., Statistischer und entscheidungsorientierter Vergleich der Testmodelle von Rasch und Birnbaum. Diagnostica, 1980, *26*, 10–20.
Falmagne, J. C. & Iverson, G., Conjoint Weber laws and additivity. Journ. Math. Psych., 1979, *20*, 164–183.
Fischer, G., Einführung in die Theorie psychologischer Tests. Bern: Huber, 1974.
Fishburn, P. C., Binary choice probabilities: on the varieties of stochastic transitivity. Journ. Math. Psych., 1973, *10*, 327–352.
Fricke, R., Über Meßmodelle in der Schulleistungsdiagnostik. Düsseldorf: Schann, 1972.
Hadley, G., Non Linear and dynamic programming. London: Addison-Wesley, 1964.
Hamerle, A. & Tutz, G., Goodness of fit tests for probabilistic measurement models. Journ. Math. Psych., 1980, *21*, 153–167.
Hehl, F. J. & Hehl, R., Persönlichkeitsskalensystem 25: Manual. Weilheim: Beltz, 1975.
Herbst, K., Ermittlung und Bewertung von Verstößen gegen den Grundsatz der spezifischen Objektivität in psychodiagnostischen Untersuchungen. Dissertation, Regensburg, 1978.
Holman, E. W., Monotonic models for asymmetric proximities. Journ. Math. Psych., 1979, *20*, 1–15.
Krantz, D. H., Rational distance functions for multidimensional scaling. Journ. Math. Psych., 1967, *4*, 226–245.
Kratz, D. H., Luce, R. D., Suppes, P. & Tversky, A., Foundations of measurement, Vol. I. New York: Academic Press, 1971.
Lehner, P. E. & Noma, E., A new solution to the problem of finding all numerical solutions to ordered metric structures. Psychometrika, 1980, *45*, 135–137.
Lord, F. M., Individualized testing and item characteristic curve theory. In D. H. Kratz, R. C. Atkinson, R. D. Luce & P. Suppes (Eds.), Contemporary developments in mathematical psychology, Vol. 2. San Francisco: Freeman, 1974.
Mokken, R. J., A theory and procedure of scale analysis. Paris: Mouton, 1971.
Rasch, G., Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute of Educational Research, 1960.
Rasch, G., On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), Proceedings of the fourth Berkeley symposium on mathematical statistics and probability. Vol. 4, Berkeley: University of California Press, 1961. S. 321–333.
McClelland, G. H. & Coombs, C. H., ORDMET: A general algorithm for constructing all numerical solutions to ordered metric structures. Psychometrika, 1975, *40*, 269–290.
Schaafsma, W., Hypothesis testing problems with the alternative restricted by a number of inequalities. Groningen: Noordhoff, 1966.
Schaafsma, W. & Smid, L. J., Most stringent somewhere most powerful tests against alternatives restricted by a number of inequalities. Ann. Math. Statist., 1966, *37*, 1161–1172.
Spada, H., Fischer, G., & Heyner, W., Denkoperationen und Lernprozesse bei Lösung von Problemstellungen aus der Mechanik. In L. H. Eckensberger (Hrsg.), Bericht über den 28. Kongreß der Deutschen Gesellschaft für Psychologie. Göttingen: Hogrefe, 1974.
Suppes, P. & Zinnes, J., Basis measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), Handbook of mathematical psychology, Vol. I. New York: Wiley, 1963.
Tversky, A. & Russo, J. E., Substitutability and similarity in binary choices. Journ. Math. Psychol., 1969, *6*, 1–12.
Weichselgartner, E., Adaptive individualisierte Psychodiagnostik auf Taschenrechner-Basis. Zeitsch. exp. angew. Psychol., 1981, *28*, 335–352.
Wood, R. L., Wingersky, M. S., & Lord, F. M., LOGIST. A computer program for estimating examinee ability and item characteristic curve parameters. Research memorandum. Educational Testing Service, Princeton New Jersey, 1976.

Liste besonderer Symbole

$\wedge$	Konjunktion
$\rightarrow$	Implikation
$\leftrightarrow$	Äquivalenz
$\forall$	Allquantor
$\exists$	Existenzquantor
$<$	Relation „kleiner“
$\leq$	Relation „kleiner oder gleich“
$\subseteq$	Teilmenge
$\cap$	Mengendurchschnitt
$\in$	Element aus
x	kartesisches Mengenprodukt
$E\{\ \}$	Erwartungswertoperator
$\mathbb{P}\{\ \}$	Wahrscheinlichkeitsmaß
$\mathbb{R}$	Menge der reellen Zahlen
α	Alpha
Π	Produktzeichen
Σ	Summenzeichen
φ	Phi
ψ	Psi