Zuerst ersch. in: Psychologie. Eine Einführung. Grundlagen, Methoden, Forschungsfelder / Jürgen Straub, Wilhelm Kempf ... (Hrsg.). München: dtv, 1997, S. 559-575

Testtheorie

WILHELM KEMPF

Psychologische Tests

Der Gedanke, daß man psychische Eigenschaften messen könne, ist sehr alt. Bereits 1883 veröffentlichte Galton ein Buch, in dem er eine Reihe von Fragen und Aufgaben vorschlug, mit deren Hilfe man den Grad der geistigen Begabung messen könne (vgl. hierzu und zum Folgenden auch Rosts Beitrag über Intelligenz). 1894 entwickelte Binet dann den ersten Intelligenztest für Kinder. Dieser Test wurde zehn Jahre später von Binet und Simon stark verbessert, indem sie eine Normierung der Aufgaben durchführten. Für jedes Lebensjahr wurde eine Reihe von Aufgaben (»Items«) zusammengestellt und an einer sehr großen Stichprobe von Kindern überprüft. Wenn 75 Prozent aller Kinder einer Altersgruppe (aber nur 25 Prozent der Kinder der nächstjüngeren Altersgruppe) die Aufgaben lösen konnten, so galten sie als »geeicht« (das heißt der Altersstufe in ihrem Schwierigkeitsgrad entsprechend).

Für jedes Lebensjahr wurde eine Serie von fünf Items zusammengestellt. Beispielsweise waren es für die Zehnjährigen folgende:

- fünf Gewichte vom leichtesten (6 g) bis zum schwersten (18 g) ordnen;
- die Zeichnung einer komplizierten geometrischen Figur aus dem Gedächtnis wiedergeben;
- mit drei vorgegebenen Worten (z. B. Schnee, spielen, Schlitten) zwei Sätze bilden;
- absurde Sätze kritisieren (z. B. »Ich habe drei Brüder – Paul, Ernst und ich«)
- sowie fünf praktische Verstandesfragen (etwa »Was muß man machen, wenn man in die Schule geht und man unterwegs merkt, daß es schon später ist als gewöhnlich?«).

Aus den Antworten der Kinder wurde ihr sogenanntes Intelligenzalter (IA) errechnet. Ausgehend von dem Jahr, bis zu dem ein Kind alle Aufgaben lösen konnte, wurde für jede gelöste Aufgabe ein

Fünftel Jahr dazugezählt. Das Intelligenzalter wurde dann zum Lebensalter (LA) in Beziehung gesetzt, zum Beispiel:

$$IA - LA = 10.4 - 10.0 = 0.4 \text{ Jahre »Intelligenz-Vorsprung«}$$

War das Intelligenzalter kleiner als das Lebensalter, so sprach man von einem »Intelligenz-Rückstand«.

William Stern (zit. n. Rohracher 1971, S. 383) kritisierte das Verfahren, da die Differenz zwischen Intelligenzalter und Lebensalter »keinen eindeutigen Sinn hat, sondern ganz Verschiedenes bedeutet, je nach der Altersstufe des geprüften Kindes.« Und weiter schreibt er: »Ein Rückstand von zwei Jahren ist beim sechsjährigen Kind ein Zeichen starker geistiger Minderwertigkeit, er verrät beim neunjährigen einen sehr viel geringeren Schwachsinnsgrad; und für das zwölfjährige Kind braucht die durch ihn ausgedrückte Schwäche überhaupt noch nicht jenseits der Grenze der Normalität zu liegen. Ein bestimmter Differenzwert hat also um so geringere Bedeutung, je höher das Alter des Kindes ist. Oder anders ausgedrückt: Bei gleichem Intelligenzgrad muß die Differenz wachsen, wenn das Lebensalter wächst.« Als Maß für die Intelligenz schlug Stern daher den sogenannten Intelligenzquotienten vor: IQ = IA/LA. Später hat es sich eingebürgert, den Intelligenzquotienten mit 100 zu multiplizieren: IQ = 100 x IA/LA.

Als man dann begann, auch Intelligenztests für Erwachsene zu entwickeln, mußte man das Vorgehen etwas modifizieren. Im Erwachsenenalter bleibt die intellektuelle Leistungsfähigkeit über einen großen Zeitraum hin konstant. Alterstypische Aufgaben sind daher nicht mehr konstruierbar. Statt dessen ging man zur Auswahl von Aufgaben über, die in ihrer Schwierigkeit hinreichend stark variieren, so daß leichte Aufgaben von vielen, schwierige Aufgaben dagegen nur von wenigen Versuchspersonen (Vpn) gelöst werden. Für jede gelöste Aufgabe erhält die Vp einen Punkt (»dichotome Testitems«). Mitunter werden die Antworten auf die Testitems auch auf einer mehrstufigen Skala bewertet (»polytome Testitems«); zum Beispiel 0 Punkte = falsch; 1 Punkt = teilweise richtig; 2 Punkte = vollständig richtig. Aus Gründen der mathematischen Einfachheit werden wir uns in der vorliegenden Einführung jedoch auf dichotome Items beschränken. Die Testleistung der Vp wird schließlich in ihrem »Testscore« ausgedrückt: in der Gesamtzahl der im Test erreichten Punkte.

Der große Erfolg der Intelligenztests hat dazu geführt, daß man

bald schon begonnen hat, auch andere psychologische Eigenschaften mittels Tests messen zu wollen, zum Beispiel Einstellungen oder Persönlichkeitseigenschaften. Auch diese Tests sind im Prinzip genauso aufgebaut. Die Items sind dort aber meist keine Aufgaben, die es zu lösen gilt, sondern Fragen, welche die Vpn beantworten. In der Klinischen Psychologie gibt es zudem solche Tests (sogenannte Symptomchecklisten), die nicht vom Patienten selbst ausgefüllt werden, sondern der behandelnde Therapeut kreuzt an, welche Symptome ein Patient zeigt. Die Items sind hier also Symptome, die auftreten können oder nicht.

Wir können somit festhalten: Ein psychologischer Test ist eine standardisierte Verhaltensstichprobe. Er besteht aus einer Anzahl (k) von Items (i=1,...,k). Die Antwort (x_{vi}) einer Vp v auf Item i wird auf einer zwei- oder mehrstufigen Skala beurteilt. Bei dichotomen Items ist $x_{vi} = 0$ oder 1. Aus den Itemantworten wird der Testscore der Vp (x_{vt}) errechnet, indem man die bei den einzelnen Items erzielten Punktzahlen aufsummiert $x_{vt} = \sum x_{vi}$.

Fragestellungen und Grundannahmen der psychologischen Testtheorie

Gegenstand der psychologischen Testtheorie ist einerseits die Konstruktion psychologischer Tests und andererseits die kritische Beurteilung psychologischer Testergebnisse. Themen der psychologischen Testtheorie sind die folgenden Eigenschaften psychologischer Tests: Homogenität, Objektivität, Reliabilität und Validität.

Die Frage nach der *Homogenität* eines Tests betrifft die Zulässigkeit der Scorebildung: Messen die Items eines Tests tatsächlich alle dieselbe Eigenschaft der Vpn, so daß es nur darauf ankommt, wieviele Items eine Person positiv beantwortet? Oder kommt es darauf an, welche Items die Person positiv beantwortet, so daß die verschiedenen Items eine eigene diagnostische Relevanz besitzen?

Die Frage nach der *Objektivität* eines Tests betrifft die Unabhängigkeit der Testergebnisse vom Testleiter: Kommen verschiedene Psychologen mit demselben Test tatsächlich zu denselben diagnostischen Aussagen? Und wie kann man das sicherstellen?

Die Frage nach der *Reliabilität* eines Tests betrifft die Meßgenauigkeit des Tests: Man muß davon ausgehen, daß die Antworten der

Vpn zufällig variieren. Mal hat eine Vp Pech und gibt eine falsche Antwort, obwohl sie die richtige Antwort doch eigentlich weiß; mal hat sie Glück und gibt die richtige Antwort, obwohl sie die Aufgabe eigentlich doch nicht so ganz beherrscht. Wie stark müssen sich die Testergebnisse zweier Vpn dann voneinander unterscheiden, daß man wirklich mit einiger Sicherheit schließen kann, daß zum Beispiel die beiden Personen tatsächlich unterschiedlich begabt sind?

Die Frage nach der *Validität* eines Tests betrifft die Gültigkeit des Tests für die jeweilige Fragestellung: Mißt zum Beispiel ein Intelligenztest tatsächlich Begabungsunterschiede der Vpn oder nur unterschiedliche Bildungschancen? Mißt er nur die Anpassung an mittelständische Bildungsnormen?

Mit der Homogenität von Tests beschäftigt sich die von Lazarsfeld (1950) begründete »*Stochastische Testtheorie*«. Objektivität, Reliabilität und Validität werden in der sogenannten »*Klassischen Testtheorie*« behandelt, die erstmals von Gulliksen (1950) methodisch-systematisch dargestellt wurde. Beiden Theorien gemeinsam ist die Annahme, daß die Itemantworten und folglich auch die daraus errechneten Testscores der Vpn keine feststehenden Größen sind, sondern zufälligen Schwankungen unterliegen. Die empirisch beobachteten Itemantworten x_{vi} und Scores x_{vt} sind also Realisationen zufälliger Variablen X_{vi} beziehungsweise X_{vt}.

Weiter wird angenommen, daß die Antworten, welche ein und dieselbe Vp auf verschiedene Items gibt, und entsprechend auch die Scores, welche sie in verschiedenen Tests erhält, voneinander unabhängig verteilt sind. Diese Annahme bezeichnet man als »lokale stochastische Unabhängigkeit«. Für die Testkonstruktion hat sie zur Folge, daß die Testitems logisch unabhängig voneinander sein müssen. Es darf also zum Beispiel nicht vorkommen, daß die Lösung eines Items als Teilschritt in der Lösung eines anderen Items enthalten ist.

Von »lokaler« stochastischer Unabhängigkeit spricht man deswegen, weil diese Unabhängigkeitsannahme immer nur für eine feste Vp gilt, also die statistische Unabhängigkeit der Antwort- und Scorevariablen X_{vi} beziehungsweise X_{vt} betrifft. In jeder Stichprobe oder Population von Vpn werden die Antworten beziehungsweise Scores der Vpn dagegen miteinander korrelieren. So werden hochintelligente Vpn bei allen Items eines Intelligenztests relativ hohe Erfolgswahrscheinlichkeiten (p_{vi}) haben und in verschiedenen Intelligenztests relativ hohe Scores (x_{vt}) erzielen, während minderbegabte Vpn bei allen Items relativ geringe Erfolgswahrscheinlichkei-

ten aufweisen und in beiden Tests relativ niedrige Scores erzielen werden. Betrachtet man die Variation der Itemantworten beziehungsweise Scores innerhalb einer Personenpopulation, so werden die Antwortvariablen ($X_{\cdot i}$) beziehungsweise Maßzahlvariablen ($X_{\cdot t}$) in dieser Population um so höher miteinander korrelieren, je stärker sich die Vpn in ihrer Intelligenz voneinander unterscheiden.

Während die stochastische Testtheorie von den einzelnen Itemantworten ausgeht und die Erfolgswahrscheinlichkeiten (p_{vi}) wahrscheinlichkeitstheoretisch modelliert, stehen in der klassischen Testtheorie die Scorevariablen (X_{vt}) im Zentrum der Theorienbildung. Könnte man ein und denselben Test beliebig oft wiederholen, so würde dieselbe Vp mal einen etwas höheren, mal einen etwas geringeren Score erzielen. Die mittlere Testleistung, welche die Vp in einer unendlich langen Serie von Testwiederholungen erzielen würde, bezeichnet man dann als »wahre« Testleistung (True Score) der Vp und schreibt dafür den griechischen Buchstaben τ (tau).

In der Praxis ist die beliebig häufige Wiederholung eines Tests freilich nicht möglich, da man mit Lern- oder Gedächtniseffekten rechnen muß. Auch wenn dies nicht der Fall wäre, könnte man den Test nicht unendlich oft wiederholen, um den True Score empirisch zu bestimmen. Denn, wie schon das Wort »unendlich« besagt, würde man dabei nie zu einem Ende kommen. Der True Score ist also keine empirisch bestimmbare, sondern eine theoretische Größe, die durch das Testergebnis der Vp jedoch statistisch »geschätzt« werden kann. Die Abweichung des als Schätzwert verwendeten Testergebnisses (x_{vt}) vom True Score der Vp bezeichnet man als den Meßfehler (f_{vt}).

Während das Testergebnis einer Vp die Realisation einer zufälligen Variablen ist, ist der True Score der Vp eine Konstante, welche diese Vp charakterisiert. Gehen wir von der Betrachtung einer einzelnen Vp zur Betrachtung einer Personenpopulation über, so werden jedoch nicht nur die Testergebnisse über die Vpn hinweg variieren, sondern verschiedene Vpn werden sich auch in ihren True Scores voneinander unterscheiden, so daß wir von einer True-Score-Variablen ($T_{\cdot t}$) sprechen können. Dasselbe gilt für die Meßfehler, die wir als Realisationen einer Fehlervariablen ($F_{\cdot t}$) auffassen können. Als Grundlage der klassischen Testtheorie ergibt sich somit die Modellgleichung

$$x_{\cdot t} = t_{\cdot t} + f_{\cdot t} \qquad (1)$$

wonach die Maßzahlvariable als Summe aus der True-Score-Variablen und einer Fehlervariablen besteht. Auf der Grundlage der getroffenen Annahmen und Definitionen läßt sich mathematisch beweisen, daß für diese Variablen die folgenden statistischen Beziehungen gelten, die in der Fachliteratur als die *Axiome der klassischen Testtheorie* bekannt sind:

1. Der mittlere Meßfehler ist gleich Null: $E(F_t) = 0$.
2. True Score und Meßfehler sind nicht miteinander korreliert: $korr(T_t,F_t) = 0$.
3. Die Meßfehler aus verschiedenen Tests t und t' sind nicht miteinander korreliert: $korr(F_t,F_t') = 0$.
4. Der Meßfehler in einem Test korreliert nicht mit dem True Score in einem anderen Test: $korr(F_t,T_t') = 0$.

Homogenität

Während es bei Intelligenztests (zunächst) plausibel erscheint, daß man die Testleistung einer Vp in einem Score zusammenfassen kann, ist dies (spätestens) bei der Einstellungsmessung oder in der klinischen Diagnostik nicht der Fall. Betrachten wir zum Beispiel die depressiven Symptome: (1) herabgesetzte Psychomotorik, (2) gehemmter Denkablauf und (3) ängstlich-gequält-klagsame Stimmung, so ergibt sich die Frage, ob sich die Patienten bezüglich des Auftretens dieser Symptome nur quantitativ – im Sinne einer stärkeren oder schwächeren Symptomatik – voneinander unterscheiden oder ob es qualitative Unterschiede zwischen den Vpn gibt – das heißt durch typische Symptomkombinationen charakterisierte Syndrome.

Wenn die Symptome einen homogenen Test für die Stärke der Depressivität bilden, dann kommt es nur darauf an, wieviele Symptome auftreten. Wenn die Symptome dagegen verschiedene Erscheinungsformen der Depression widerspiegeln, dann kommt es darauf an, welche Symptome auftreten. Die Bildung eines Summenscores macht dann wenig Sinn. Man sagt, der Test sei heterogen. Welches der Fall ist, kann mit Hilfe der auf Lazarsfeld (1950) zurückgehenden *Latent-Class-Analyse* (LCA) empirisch untersucht werden.

Die LCA geht davon aus, daß sich die Vpn in eine Anzahl von

Gruppen (g=1,...,h) einteilen lassen, die sich quantitativ oder qualitativ voneinander unterscheiden. Welcher Gruppe beziehungsweise »Klasse« eine Vp angehört, ist zunächst nicht bekannt, sondern soll erst festgestellt werden. Man spricht daher von »latenten« Klassen (im Unterschied zu »manifesten« Klassen, bei denen die Vpn nach einem direkt beobachtbaren Merkmal klassifiziert werden, zum Beispiel nach dem Geschlecht).

Die verschiedenen Klassen zeichnen sich dadurch aus, daß jedes der untersuchten *Symptome* (i) in jeder *Klasse* (g) eine *für die Klasse charakteristische Auftrittswahrscheinlichkeit* (p_{gi}) hat. Die *Klassengröße* (p_g) beschreibt den relativen Anteil der Vpn, welche einer bestimmten Klasse (g) angehören. Die *Wahrscheinlichkeit* (p_{vi}), mit welcher das Symptom i bei einer *zufällig herausgegriffenen Vp v* auftritt, kann dann durch die folgende Modellgleichung beschrieben werden:

$$p_{vi} = \sum p_g p_{gi} \qquad (2)$$

Mit Hilfe entsprechender Computerprogramme (z.B. WINMIRA) kann man dann ausrechnen, wieviele Klassen zur Beschreibung der empirisch beobachteten Antworten von n Vpn auf k Items benötigt werden. Man erhält so die *Klassenanzahl* (h). Weiterhin erhält man die Klassengrößen (p_g) und die *klassenspezifischen Auftrittswahrscheinlichkeiten einer positiven Antwort auf die verschiedenen Items* (p_{gi}). Diese sind in einer Tabelle der folgenden Art darstellbar:

g	p_g	p_{g1}	p_{g2}	p_{g3}	$E(X_{gt})$
1	0.06 = 6%	1.00	1.00	1.00	3.0
2	0.17 = 17%	0.90	0.93	0.05	1.88
3	0.17 = 17%	0.00	0.06	1.00	1.06
h=4	0.60 = 60%	0.14	0.10	0.05	0.29
Ges.	1.00 = 100%	0.30	0.29	0.27	

Tab. 1 Latent-Class-Analyse der Symptome: (1) herabgesetzte Psychomotorik, (2) gehemmter Denkablauf und (3) ängstlich-gequält-klagsame Stimmung.

Tabelle 1 zeigt, daß 4 latente Klassen identifiziert wurden, deren erste mit 6 Prozent der Patienten sehr klein ist, die Klassen 2 und 3 umfassen je 17 Prozent der Patienten, und Klasse 4 ist mit 60 Prozent der Patienten bei weitem am größten.

Ob sich die Klassen quantitativ oder qualitativ voneinander unterscheiden, können wir in zwei Schritten beurteilen. In einem ersten Schritt betrachten wir die *mittleren Summenscores innerhalb der Klassen* ($E(X_{gt})$). Aus diesen sehen wir, daß sich die Klassen tatsächlich in ihrem Summenscore, das heißt quantitativ unterscheiden. Klasse 1 ist eine symptomreiche Klasse. Sie enthält Patienten, die offensichtlich schwer depressiv sind. Die Patienten in Klasse 2 und 3 zeigen weniger Symptome, und Klasse 4 ist eine symptomarme Klasse. Sie enthält Patienten, die schon fast gesund sind.

Zur Feststellung der Homogenität des Tests reicht dies aber noch nicht aus. Wir fertigen daher in einem zweiten Schritt eine graphische Darstellung der *klassenspezifischen Symptomwahrscheinlichkeiten* an (s. u. *Abbildung 1*), wobei die Symptome (aufgrund ihrer Auftrittswahrscheinlichkeiten in der Gesamtstichprobe vom leichtesten zum schwierigsten geordnet) auf der Abszisse und die klassenspezifischen Symptomwahrscheinlichkeiten auf der Ordinate eines Koordinatensystems aufgetragen und miteinander verbunden werden, so daß man für jede Klasse eine Profillinie erhält, welche die für die Klasse typische Symptomatik beschreibt.

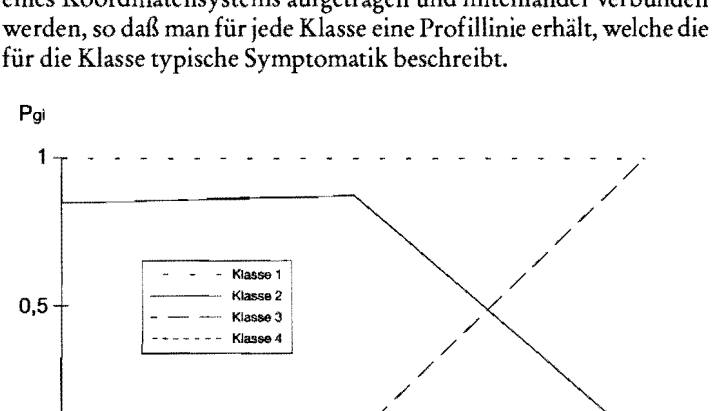

Abb. 1 Latent-Class-Analyse der Symptome: (1) herabgesetzte Psychomotorik, (2) gehemmter Denkablauf und (3) ängstlich-gequält-klagsame Stimmung.

Die Profillinie der symptomreichen Klasse 1 zeigt, daß hier alle 3 Symptome mit an Sicherheit grenzender Wahrscheinlichkeit auftreten. Die Profillinie der symptomarmen Klasse 4 zeigt, daß hier alle drei Symptome nur mit geringer Wahrscheinlichkeit auftreten, und zwar um so seltener, je schwieriger ein Symptom ist. Die Profillinie zeigt einen monoton fallenden Verlauf.

Soweit macht der Test den Eindruck von Homogenität. Betrachten wir jedoch die Profillinien der Klassen 2 und 3, so sehen wir, daß sich diese nicht nur quantitativ voneinander unterscheiden, sondern geradezu durch eine entgegengesetzte Symptomatik charakterisiert sind. Die Patienten in Klasse 2 zeigen mit hoher Wahrscheinlichkeit eine herabgesetzte Psychomotorik und einen gehemmten Denkablauf, kaum jedoch eine ängstlich-gequält-klagsame Stimmung. Bei den Patienten in Klasse 3 ist es gerade umgekehrt.

Der Test ist also heterogen. Mit den Klassen 2 und 3 identifiziert er zwei qualitativ verschiedene Syndrome der Depression. Die Patienten in diesen beiden Klassen unterscheiden sich zwar auch in der (mittleren) Anzahl der gezeigten Symptome voneinander, der Score allein ist aber nicht geeignet, um diese qualitativ verschiedene Symptomatik darzustellen.

Wenn ein Test nicht homogen ist, können wir die diagnostische Beurteilung der Vpn nicht (oder zumindest nicht ohne Informationsverlust) auf der Grundlage ihres Scores treffen. Notwendige Voraussetzung für Homogenität ist, daß die Profillinien alle monoton fallend sein müssen und sich nicht überschneiden dürfen. Je stärker diese Voraussetzung verletzt ist, desto größer ist der Informationsverlust, welcher mit der Scorebildung einhergeht.

Soll die gesamte Testinformation über die Vpn durch ihren Score erschöpfend beschrieben werden – so, daß es überhaupt nicht mehr darauf ankommt, welche Items eine Vp gelöst hat –, so sind noch strengere Voraussetzungen zu fordern. Wie Rasch (1960, 1968) gezeigt hat, ist dies genau dann der Fall, wenn sich die Wahrscheinlichkeitsverteilung der Itemantworten der Vpn durch die Modellgleichung

$$p_{vi} = \exp(\theta_v - \delta_i) / ((1 + \exp(\theta_v - \delta_i)) \qquad (3)$$

beschreiben läßt, so daß die Wahrscheinlichkeit einer positiven Antwort als Funktion zweier Parameter darstellbar ist, deren erster (θ_v; θ = griech. Buchstabe theta) die durch den Test erfaßte »latente« Eigenschaft der Vpn mißt. Je größer dieser Personenparameter ist,

desto größer ist die Wahrscheinlichkeit, mit welcher die Vp ein Item positiv beantwortet. Der andere Parameter (δ_i; δ = griech. Buchstabe delta) beschreibt die Itemschwierigkeit. Je größer dieser Itemparameter ist, desto geringer ist die Wahrscheinlichkeit, mit welcher das Item positiv beantwortet wird (vgl. *Abbildung 2*). Latente Personeneigenschaft und Itemschwierigkeit werden auf einer gemeinsamen Differenzenskala gemessen, das heißt, sie haben eine feste Maßeinheit, aber einen willkürlich festgesetzten Nullpunkt. Ist $\theta_v = \delta_i$, so ist $p_{vi} = {}^1/_2$.

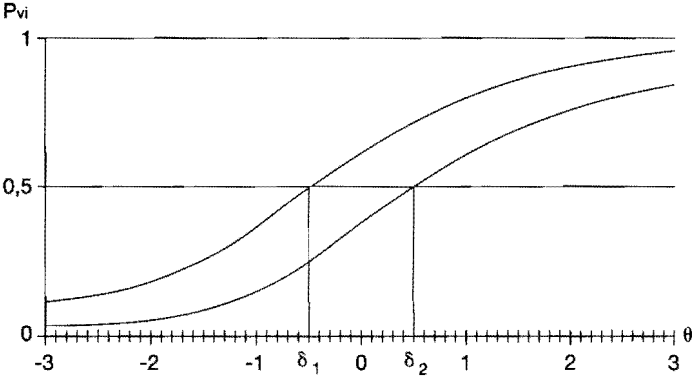

Abb. 2 Itemcharakteristik des Rasch-Modells.

Ob sich die empirischen Daten durch diese Modellgleichung beschreiben lassen, kann mit Hilfe statistischer Tests geprüft werden, für deren Berechnung verschiedene Computerprogramme zur Verfügung stehen.

Dabei ist zu berücksichtigen, daß Homogenität im Sinne des *Rasch-Modells* ein Idealerfordernis darstellt, das in der testpsychologischen Praxis nur selten erfüllt ist. Während es eine Vielzahl von Anwendungen des Rasch-Modells in der testdiagnostischen Grundlagenforschung gibt, vermögen nur die wenigsten der auf dem Markt befindlichen Tests diesem Erfordernis gerecht zu werden. Bei der großen Majorität der gebräuchlichen Tests geht die Scorebildung deshalb mit einem mehr oder minder erheblichen Informationsverlust einher. Als psychologischer Diagnostiker sollte man sich daher in der Regel nicht nur mit der Beurteilung der Scores zufriedengeben.

Objektivität

Unter der Objektivität eines Tests versteht man den Grad, in dem die Ergebnisse eines Tests unabhängig vom Testleiter sind. Ein Test wäre demnach vollkommen objektiv, wenn verschiedene Psychologen bei derselben Vp zum selben Ergebnis gelangten. Lienert (Lienert & Raatz 1994) spricht deshalb auch von »interpersoneller Übereinstimmung« der Untersucher und unterscheidet – je nachdem, in welcher Phase der Testdurchführung allfällige Nicht-Übereinstimmungen auftreten können – drei verschiedene Aspekte der Objektivität.

Die sogenannte *Durchführungsobjektivität* betrifft die mögliche Beeinflussung der Testergebnisse durch zufällige oder systematische Verhaltensvariationen des Testleiters während der Testdurchführung. Soll die Durchführungsobjektivität maximal hoch werden, dann müssen alle Vpn den Test unter denselben – beziehungsweise unter vergleichbaren – Bedingungen bearbeiten. Hierzu gehören die Standardisierung der Untersuchungssituation (allgemeine Arbeitsbedingungen, Beleuchtungs- und Sehverhältnisse, Arbeitsmaterial, Bearbeitungszeit etc.), sowie die Standardisierung der Testinstruktion, die zum Beispiel schriftlich vorgegeben wird.

Aber: Dasselbe ist nicht für alle das gleiche. Es kommt darauf an, daß alle Vpn ein gleich gutes Verständnis davon entwickeln, was im Test von ihnen erwartet wird. Manche können besser lesen, manche schlechter. Für manche ist die Instruktion »zu abstrakt« formuliert, für andere »zu primitiv«. Will man die Durchführungsobjektivität optimieren, so kommt man daher nicht umhin, die Instruktion flexibel zu gestalten, an die jeweiligen Vpn anzupassen und sich durch Nachfragen des Verständnisses zu versichern. Insbesondere bei der Testung von Kindern ist es unumgänglich, erst einen positiven Rapport herzustellen, bevor man irgendwelche Anforderungen an sie stellen kann. In eine griffige Formel gefaßt: »Never ask a child a question before it has smiled at you.«

Die sogenannte *Auswertungsobjektivität* betrifft die Bewertung der Itemantworten durch den Testleiter (z.B. als »richtig« oder »falsch«). Sieht der Test frei formulierte Antworten der Vpn vor (sogenannte »offene« Fragen), so ist hierfür ein präzises Regelsystem mit Beispielen und Gegenbeispielen vonnöten, das den Beurteilungsspielraum möglichst einengt.

Eine Optimierung der Auswertungsobjektivität kann durch so-

genannte »multiple-choice« Fragen erzielt werden, bei denen die Vp aus mehreren vorgegebenen Antwortmöglichkeiten diejenige auswählt, die ihr zutreffend erscheint. Dieser Vorteil wird jedoch um den Nachteil erkauft, daß die richtige Antwort bei multiple-choice-Fragen auch durch bloßes Raten zustandekommen kann.

Zur Kontrolle der Auswertungsobjektivität kann man einen Test durch zwei unabhängige Untersucher auswerten lassen. Zur Beurteilung der Auswerterübereinstimmung berechnet man dann zum Beispiel den Koeffizienten Kappa (Cohen 1960; vgl. Bortz, Lienert & Boehnke 1990).

Die sogenannte *Interpretationsobjektivität* betrifft den Grad der Unabhängigkeit der Interpretation der Testergebnisse von der Person des Diagnostikers. Um einen objektiven Bewertungsmaßstab für den Testscore einer Vp zu erzielen, wird dieser zur Verteilung der Testscores in der Referenzpopulation (z.B. in der Gruppe der Gleichaltrigen) in Beziehung gesetzt. Als Vergleichsnormen dienen dabei Prozentrangtabellen, Standardwerte oder IQ-Punkte.

Prozentrangtabellen geben an, wie groß der Anteil der Vpn in der Referenzpopulation ist, die den jeweiligen Score nicht überschreiten. Standardwerte und IQ-Punkte transformieren die Scoreverteilung in eine Normalverteilung mit dem Mittelwert 100 und einer Standardabweichung von 10 (Standardwerte) beziehungsweise 15 (IQ-Punkte). Da die Normalverteilung tabelliert ist, kann man daraus jederzeit auf den Prozentrang des jeweiligen Scores rückschließen. Die Normalverteilungstransformation bietet darüber hinaus jedoch den Vorteil, daß man anhand des (transformierten) Testwertes sofort einen groben Eindruck gewinnt, wie dieser Wert zu beurteilen ist. So liegen zwei Drittel aller Testwerte in dem Intervall von +/- einer Standardabweichung um den Mittelwert. Standardwerte zwischen 90 und 110 beziehungsweise IQ-Punkte zwischen 85 und 115 können daher unmittelbar als durchschnittliche Testleistung erkannt werden, und so weiter.

Zur Erzielung von Interpretationsobjektivität bedarf es aber nicht nur der Vereinheitlichung der Interpretation der Testscores, sondern auch der Vereinheitlichung der daraus abzuleitenden diagnostischen Urteile. In der klinischen Diagnostik bedient man sich hierzu des Diagnostischen und Statistischen Manuals Psychischer Störungen (DSM-III), das zum Beispiel den Schweregrad intellektueller Beeinträchtigung wie folgt definiert:

	IQ
leicht	50–55 bis etwa 70
mäßig	35–40 bis etwa 50–55
schwer	20–25 bis etwa 35–40
schwerst	unter 20–25

Tab. 2 Schweregrad intellektueller Beeinträchtigung nach DSM-III.

Geistige Behinderung ist laut DSM-III jedoch nicht nur durch deutlich unterdurchschnittliche intellektuelle Leistungsfähigkeit (IQ unter 70) definiert, sondern auch durch gleichzeitige Defizite beziehungsweise Beeinträchtigungen der sozialen Anpassungsfähigkeit der Vp, das heißt ihrer Leistungsfähigkeit bei der Erfüllung sozialer Normen, die ihr soziales Umfeld von Personen ihres Alters erwartet bezüglich Anpassung und Verantwortung, Kommunikation, alltagspraktischer Fertigkeiten sowie persönlicher Unabhängigkeit und Autonomie. Das Ausmaß der Beeinträchtigung auf diesem Gebiet der sozialen Anpassungsfähigkeit spielt für die Beurteilung des Schweregrades der geistigen Behinderung ebenfalls eine Rolle.

Reliabilität

Unter der Reliabilität eines Tests versteht man den Grad seiner Meßgenauigkeit. Der *Reliabilitätskoeffizient* ist definiert als das Quadrat der Korrelation zwischen True-Score und Maßzahlvariable und beschreibt den Anteil der True-Score-Varianz an der Maßzahlvarianz in der Referenzpopulation:

$$\text{korr}^2(X_t, T_t) = \text{var}(T_t)/\text{var}(X_t). \qquad (4)$$

Da der True-Score eine theoretische Größe ist, kann diese Korrelation jedoch nicht direkt ausgerechnet werden. Eine intuitiv einleuchtende Methode zur Reliabilitätsbestimmung besteht jedoch in der Testwiederholung. Dabei wird der Test denselben Vpn zweimal vorgegeben und die Korrelation zwischen den Testergebnissen aus den beiden Meßreihen berechnet. Voraussetzung dafür ist, daß das

gemessene Merkmal über den Zeitraum zwischen den beiden Testungen konstant bleibt und sich auch die Meßgenauigkeit des Tests durch die wiederholte Testvorgabe nicht verändert. Sind diese Voraussetzungen erfüllt, so spricht man von parallelen Testungen, und die so berechnete *Retest-Reliabilität* stimmt mit dem oben definierten Reliabilitätskoeffizienten numerisch überein.

Eine andere Methode zur Reliabilitätsbestimmung ist die sogenannte *Paralleltestmethode*. Dabei wird nicht derselbe Test zweimal vorgegeben, sondern zwei verschiedene Tests, die dasselbe Merkmal gleich gut messen. Ist diese Voraussetzung erfüllt, so spricht man bei gleicher Schwierigkeit der Tests von parallelen Tests. Unterscheiden sich die Tests in ihrer Schwierigkeit, so spricht man von essentiell parallelen Tests. Unterscheiden sich zwei Tests zwar in ihrer Meßgenauigkeit, messen aber dennoch dasselbe Merkmal der Vpn, so spricht man bei gleicher Schwierigkeit der Tests von tau-äquivalenten Tests, bei unterschiedlicher Schwierigkeit von essentiell tau-äquivalenten Tests.

Ist die Reliabilität eines Tests nicht zufriedenstellend, so kann sie durch Testverlängerung verbessert werden. Wird ein Test zum Beispiel durch Hinzufügen eines Paralleltests auf doppelte Testlänge verlängert, so nimmt die Reliabilität des Tests, wie in *Abbildung 3* gezeigt, zu:

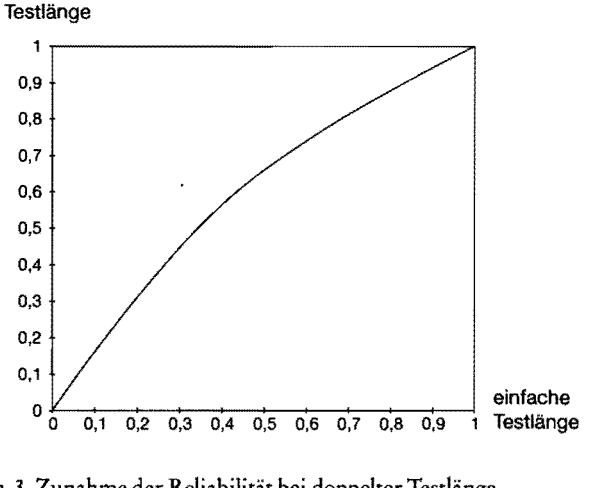

Abb. 3 Zunahme der Reliabilität bei doppelter Testlänge.

Diese durch die sogenannte Spearman-Brown-Formel beschriebene Beziehung ist zugleich Grundlage der *Split-Half Methode* der Reliabilitätsbestimmung. Dabei wird der Test in zwei parallele Testhälften zerlegt, so daß die Korrelation zwischen den Scores aus den beiden Testhälften deren Reliabilität ergibt. Aus dieser läßt sich mittels der Spearman-Brown-Formel schließlich die Reliabilität des ursprünglichen Tests errechnen.

Auf der Zerlegung des Tests in zwei oder mehrere Testteile beruht auch die sogenannte interne *Konsistenzanalyse*, die durch Berechnung des sogenannten Koeffizienten Alpha (Cronbach 1951) die Angabe einer unteren Schranke für die Reliabilität des Tests ermöglicht. Der Vorteil der internen Konsistenzanalyse gegenüber der Split-Half Methode besteht darin, daß ihre Anwendung nicht an die Voraussetzung paralleler Testteile gebunden ist. Der Koeffizient Alpha gibt die Reliabilität des Tests bereits dann exakt wieder, wenn die Testteile (lediglich) essentiell tau-äquivalent sind. Ist auch diese Voraussetzung nicht erfüllt, so ist die Reliabilität jedenfalls besser als der für Alpha errechnete Zahlenwert.

Ist die Reliabilität eines Tests bekannt, so kann man daraus den sogenannten *Standardmeßfehler* des Tests berechnen, das heißt die Standardabweichung des Meßfehlers in der Referenzpopulation. Dieser bildet die Grundlage für die inferenzstatistische Beurteilung von Testscores. So kann man zum Beispiel bei gegebenem Testergebnis einer Vp einen Vertrauensbereich angeben, durch welchen der True-Score der Vp mit einer vorgegebenen Sicherheit (z.B. 95 Prozent oder 99 Prozent) überdeckt wird. Oder man kann Hypothesen über den True-Score einer Vp auf statistische Signifikanz prüfen. Man kann die Testergebnisse zweier Vpn inferenzstatistisch miteinander vergleichen und so weiter.

Validität

Die Validität eines Tests gibt an, wie gut ein Test jenes Merkmal mißt, das er zu messen beansprucht. Dabei ist Vorsicht gegenüber dem bloßen Augenschein angebracht. Was zum Beispiel auf den ersten Blick als Intelligenzleistung erscheint, kann sich bei genauerer theoretischer und empirischer Analyse als hochgradig schicht- oder kulturabhängig erweisen. Bloße Augenschein-Validität ist für die

Konstruktion und Auswahl von Testitems daher kein hinreichendes Kriterium. Lienert (Lienert & Raatz 1994) unterscheidet drei Aspekte der Validität: inhaltliche Validität, kriteriumsbezogene Validität und Konstruktvalidität.

Unter *Inhaltsvalidität* versteht man, daß die Items eines Tests das zu messende Merkmal repräsentativ erfassen. Zum Beispiel ist ein Schulleistungstest inhaltlich valide, wenn seine Aufgaben eine repräsentative Auswahl aus dem Unterrichtsstoff darstellen.

Wenn man in der klassischen Testtheorie von Validität spricht, so ist damit die *Kriteriumsvalidität* gemeint. Diese wird bestimmt, indem man die Testergebnisse einer Stichprobe von Vpn mit einem Außenkriterium korreliert, das – vom Test unabhängig erhoben – das zu erfassende Merkmal valide widerspiegelt. So kann man zum Beispiel die Validität eines Intelligenztests für die Prognose von Schulleistungen bestimmen, indem man die Testergebnisse mit den Schulnoten korreliert. Wie man mathematisch zeigen kann, ist die so definierte Kriteriumsvalidität nie größer als die Quadratwurzel aus der Reliabilität des Tests. Reliabilität von Tests ist also eine notwendige Voraussetzung für valide Testergebnisse.

Im Gegensatz zu Inhalts- und Kriteriumsvalidität ist die *Konstruktvalidität* weniger pragmatisch als theoretisch orientiert. Ihre Bedeutung liegt nicht in unmittelbarer praktisch-diagnostischer Verwertbarkeit, sondern in der theoretischen Klärung dessen, was der Test mißt. Dies geschieht unter anderem auf der Grundlage der empirischen Untersuchung der Korrelationen des Tests mit Außenkriterien sowie auf der Grundlage seiner Korrelationen mit validitätsverwandten und validitätsdivergenten Tests (die einen ähnlichen beziehungsweise abweichenden Validitätsanspruch haben).

Literatur

Bortz, Jürgen, Gustav A. Lienert, Klaus Boehnke (1990): *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer.

Cohen, Jacob (1960): A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37–46.

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

Gulliksen, Harold (1950): *Theory of Mental Tests*. New York: Wiley.

Lazarsfeld, Paul F. (1950): Logical and mathematical foundations of latent

structure analysis. In: Samuel A. Stouffer, Louis Guttman, Edward A. Suchman, Paul F. Lazarsfeld, Shirley A. Star, John A. Clausen (Hg.): *Studies in social psychology in world war II.* Bd. IV. Princeton/N.J.: Princeton University Press.

Lienert, Gustav A. (1961): *Testaufbau und Testanalyse.* Weinheim: Beltz.

Rasch, Georg (1960): *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Rasch, Georg (1968): *A mathematical theory of objectivity and its consequences for model construction.* Proceedings of the European Meeting on Statistic, Econometrics and Management Science, Amsterdam, 2–7 september 1968.

Rohracher, Hubert (1971): *Einführung in die Psychologie* (10. Aufl.). Wien: Urban & Schwarzenberg.

Lehrbücher:

Fischer, Gerhard H. (1974): *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

Lienert, Gustav A., Ulrich Raatz (1994): *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Beltz.

Lord, Frederic M., Melvin R. Novick (1968): *Statistical Theories of Mental Test Scores.* Reading (MA): Addison-Wesley.

Rost, Jürgen (1996): *Lehrbuch Testtheorie Testkonstruktion.* Bern: Huber.

Steyer, Rolf, Michael Eid (1993): *Messen und Testen. Ein Lehrbuch.* Berlin: Springer.