Klassische Testtheorie

Heinrich Stumpf

Die Methoden der *klassischen* Testtheorie sowie ihrer Verallgemeinerungen und Modifikationen gehen auf eine im 19. Jahrhundert (Fechner, 1860; Wundt, 1862) begonnene Tradition zurück: In Analogie zu den entsprechenden Methoden der Naturwissenschaften wurden Verfahren der *Messung* in die Psychologie eingeführt. Diese Tradition verzweigte sich bald in die *reizzentrierten* Methoden der *Psychophysik* und die *individuumzentrierten* Verfahren der *Psychometrie* (Galton, 1883; J. McKeen Cattell, 1890). Die klassische Testtheorie (s. besonders Gulliksen, 1950; Lord & Novick, 1968; Allen & Yen, 1979; Lord, 1980; Kristof, 1983) stellt innerhalb der Psychometrie den ältesten und bis heute am weitesten verbreiteten Ansatz dar. Sie steht jedoch heute neben anderen testtheoretischen Modellen, wie den *Item-Response-* oder *Latent-Trait-*Modellen (s. die Beiträge von Scheiblechner und Roskam, in diesem Band).

Gegenüber diesen Ansätzen kann man die klassische Testtheorie insgesamt als *ein Arsenal pragmatisch orientierter Prinzipien oder Regeln zur Konstruktion, Erprobung und Evaluation psychometrischer Tests und zur Interpretation von Testergebnissen* charakterisieren. Im Gegensatz zur abstrakten Meßtheorie (s. Niederée & Narens, in diesem Band) wählt die klassische Testtheorie keinen *axiomatischen*, sondern einen *pragmatischen* Ansatz zur Messung von Personenmerkmalen. Sie betrachtet die ihr zur Messung vorgegebenen Merkmale als kontinuierliche Größen, die bei einer Person innerhalb eines theoretisch gesetzten Zeitrahmens, zumindest aber für die Dauer der Testung, invariant bleiben, hinsichtlich derer aber graduelle Unterschiede zwischen verschiedenen Personen bestehen. Man geht zunächst davon aus, daß die Merkmale grundsätzlich meßbar sind; ob diese Annahme berechtigt ist, wird nach dem Erfolg oder Mißerfolg bei der Erprobung und Evaluation von Testverfahren für die betreffenden Eigenschaften beurteilt. Die Messung erfolgt per conventionem in der Regel durch die Abzählung derjenigen Reaktionen des Probanden auf die Items eines Tests, die als „richtig", „symptomatisch" o. ä. gelten. Die Ergebnisse dieser Abzählung werden in anschließenden Datenanalysen meist wie intervallskalierte Meßwerte behandelt. Da für diese Annahmen keine meßtheoretische Begründung angegeben wird, wird die Diagnostik im Sinne der klassischen Testtheorie auch als *vereinbarte Messung* oder *Messung per fiat* bezeichnet. Ihre Rechtfertigung liegt letztlich in der Brauchbarkeit der Testergebnisse als Datenbasis zur psychologischen Theorienbildung und zur praktisch-diagnostischen Beurteilung von Personen, etwa in Form der Vorhersage zukünftigen Verhaltens (z.B. schulischer oder beruflicher Leistungen).

Zur Beurteilung der Brauchbarkeit dieser Ergebnisse formuliert die klassische Testtheorie eine Reihe von *Testgütekriterien* (insbesondere *Reliabilität* und *Validität*,

https://doi.org/10.25521/HQM31

s.u.) und eine Vielzahl von Regeln zur Evaluierung von Tests hinsichtlich dieser Kriterien sowie zur Optimierung von Testgüteeigenschaften. Als Voraussetzung dazu expliziert sie ein System von Grundannahmen zur Messung psychischer Eigenschaften. Diese Annahmen werden auch *Axiome* genannt. Dieser Teil der Testtheorie problematisiert aber nicht die eigentlichen meßtheoretischen Voraussetzungen, sondern er stellt lediglich eine formale Beschreibung der alltäglichen Beobachtung dar, daß Messungen im Bereich der differentiellen Psychologie im allgemeinen ein beträchtliches Ausmaß an Unzuverlässigkeit (Unreliabilität) anhaftet, und ist im wesentlichen die Grundlage der *Reliabilitätstheorie*. In diesem Teil der klassischen Testtheorie kommt formal gesehen keine Beschränkung auf Daten aus dem Bereich der Psychologie zum Ausdruck; entsprechende Ansätze finden sich auch auf anderen Gebieten, in denen Messung nur mit einem vergleichsweise beschränkten Grad an Reliabilität möglich ist.

Charakteristisch für die Psychometrie ist der Umstand, daß neben dem Reliabilitätsproblem dem *Validitätsproblem* entscheidende Bedeutung zukommt, welches sich vereinfacht auf die Frage bringen läßt, ob die Testergebnisse im Sinne der Konstrukte interpretiert werden können, die man zu messen beabsichtigt, d.h. inwieweit Tests ihrem eigentlichen Zweck gerecht werden.

1 Grundannahmen der klassischen Testtheorie

Grundlegend für die klassische Testtheorie ist die Annahme, daß das (beobachtete) Ergebnis eines Probanden in einem Test additiv aus zwei Komponenten zusammengesetzt ist, dem *wahren Wert* (*true score*) und dem *Meßfehler* (*error score*). Diese beiden Komponenten sind nicht direkt beobachtbar, und Versuche, sie operational zu definieren, führen zu keinem befriedigenden Ergebnis, weil man dabei stets von unrealistischen Voraussetzungen ausgehen muß, wie etwa der beliebigen Wiederholbarkeit einer Testung unter denselben Bedingungen. Deshalb führt man die beiden Größen zunächst nur durch bloße formale Definitionen ein. Novick (1966) konnte zeigen, daß sich auf diese Weise die wichtigsten Lehrsätze der klassischen Testtheorie – die sogenannten „Axiome", die im Gegensatz zu den Axiomen der Meßtheorie nur Aussagen bezüglich der *Meßwerte* darstellen – aus einigen wenigen einfacheren Annahmen herleiten lassen[1] (s. besonders Lord & Novick, 1968; Fischer, 1974; Lord, 1980).

Mit x_{ik} sei der beobachtete Wert eines beliebigen Probanden i aus einer Population I in einem Test k bezeichnet. Ferner sei unterstellt, die Person könne den Test beliebig oft bearbeiten, ohne daß das Verhalten in einer Testung das Ergebnis in einer anderen beeinflußte. Diese verschiedenen Testungen werden hier mit dem Index l bezeichnet, und man kann jedes Testergebnis als Realisierung einer Zufallsvariablen X_{ikl} auffassen. Praktisch wird eine solche experimentelle Situation in der Regel nicht zu realisieren sein, aber man begnügt sich mit der hypothetischen Konstruktion von X_{ikl} als nützlicher Grundlage für einige Definitionen und Ableitungen, deren Brauchbarkeit zur Beschreibung empirischer Sachverhalte sich erst später erweisen soll. Der wahre Wert τ_{ik} der Person i im Test k wird nun als der Erwartungswert von X_{ikl}

[1] Die vorliegende Darstellung folgt Lord (1980, S.3 ff.); Leser finden die Ableitungen in ausführlicher Form bei Lord und Novick (1968, S. 30 ff.).

definiert:

$$\tau_{ik} := E(X_{ikl}). \tag{1}$$

Die Meßfehler (E_{ikl}) bei den einzelnen hypothetischen Testungen werden definiert als *die Differenzen zwischen den jeweils beobachteten Werten und dem (konstanten) wahren Wert des Probanden*:

$$E_{ikl} := X_{ikl} - \tau_{ik}. \tag{2}$$

Aus diesen Definitionen folgt, *daß der Erwartungswert des Fehlers gleich Null ist*:

$$E(E_{ikl}) = E(X_{ikl} - \tau_{ik}) = E(X_{ikl}) - E(\tau_{ik}) = \tau_{ik} - \tau_{ik} = 0. \tag{3}$$

Wenn nach Gleichung (3) für alle Werte, die T_{*k} über verschiedene Personen annehmen kann, die Fehlerwerte denselben Erwartungswert haben – nämlich Null – dann müssen in der Population I die beiden Zufallsvariablen der wahren Werte (T_{*k}) und der Meßfehler (E_{*k}) zu Null korrelieren:

$$\rho(T_{*k}, E_{*k}) = 0. \tag{4}$$

Weiterhin geht die klassische Testtheorie davon aus, daß die Meßfehler in einem Test (E_{*k}) unkorreliert sind mit den Meßfehlern (E_{*m}) und den wahren Werten (T_{*m}) in einem beliebigen anderen Test:

$$\rho(E_{*k}, E_{*m}) = 0, \tag{5}$$

$$\rho(E_{*k}, T_{*m}) = 0. \tag{6}$$

Wenn nach Gleichung (4) die wahren Werte und Meßfehler in einem Test zu Null korrelieren, so bedeutet dies, daß ihre Varianzen zusammengenommen die Varianz der beobachteten Werte ergeben:

$$\sigma^2(T_{*k}) + \sigma^2(E_{*k}) = \sigma^2(T_{*k} + E_{*k}) = \sigma^2(X_{*k}). \tag{7}$$

Aus den Gleichungen (2), (4) und (7) läßt sich die *Kovarianz der beobachteten mit den wahren Werten* ableiten als:

$$\begin{aligned} \sigma(X_{*k}, T_{*k}) &= \sigma((T_{*k} + E_{*k}), T_{*k}) = \\ \sigma^2(T_{*k}) &+ \sigma(E_{*k}, T_{*k}) = \sigma^2(T_{*k}). \end{aligned} \tag{8}$$

Anhand der Gleichung (8) definiert man die *Reliabilität* (Zuverlässigkeit) eines Tests k als das Quadrat der Korrelation zwischen den beobachteten und den wahren Werten in einer Population I:

$$\rho^2(X_{*k}, T_{*k}) = \frac{\sigma^2(X_{*k}, T_{*k})}{\sigma^2(X_{*k})\sigma^2(T_{*k})} = \frac{\sigma^2(T_{*k})}{\sigma^2(X_{*k})} = 1 - \frac{\sigma^2(E_{*k})}{\sigma^2(X_{*k})}. \tag{9}$$

In der obigen Gleichung ist die Reliabilität noch als Beziehung zwischen einer nicht beobachtbaren (T_{*k}) und einer beobachtbaren (X_{*k}) Variablen ausgedrückt. Sie läßt sich jedoch auch anhand von Daten schätzen (vgl. Abschnitt 2.2), so daß sich das Verhältnis der Varianz der wahren Werte zu der der beobachteten Werte in einem

Test empirisch beurteilen läßt; mit der Reliabilitätsschätzung hat man zugleich eine Schätzung der Fehlervarianz $\sigma^2(E_{*k})$, deren Quadratwurzel man als *Standardmeßfehler* bezeichnet.

Die Gleichungen (1) bis (6) enthalten die Grundannahmen der klassischen Testtheorie. Wie man an der Zerlegung der Varianz des beobachteten Werts (Gleichung 7) erkennt, läßt sich das Modell der klassischen Testtheorie als ein Spezialfall der Faktorenanalyse auffassen (vgl. z.B. Lord & Novick, 1968, S. 530 ff.; Schönemann & Borg, in diesem Band). Von dieser Eigenschaft der Testtheorie wird bei ihrer Anwendung in vieler Hinsicht Gebrauch gemacht.

Zu den Axiomen treten die folgenden Definitionen äquivalenter Messungen in einer Population I:

1. Zwei Tests k und m heißen *streng parallel*, wenn (a) der wahre Wert jeder Person i im Test k gleich ihrem wahren Wert in m ist und wenn (b) die Meßfehlervarianzen der Person in beiden Tests gleich sind:

$$\tau_{ik} = \tau_{im} \quad \text{und} \quad \sigma^2(E_{ik}) = \sigma^2(E_{im}). \tag{10}$$

   Parallele Tests messen dieselbe Eigenschaft gleich zuverlässig.

2. Zwei Tests heißen *τ-äquivalent*, wenn nur die erste der beiden oben genannten Bedingungen (a) und (b) erfüllt ist:

$$\tau_{ik} = \tau_{im}. \tag{11}$$

   Die beiden Tests messen dieselbe Eigenschaft, aber u.U. mit unterschiedlicher Zuverlässigkeit.

3. Zwei Tests heißen *essentiell τ-äquivalent*, wenn für jede Person aus I der wahre Wert im ersten Test durch eine additive Konstante (a_{km}) in den wahren Wert im zweiten Test zu transformieren ist; die Reliabilitäten brauchen nicht gleich zu sein:

$$\tau_{ik} = \tau_{im} + a_{km}. \tag{12}$$

Zwei Tests, die als parallel gelten, ohne daß ihre Äquivalenzeigenschaften bekannt sind, werden als *nominell parallel* bezeichnet. Verfahren zur statistischen Überprüfung von Äquivalenzannahmen beschreibt z.B. Rasmussen (1988).

Aus der Definition streng paralleler Tests läßt sich u.a. eine Interpretation der Reliabilität ableiten: Die Reliabilität eines Tests k ist gleich der Korrelation der beobachteten Werte X_{*k} mit denen in einem parallelen Test X_{*m}.

Zur Veranschaulichung des oben bloß definitorisch und folgernd („syntaktisch") eingeführten Modells und im Vorgriff auf einige der dagegen vorgebrachten grundsätzlichen Einwände seien hier einige Anmerkungen zu inhaltlichen („semantischen") Interpretationen der Grundbegriffe und -aussagen und zu einigen allgemeinen Eigenschaften des Modells gemacht. Ganz allgemein gilt, daß die Probleme der klassischen Testtheorie „nicht so sehr im syntaktischen Teil" liegen, den man „im wesentlichen mit dem gesamten Axiomensystem und Formelapparat der Testtheorie gleichsetzen kann" – „als vielmehr im semantischen" (Fischer, 1974, S. 26).

Der wahre Wert repräsentiert per definitionem die Gesamtheit aller systematischen Varianzkomponenten, der Meßfehler die Gesamtheit aller unsystematischen Varianzanteile, die das Testergebnis bestimmen. Die systematischen Varianzanteile sind jedoch nicht ausschließlich solche, die auf der Variation des zu messenden Merkmals beruhen, sondern auch solche, die auf spezifischen Eigenschaften der Meßmethode (Test) beruhen. So sind etwa die Ergebnisse in vielen Skalen von Persönlichkeitsfragebögen mit sozialer Erwünschtheit konfundiert. Das klassische Modell rechnet derartige systematische Varianzanteile dem wahren Wert zu, nicht dem Meßfehler; sie sind aber merkmals*untypisch*. Der wahre Wert ist also – trotz seiner Bezeichnung – im allgemeinen nicht gleichzusetzen mit dem Ergebnis einer „perfekten" (d.h. gänzlich validen (s. Abschnitt 2.2)), sondern nur mit dem Resultat einer gänzlich reliablen (vgl. Abschnitt 2.1) Messung.

Der wahre Wert ist *nur in bezug auf den jeweils verwendeten Test definiert*; er ist damit abhängig von dem jeweils verwendeten Satz von Items. Die Grundintention etwa des Rasch-Modells (Rasch, 1960), Personenparameter unabhängig von der Zusammensetzung der jeweils vorgegebenen Itemmengen zu schätzen, ist der klassischen Testtheorie in dieser Form fremd. Allerdings wirkt sich dieses Bestreben hier indirekt aus in den Ansätzen zur Konstruktion äquivalenter Testformen, zur Transformierung von Rohwerten auf einheitliche Standardskalen und in verschiedenen Arten des „*equating*" von Testwerten, d.h. der Transformation der Werte eines Tests auf die Skala eines anderen (vgl. im einzelnen z.B. die Übersichten bei Angoff, 1971; Allen & Yen, 1979, S. 148 ff.; Holland & Rubin, 1982).

Die entscheidenden psychometrischen Eigenschaften (wie Reliabilität und Validität), die Tests zugeschrieben werden, sind abhängig von den Verteilungen der wahren Werte und Meßfehler in der Population (bzw. Stichprobe) der jeweils Getesteten. Diese Populationsabhängigkeit wird an den Gleichungen (7) bis (9) deutlich. Besonders wichtig ist in diesem Zusammenhang, daß die Reliabilität auch von der Verteilung der *wahren Werte* in der Population abhängt; entsprechendes gilt für die Validität (vgl. Abschnitt 2.3). Das Prinzip der Populations- bzw. Stichprobenabhängigkeit ist dem Sozialwissenschaftler vertraut, für den Meßtheoretiker muß es jedoch befremdlich erscheinen, daß grundlegende Eigenschaften der Meßmethoden abhängig sind von der Grundgesamtheit der Objekte, die gemessen werden. Die klassische Testtheorie begnügt sich jedoch damit, daß Aussagen, die über Tests oder anhand von Testergebnissen gemacht werden, nur im Bezugsrahmen einer bestimmten Population gelten. Nur in einigen Spezialfällen des „*test equating*" (s. besonders Braun & Holland, 1982) wird versucht, über eine Population hinaus zu generalisieren. Sollen anhand eines Tests Aussagen über Probanden aus einer anderen Grundgesamtheit gemacht werden, so bedarf es einer erneuten Erprobung und Evaluierung, oft auch einer – zumindest teilweisen – Neukonstruktion.

Das Postulat der Nullkorrelation zwischen dem wahren Wert und dem Meßfehler im oben definierten Sinne ist in bestimmten Fällen unrealistisch. So ist dann, wenn der wahre Wert in der Nähe der „Testdecke" (d.h. des maximal erreichbaren Wertes in einem Test) liegt, die Varianz des Meßfehlers geringer zu veranschlagen als wenn der *true score* weiter von der Decke entfernt liegt (Lumsden, 1976). Dieser Umstand steht in Konflikt mit der gängigen Praxis, einen Test durch einen einzigen Standardmeßfehler zu charakterisieren. So sind denn auch verschiedene Methoden

beschrieben worden, Standardmeßfehler für unterschiedliche Niveaus des beobachteten Werts zu schätzen (vgl. z.B. Feldt, Steffen & Gupta, 1985; Blixt & Shama, 1986), ein Vorgehen, das besonders dann angezeigt ist, wenn diagnostische Entscheidungen nach verschiedenen kritischen Testwerten getroffen werden sollen (American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME), 1985).

Die Annahme des Intervallskalenniveaus bei der Analyse von Testwerten bleibt ein Postulat. Da empirische Beziehungen, die diese Annahme schlüssig begründen könnten, kaum befriedigend zu operationalisieren sind, nimmt man sich gleichsam die Freiheit, derartige Relationen „per fiat" vorauszusetzen „und so etwa gleichen Differenzen gleiche psychologische Bedeutung zuzusprechen" (Kristof, 1983, S. 599).

Weitere Probleme ergeben sich, wenn die klassische Testtheorie auf Anwendungsgebiete übertragen wird, für die sie ursprünglich nicht formuliert wurde. In diesem Zusammenhang sind besonders die Schwierigkeiten zu nennen, die sich bei der Anwendung des klassischen Ansatzes in der Veränderungsmessung (s. z.B. Fischer, 1974, S. 128 ff.) und bei kriteriumsorientierten Tests (s. z.B. Klauer, 1983) ergeben.

Trotz dieser Probleme und Unzulänglichkeiten ist die klassische Testtheorie nach wie vor der dominierende Ansatz in der Psychometrie; viele Verfahren, die auf ihrer Grundlage entwickelt wurden, haben sich in der Praxis bewährt, so daß man klassisch konstruierten Tests nicht generell die Tauglichkeit in Psychologie und Pädagogik absprechen und der Testtheorie insgesamt den Fortbestand und weiteren Ausbau nicht verweigern sollte (vgl. Klauer, 1978, S. 7; Kristof, 1983, S. 599).

2 Testgütekriterien

Die *Gütekriterien psychologischer Tests* wurden in ihren Einzelheiten weitgehend unabhängig von den Grundannahmen der klassischen Testtheorie formuliert. Mit Lienert und Raatz (1994) unterscheidet man als Hauptgütekriterien die *Objektivität*, *Reliabilität* und *Validität* und als Nebengütekriterien die *Normierung*, *Vergleichbarkeit*, *Ökonomie* und *Nützlichkeit*. Seit Ende der sechziger Jahre wird als weiteres Kriterium zunehmend die *Fairneß* diskutiert (s. Abschnitt 2.4). Bezüglich der Nebengütekriterien muß an dieser Stelle auf Lienert und Raatz (1994, S. 7 ff.) verwiesen werden.

2.1 Objektivität

Unter *Objektivität* versteht man die Unabhängigkeit der Testergebnisse von situativen Einflüssen bei der Testdurchführung, von der Person des Auswerters und desjenigen, der die Ergebnisse interpretiert; demgemäß spricht man auch von *Durchführungs-*, *Auswertungs-* und *Interpretationsobjektivität* (Lienert & Raatz, 1994, S. 7f.; Michel & Conrad, 1982, S. 15 f.). Die Absicherung der Objektivität erfolgt im allgemeinen durch Vorgabe präziser, möglichst einfacher Verfahrensvorschriften. Je nach Art des Tests und seines Anwendungskontextes können sich aus der Objektivitätsforderung besondere praktische Probleme ergeben, wie z.B. die der Betrugssicherheit, der Vermeidung von Testleitereffekten, der Fehlersicherheit bei der Auswertung etc. Neben den genannten Momenten werden von einigen Autoren noch zusätzliche

Aspekte mit dem Objektivitätsbegriff verbunden, so z.B. das Kriterium der Undurchschaubarkeit des Tests für den Probanden (Cattell, 1958; Häcker, 1982). Diese zusätzlichen Kriterien sind aber nur auf bestimmte Klassen von Tests anwendbar.

2.2 Reliabilität

Die Schätzung der Reliabilität eines Tests erfolgt in der Praxis meist dadurch, daß man Ergebnisse zumindest zweier (wenigstens nominell) äquivalenter Messungen des zugrunde liegenden Merkmals in Beziehung zueinander setzt. Dabei kann die Reliabilität im Prinzip auf drei verschiedene Arten geschätzt werden:

- als *Korrelation* der Ergebnisse in zwei oder mehr *Paralleltests (Paralleltestreliabilität)*,
- als *Stabilität* der Ergebnisse bei wiederholter Anwendung eines Tests und
- als *Konsistenz* der Ergebnisse in verschiedenen Teilen eines Tests.

Ein Verfahren der Reliabilitätsschätzung, das ohne Testwiederholung und -aufteilung auskommt und so keiner der drei genannten Schätzverfahren zugeordnet werden kann, ist die Faktorenanalyse (vgl. Schönemann & Borg, in diesem Band). Faktorisiert man eine Batterie von Tests, so stellt die Kommunalität eines Untertests im gemeinsamen Faktorenraum zusammengenommen mit der Spezifität der jeweils betrachteten Variablen eine Schätzung der Reliabilität dar (vgl. im einzelnen z.B. Überla, 1968, S. 54 ff.). Diese Form der Reliabilitätsbestimmung ist aber recht grob und wird selten angewandt.

Die Definition in (9) darf nicht darüber hinwegtäuschen, daß es einen einzigen, allgemeinverbindlichen Reliabilitätskoeffizienten für einen Test ebensowenig gibt wie einen alleingültigen Validitätskennwert (vgl. hierzu Abschnitt 2.3). So hat etwa ein Stabilitätskoeffizient einen ganz anderen Stellenwert, wenn es um eine langfristige Leistungsprognose geht, als wenn es dem Diagnostiker um die Einschätzung einer akuten psychischen Belastung zu tun ist. Es ist nicht zuletzt eine inhaltliche Frage, welche Varianzanteile man dem wahren Wert und welche man dem Meßfehler zuordnet (vgl. besonders Stanley, 1971). So rechnet man etwa tageszeitlich bedingte Schwankungen von Testwerten meist dem Meßfehler zu; es gibt jedoch eine Reihe von diagnostischen Fragestellungen, bei denen die Tageszeit der Testabnahme als konstitutiv für den wahren Wert angesehen wird; in solchen Fällen sollte auch die Erhebung von Daten zur Reliabilitätsschätzung an die entsprechende Tageszeit gebunden sein.

Die theoretisch eleganteste Interpretation der Reliabilität eines Tests ist die *Korrelation des beobachteten Testwerts mit dem in einem echten Paralleltest*. Allein zur Schätzung der Reliabilität eines Tests wird die Paralleltestmethode aber in der Praxis nur selten angewendet. Zwar gibt es differenzierte Verfahren zur Konstruktion äquivalenter Testformen, deren Anwendung ist aber recht aufwendig und bietet in der Regel keine Gewähr dafür, daß die verschiedenen Testformen in einer Kreuzvalidierung tatsächlich den Kriterien strenger Parallelität genügen. Man bedient sich daher solcher Methoden in der Regel nur dann, wenn aus praktisch-diagnostischen Gründen mehrere äquivalente Testformen erforderlich sind. Häufig werden aber in

der Literatur als Reliabilitätsschätzungen Korrelationen zwischen nominell parallelen Testformen mitgeteilt, die man auch als „*alternate forms reliabilities*" bezeichnet. Selbst wenn die Testformen aber streng parallel sind, können sich aus der Vorgabe ähnlicher Items Probleme bezüglich der Vergleichbarkeit der Testungen ergeben. Dies ist insbesondere dann der Fall, wenn die Testformen eine Funktion prüfen, deren Vollzug für den Probanden relativ ungewohnt ist. In solchen Fällen kann von der ersten zur zweiten Testung ein erheblicher Übungseffekt eintreten.

Zur Schätzung der *Stabilität* (oder *Retestreliabilität*) wird ein und derselbe Test zweimal oder öfter vorgegeben, und die Ergebnisse werden korreliert. Die Länge der Retest-Intervalle richtet sich nicht nur nach psychometrischen, sondern auch nach inhaltlich-psychologischen Gesichtspunkten. Im Rahmen der Persönlichkeitsdiagnostik beispielsweise werden oft Stabilitätskoeffizienten für *verschieden lange Intervalle* (Tage bis Jahre) berechnet. Formal gesehen stellt die zweimalige Vorgabe eines Tests zwar einen Spezialfall der Anwendung paralleler Testformen dar, denn jeder Test ist zu sich selbst parallel, aus der bloßen Wiederholung der Vorgabe ergibt sich aber eine Reihe von Problemen, die dazu führen können, daß die beiden Testungen nicht äquivalent sind, wie im Falle des oben beschriebenen Übungseffekts. Eine ausführliche Darstellung dieser Probleme findet der Leser bei Stanley (1971, S. 404 ff.). Bei langen Retest-Intervallen kommt als weiteres Problem oft eine selektive Schrumpfung der Stichprobe hinzu, die die Repräsentativität der erhobenen Daten in Frage stellt. In diesem Fall ergibt sich zudem oft die Notwendigkeit, Unreliabilität des Tests und personenspezifische Variation des wahren Werts zu unterscheiden. Ein Verfahren dazu beschreibt Heise (1969).

Die Bestimmung der *internen Konsistenz* von Tests ist die am häufigsten angewandte Methode der Reliabilitätsschätzung. Dies beruht nicht zuletzt darauf, daß man bei Anwendung dieser Methode mit einer einmaligen Vorgabe des Tests auskommt. Die Analyse der internen Konsistenz beruht auf der rechnerischen Aufteilung des Tests und der Schätzung der Reliabilität anhand der Beziehungen, die zwischen den Testteilen bestehen. Theoretisch kommen als Verfahren der Aufteilung alle denkbaren Prozeduren von der Bildung von *Testhälften* bis hin zur Betrachtung der *einzelnen Items als Testteile* in Betracht. In der praktischen Anwendung werden aber meist nur die beiden genannten Extremfälle realisiert.

Historisch gesehen wurden zuerst Verfahren entwickelt, die von einer Form der Testhalbierung, meist einer Aufteilung in Items mit ungerader und gerader Ordnungszahl (*Odd-Even-Methode*), ausgehen und aus der Korrelation der Testhälften nach der „*prophecy formula*" von Spearman und Brown die sog. *Halbierungsreliabilität* des gesamten Tests schätzen (vgl. Stanley, 1971, pp. 370 ff.). Von diesen Verfahren gibt es zahlreiche Varianten für spezielle Anwendungsfälle (s. Lienert & Raatz, 1994). Die Methoden lassen sich unter bestimmten Bedingungen als Spezialfälle des weiter unten beschriebenen Verfahrens der *klassischen Konsistenzanalyse* darstellen (vgl. Lord & Novick, 1968, S. 219 ff.).

Die Konsistenzanalyse ist historisch gesehen aus Ansätzen hervorgegangen, die Verfahren der Reliabilitätsschätzung aus Testhälften zu generalisieren auf Fälle, in denen mehr als zwei Testteile betrachtet werden (Kuder & Richardson, 1937). Die Konsistenzschätzung läßt sich anhand des α-Koeffizienten nach Cronbach (1951) veranschaulichen. Dieser Index ist eine untere Schätzung der Reliabilität; er ist dann

gleich der Reliabilität, wenn die Testteile zumindest essentiell τ-äquivalent sind. Angenommen, ein Test, dessen beobachtete Werte die Varianz σ^2 haben, sei in M verschiedene Teile $X_m (m = 1, \ldots, M)$ aufgeteilt, die jeweils die Varianz σ_m^2 haben. Die Reliabilität des Tests ist dann mindestens gleich:

$$\alpha := \left(\frac{M}{M-1}\right)\left(\frac{\sigma^2 - \sum_{m=1}^{M} \sigma_m^2}{\sigma^2}\right). \tag{13}$$

Als „Teile des Tests" kann man Testhälften bis hin zu einzelnen Items betrachten. Meist wird bei der Berechnung des α-Koeffizienten von Items ausgegangen. Sind diese dichotom (0,1) kodiert, so entspricht α einem bereits von Kuder und Richardson (1937) abgeleiteten Koeffizienten (KR20). α ist unter anderem eine Funktion der mittleren Interkorrelation der Items. Der α-Kennwert ist das heute gebräuchlichste Konsistenzmaß, es gibt aber noch eine Vielzahl weiterer Konsistenzindizes (s. z.B. Cronbach, 1988a).

Bei der Schätzung der Konsistenz eines Tests steht als Alternative zur Berechnung von Indizes wie α die konfirmatorische Faktorenanalyse des Tests auf Itemebene zur Verfügung (Fleishman & Benson, 1987; Reuterberg & Gustafsson, 1992; vgl. auch Rietz, Rudinger & Andres, in diesem Band). Die Reliabilität läßt sich aus Parameterschätzungen des faktorenanalytischen Modells bestimmen. Ein wichtiger Vorzug dieses Verfahrens liegt darin, daß sich damit Annahmen zur Äquivalenz der Testteile empirisch prüfen lassen. Die wenigen bisher vorliegenden Vergleiche dieser beiden Arten der Konsistenzanalyse sprechen aber dafür, daß α gegenüber Verletzungen der Äquivalenzannahme relativ robust ist.

Verfahren zur Prüfung statistischer Hypothesen über α-Koeffizienten (wie der Annahme der Gleichheit zweier α-Indizes in verschiedenen Populationen oder der Annahme, daß α in einer Population einen bestimmten Wert hat) beschreiben Feldt, Woodruff und Salih (1987), eine Übersicht über Methoden zur Prüfung der Gleichheit oder Unterschiedlichkeit der α-Koeffizienten verschiedener Tests in einer Population geben Woodruff und Feldt (1986).

Der Koeffizient α wird oft auch als „Homogenitätsindex" bezeichnet. Dabei ist jedoch zu beachten, daß der Homogenitätsbegriff in der klassischen Testtheorie zwar nicht ganz einheitlich, im allgemeinen jedoch enger gefaßt wird als der Konsistenzbegriff. Konsistenz im Sinne des α-Koeffizienten bedeutet lediglich, daß die *durchschnittliche* Interkorrelation der Testteile hoch ist, einzelne Teile des Tests können untereinander aber immer noch niedrig korrelieren. Homogenität bedeutet darüber hinaus, daß der Test faktorenanalytisch gesehen eindimensional ist. Green, Lissitz und Mulaik (1977) zeigten, daß ein Test mit vergleichsweise hohem α immer noch multifaktoriell sein kann. Derartige Beobachtungen haben zur Entwicklung einer Reihe spezieller *Homogenitätsindizes* geführt (s. Hattie, 1985), die zumeist auf Ergebnissen von Faktorenanalysen oder auf der Analyse der Verteilung der Iteminterkorrelationen (Piedmont & Hyland, 1993) beruhen.

In der Praxis wird die Konsistenzschätzung meist mit einer *Item-* und *Testanalyse* verbunden. Dabei werden neben *Itemschwierigkeits-* auch *Trennschärfe*-Indizes, d.h. Korrelationen der Items mit dem Gesamtwert des Tests (s. im einzelnen Lord & Novick, 1968, S. 327 ff.; Lienert, 1969, S. 87 ff.) berechnet, anhand derer man durch Itemauswahl eine systematische Konsistenzmaximierung betreiben kann. Mit

einer Anzahl spezieller Verfahren lassen sich darüber hinaus Untermengen von Items identifizieren, die maximale Konsistenz oder Homogenität aufweisen (z.B. Serlin & Kaiser, 1978; Bührer & Kohr, 1984). Nach Anwendung derartiger Maximierungstechniken sind aber *Kreuzvalidierungen* der Ergebnisse erforderlich.

2.3 Validität

Unter der *Validität* eines Tests versteht man die Gültigkeit der Interpretationen, Schlußfolgerungen und Vorhersagen sowie die Angemessenheit der diagnostischen Entscheidungen, die aus den Testergebnissen abgeleitet werden. Den Prozeß der Untersuchung der Testgültigkeit bezeichnet man als *Validierung* (z.B. Cronbach, 1971). Ganz allgemein betrachtet kann man die Validität von Testergebnissen als einen Spezialfall der Gültigkeit von Resultaten psychologischer Forschung allgemein auffassen (Brinberg & McGrath, 1982; Messick, 1988). In den klassischen Arbeiten aus dem Bereich der Psychometrie (z.B. Cronbach & Meehl, 1955; Loevinger, 1957; Cronbach, 1971) erfuhr das Validitätskonzept jedoch eine eigene, charakteristische Ausprägung.

Üblicherweise werden drei globale Aspekte der Validität unterschieden, nämlich *Inhalts- (Kontent-), kriterienbezogene* und *Konstruktvalidität*. Je nachdem, zu welchem Zweck ein Test verwendet werden soll, erstreckt sich die Validierung auf alle drei Bereiche, oder sie arbeitet einen oder zwei Aspekte besonders heraus (z.B. AERA, APA & NCME, 1985). Im Idealfall schließt eine Validierung aber Studien in allen drei Bereichen ein (z.B. AERA, APA & NCME, 1985; Messick, 1988).

Zur Überprüfung der *Inhaltsvalidität* oder zur Erstellung eines inhaltsvaliden Tests muß zunächst eine Grundgesamtheit (ein „Universum") von Items definiert werden, in bezug auf welche anhand der Testergebnisse Aussagen gemacht werden sollen. Geht es beispielsweise in einem Test um die Beherrschung des kleinen Einmaleins, so ist der Bereich, auf den geschlossen werden soll, durch die Gesamtheit aller Aufgaben zum kleinen Einmaleins gegeben. Man versteht unter *Inhaltsvalidität* eines Tests das Ausmaß, in dem die Testitems repräsentativ für eine derartige zuvor definierte Grundgesamtheit sind (Klauer, 1984). Es sind eine Reihe von Verfahren beschrieben worden, mit denen man, ausgehend von vorgegebenen Inhalten (etwa Lehrstoffen zu bestimmten Themen), Grundmengen von Items definieren und die Items selbst generieren kann (Roid & Haladyna, 1982; Feger, 1984; Schott & Wieberg, 1984). In vielen Bereichen ermöglicht es auch die Facettentheorie (s. Borg, in diesem Band), die Itemgrundgesamtheit differenzierter und exakter zu beschreiben, als dies durch ad hoc-Verfahren möglich ist (Edmundson, Koch & Silverman, 1993).

Liegt ein Item-Universum fest, so bedarf es der Anwendung von *Sampling*-Vorschriften, um repräsentative Stichproben von Items zu erzeugen. Oft genügen dazu einfache Zufallsprozeduren nicht, sondern man muß – bei in sich nach inhaltlichen oder formalen Gesichtspunkten gegliederten Item-Grundmengen – Quotenverfahren anwenden (Klauer, 1984). Grundsätzlich gilt es bei der Konstruktion kontentvalider Tests, nicht nur die Inhalte der Items festzulegen, sondern auch die Art, wie der Proband die Items zu bearbeiten hat – also z.B. im Mehrfachwahl-Modus oder in Form offener Beantwortung – und wie die Reaktionen des Probanden ausgewertet und zu Testwerten zusammengefaßt werden sollen. Zu Recht ist darauf hingewiesen worden, daß selbst dann, wenn die Inhalte eines Tests die intendierte Grundgesamtheit

angemessen repräsentieren, der Test möglicherweise immer noch keinen Rückschluß auf die Grundgesamtheit erlaubt. Dies ist z.B. dann der Fall, wenn Mehrfachwahl-Aufgaben so schlecht konstruiert sind, daß sie sich allein durch Betrachtung der Items und ohne Rekurs auf die Inhalte lösen lassen. Manche Autoren ziehen es daher vor, nicht von Kontent*validität* sondern von Kontent*repräsentativität* zu sprechen und Fragen, die in den Bereich der Konstruktvalidierung (s.u.) gehören, vom Problem der Generierung repräsentativer Itemmengen zu unterscheiden. Aus diesen Gründen wird auch vielfach die Auffassung vertreten, daß eine Überprüfung der Inhaltsvalidität allein als Rechtfertigung für den Einsatz eines Tests im Regelfall nicht ausreichend ist (vgl. besonders Messick, 1975, 1988).

Die Vielzahl der vorgeschlagenen Verfahren zur Erstellung kontentvalider Itemsätze darf nicht darüber hinwegtäuschen, daß sich längst nicht in allen Fällen die Item-Universa so präzise definieren lassen, daß die erwähnten Verfahren zur Itemgenerierung anwendbar sind. Dies gilt z.B. für weite Bereiche der Persönlichkeits- und der Berufseignungsdiagnostik. In solchen Fällen ist man nach wie vor auf intuitive Verfahren der Itemerstellung und die nachträgliche Überprüfung der Inhaltsvalidität – vornehmlich durch Expertenurteil – angewiesen.

Inhaltsvalidität ist naturgemäß für lernzielorientierte Tests ein vorrangiges Gütekriterium. Sie ist aber auch für die Konstruktion praktisch jedes anderen Tests von Bedeutung; selbst wenn es bei einer Testentwicklung nur darum geht, ein isoliertes Kriterium, wie etwa den Erfolg in einem bestimmten Beruf, vorherzusagen, stellen eine genaue Beschreibung der Berufsanforderungen und die Erzeugung eines dafür repräsentativen Itempools eine günstige Voraussetzung dafür dar, mit dem zu konstruierenden Test eine akzeptable Vorhersagevalidität (s.u.) zu erreichen.

Bei der Schätzung der *kriterienbezogenen Validität* wird das Testergebnis in Beziehung gesetzt zu einer externen Variablen (Kriterium), die entweder selbst ein direktes Maß des zu erfassenden Merkmals ist oder zu der aus praktischen oder theoretischen Gründen durch den Test Aussagen (meist: Vorhersagen) gemacht werden sollen. Liegt das Kriterium zum Zeitpunkt der Testung bereits vor oder wird es zur selben Zeit erhoben, spricht man von *konkurrenter* oder *Übereinstimmungsvalidität*, fällt es erst später an, von *prognostischer* („prädiktiver") oder *Vorhersagevalidität*. Oft wird eine konkurrente Validierung auch stellvertretend für eine prognostische durchgeführt (wenn man z.B. Berufstätige einen Test bearbeiten läßt, der zur Prognose des Erfolgs in demselben Beruf verwendet werden soll). Technische Einzelheiten der kriterienbezogenen Validierung, der Gewichtung von Testergebnissen und der Auswahl von Testitems unter dem Gesichtspunkt der Validitätsmaximierung sind anderenorts ausführlich dargestellt worden (s. z.B. Lord & Novick, 1968; Wiggins, 1973; Ulrich, 1985).

Obwohl das Konzept der kriterienbezogenen Validität auf einem einfachen Grundgedanken beruht und vordergründig betrachtet mit einem Minimum an theoretischen Voraussetzungen auszukommen scheint, führt es bei konsequenter Anwendung zu einer ganzen Reihe von Problemen, von denen im folgenden nur einige erwähnt werden können.

Es gibt strenggenommen nicht *die* kriterienbezogene Validität eines Tests, sondern eine Vielzahl von Validitäten in bezug auf verschiedene Kriterien. Diese Validitäten sind ihrerseits von zahlreichen Randbedingungen abhängig. Findet etwa

eine Auslese von Personen anhand der Testergebnisse statt, so ändern sich die Validitäten in Abhängigkeit vom Anteil der Ausgewählten und Abgelehnten; ändert sich die Verteilung der Kriteriumsmaße, so ändern sich auch die Validitäten (sind die Kriteriumsmaße z.B. dichotom (z.B. Erfolg oder Mißerfolg in einem Examen), so variieren die Validitäten mit den Erfolgsquoten (vgl. hierzu im einzelnen z.B. Lord & Novick, 1968, S. 140 ff., 275 ff., sowie Wiggins, 1973). Derartige Beobachtungen sind historisch gesehen einer der wichtigsten Beweggründe gewesen, ausgehend vom Konzept der kriterienbezogenen Testgültigkeit eine allgemeinere Validitätstheorie zu entwickeln (Cronbach & Meehl, 1955; Loevinger, 1957) und die Validitätsbeurteilung eines Tests nicht auf die Betrachtung einiger isolierter Kriteriumskorrelationen zu beschränken.

Das Ergebnis einer kriterienbezogenen Validierung hängt entscheidend von der *Operationalisierung des Kriteriums* ab; da Kriteriumsmaße ihrerseits fehlerbehaftet sind, wurde analog zu dem Kanon der klassischen Testtheorie ein System von Gütemaßstäben für Kriteriumsmaße entwickelt (Guion, 1965, S. 119 f.).

Ein immer wiederkehrendes Problem in Studien zur prognostischen und in stellvertretend dafür durchgeführten Untersuchungen zur konkurrenten Validität besteht in der *Varianzeinschränkung* der Prädiktoren oder der Kriteriumsmaße. Selbst wenn in einer Erprobungsphase die Anwendung eines Tests (etwa zur Berufseignung) nicht formell mit Interventionsmaßnahmen wie Auslese, Plazierung oder Beratung verbunden ist, ermöglichen die erhobenen Test- und Kriteriendaten nur in relativ seltenen Fällen unvoreingenommene Schätzungen der Populationsvarianzen, da beispielsweise vor allem Probanden mit ungünstigen Testergebnissen die jeweilige Berufstätigkeit oder Ausbildung gar nicht erst aufgenommen haben und Probanden mit schlechten Berufsleistungen bis zum Zeitpunkt der Kriterienerhebung eher aus der betreffenden Tätigkeit ausgeschieden sind als Erfolgreiche. Die damit verbundene Einschränkung der Varianz auf der Prädiktor- oder Kriterienseite bedingt oft eine Unterschätzung der tatsächlichen Prognosekraft eines Tests (vgl. z.B. Kaufman, 1972). Seit längerem lagen Verfahren zur rechnerischen Korrektur von Varianzeinschränkungen vor (s. z.B. Alexander, Carson, Alliger & Barrett, 1984), deren Anwendung in der Regel aber die Kenntnis der uneingeschränkten Populationsvarianzen voraussetzte. Hanges, Rentsch, Yusko und Alexander (1991) aber schlugen eine Prozedur vor, die Korrekturen von Validitätskoeffizienten auf Varianzeinschränkung mit einem Minimum an Information über die Stichprobendaten hinaus erlaubt.

In Studien zur kriterienbezogenen Validität macht man zudem nicht selten die Beobachtung, daß die Korrelation zwischen Testwert und Kriteriumsmaß gruppenspezifisch ist, daß also der Test das Kriterium für eine Gruppe von Probanden besser vorhersagt als für eine andere. In diesem Falle spricht man von *differentieller Validität*. Befunde dieser Art haben dazu geführt, daß man in Validitätsstudien oft routinemäßig prüft, ob die kriterienbezogene Validität des Tests in verschiedenen Untergruppen der Population dieselbe ist (z.B. Stumpf & Nauels, 1988).

Unter *Konstruktvalidität* versteht man die Gültigkeit der Interpretation der Ergebnisse eines Tests im Sinne eines oder mehrerer psychologischer Konstrukte. Im Falle eines Raumvorstellungstests beispielsweise beinhaltet die Konstruktvalidierung die Prüfung, inwieweit die Ergebnisse tatsächlich im Sinne hoher oder niedriger Raumvorstellungsfähigkeit interpretiert werden können und nicht etwa Indikatoren

anderer Fähigkeiten (wie differenzierte Wahrnehmung oder schlußfolgerndes Denken) sind. Eine Konstruktvalidierung ist in der Regel ein langwieriger, im Prinzip oft nicht abschließbarer Prozeß. Die Konstruktvalidität läßt sich in der Regel nicht in Form eines einzelnen Koeffizienten darstellen. Vielfach wird sie nicht allein in quantitativer Weise (z.B. durch Korrelationen des Tests mit anderen Variablen, Faktorenladungen o.ä.), sondern auch in qualitativer Form (z.B. durch den Nachweis, daß vermutete kognitive Prozesse bei der Lösung von Testaufgaben tatsächlich ablaufen) beschrieben. Eine Konstruktvalidierung beginnt meist mit dem Anspruch eines Testautors, mit seinem Verfahren ein bestimmtes Merkmal zu erfassen. Das Merkmal wird durch ein theoretisches Konzept beschrieben, das seinerseits in ein System von Begriffen und Aussagen, die diese Begriffe (Konstrukte) erläutern sowie miteinander und mit empirischen Beobachtungen verbinden, eingebettet ist. Ein solches System braucht keine vollständig entwickelte Theorie zu sein, es kann sich dabei um eine skizzenhafte Vorstufe zu einer Theorie handeln, die man mit Cronbach eine „Konstruktion" nennt. So ist beispielsweise das Konstrukt der Studieneignung für das Fach Humanmedizin in ein System von Konzepten der Studieneignung für die verschiedensten Fächer eingebettet. In dieser Konstruktion wird z.B. davon ausgegangen, daß die Eignung für ein Medizinstudium positiv mit der Eignung für ein naturwissenschaftliches Studienfach korreliert und daß Personen mit hohen Ausprägungen dieser Merkmale ceteris paribus ein höheres Maß an Studienerfolg erwarten lassen als Personen mit niedrigen Ausprägungen.

Im allgemeinen bringt ein Testautor zu Beginn der Konstruktvalidierung einige Befunde dazu bei, daß die Ergebnisse in seinem Test tatsächlich im Sinne des entsprechenden Konstrukts interpretiert werden können, etwa dadurch, daß er Untersuchungsergebnisse zur Konvergenz verschiedener Indikatoren für das zugehörige Merkmal anführt. Die Validierung nimmt dadurch ihren Fortgang, daß *plausible Alternativhypothesen* (Cronbach, 1971) zu der ursprünglich vorgeschlagenen Interpretation aufgestellt und überprüft werden, z.B. die, daß die Indikatorenwerte für ein bestimmtes Merkmal mit anderen Merkmalen konfundiert sind oder daß in bestimmten Fällen ein anderes Konstrukt das Testergebnis besser erklärt, daß z.B. die Antworten in einem Persönlichkeitsfragebogen bei bestimmten Personen nichts anderes widerspiegeln als soziale Erwünschtheit. Damit wird ein Forschungsprozeß eingeleitet, der im allgemeinen zumindest zu einer Präzisierung der ursprünglichen Interpretation des Tests führt, möglicherweise aber auch zur Aufgabe des Tests, des zugehörigen Konstrukts oder der gesamten Konstruktion.

Analysen, die für eine Konstruktvalidierung relevant sein können, lassen sich einteilen in *logisch-inhaltliche*, *experimentelle* und *korrelationsstatistische* (Cronbach, 1971). Strenggenommen können logisch-inhaltliche Überlegungen den Validitätsanspruch nicht immer direkt widerlegen, sie können aber zu plausiblen Alternativhypothesen führen, die die Interpretation von Testergebnissen in Frage stellen, wie im obigen Beispiel.

Das Feld experimenteller Untersuchungen zur Konstruktvalidität von Tests ist sehr weit. Es umfaßt beispielsweise kognitionspsychologische Untersuchungen zu den bei der Bearbeitung von Aufgaben ablaufenden Lösungsprozessen (s. z.B. Klieme, 1989) oder Ansätze, Testergebnisse experimentell zu beeinflussen, wie dies etwa in Untersuchungen zur Bearbeitung von Persönlichkeitsfragebögen unter Täuschungs-

instruktionen geschieht (s. z.B. Stumpf, Angleitner, Wieck, Jackson & Beloch-Till, 1985, S. 75 ff.).

Von den *korrelationsstatistischen* Verfahren zur Konstruktvalidierung sei hier das *Multitrait-Multimethod* (MTMM-) Modell nach Campbell und Fiske (1959) erwähnt. Dieser Ansatz fordert für jeden Test den Nachweis *konvergenter* (hoher Korrelationen verschiedener Indikatoren für ein Merkmal) und *diskriminanter* Validität (niedriger Korrelationen von Indikatoren für verschiedene Merkmale). Das MTMM-Modell geht insofern über den ursprünglichen Ansatz der klassischen Testtheorie hinaus, als es neben der unsystematischen Meßfehlervarianz grundsätzlich *zwei* Arten der systematischen Varianz von Testwerten unterscheidet, die sich dem Modell zufolge additiv verhalten, nämlich merkmalsspezifische (valide) und methodenspezifische.

Die Anwendung des MTMM-Modells setzt voraus, daß für mehrere Merkmale Indikatorenwerte aus mehreren möglichst unähnlichen Meßmethoden (z.B. Persönlichkeitsfragebögen, Selbsteinschätzungen, Fremdeinschätzungen) vorliegen; die Interkorrelations- und Reliabilitätskoeffizienten der Indikatorenwerte werden zur *Multitrait-Multimethod-Matrix* zusammengestellt. Die Validitätsbeurteilung erfolgte anfangs durch systematische Vergleiche der Korrelationen (Campbell & Fiske, 1959). Heute verwendet man aber zur Analyse von MTMM-Daten meist weiterentwickelte Verfahren (s. die Übersicht bei Schmitt & Stults, 1986), wie varianz- oder faktorenanalytische Techniken, insbesondere konfirmatorische Faktorenanalysen (s. z.B. Jöreskog, 1974; Marsh, 1989). Nicht selten ergeben sich bei konfirmatorischen Faktorenanalysen von MTMM-Matrizen aber Probleme der Eindeutigkeit und Schätzbarkeit von Parametern (z.B. Marsh, 1989; Grayson & Marsh, 1994).

Anders als der Begriff der Reliabilität hat das Validitätskonzept im Laufe seiner Geschichte eine Reihe grundlegender Revisionen erfahren (s. z.B. Anastasi, 1986; Angoff, 1988). Eine der wichtigsten Neuerungen in letzter Zeit ist der Vorschlag, bei der Validierung eines Tests nicht nur psychologische Fragen zu behandeln, sondern auch die politischen und ökonomischen Konsequenzen des Testeinsatzes zu prüfen (Cronbach, 1988b; Messick, 1988).

2.4 Testfairneß

Das Gütekriterium der Fairneß eines Tests wurde im Rahmen der Diskussion um *Test-* und *Item-Bias* in die Psychometrie eingeführt. Anlaß für diese Innovation war die Beobachtung, daß in den USA Personen aus bestimmten ethnischen Minoritäten in den großen nationalen Testprogrammen bisweilen ungünstig abschneiden. Obwohl die Testfairneß eng mit der Validität zusammenhängt (s. z.B. Shepard, 1982), wird sie inzwischen als selbständiges Kriterium zur Beurteilung von Tests und Testitems angesehen (Möbus, 1983). Unter Unfairneß oder *bias* versteht man in der Psychometrie einen konstanten oder systematischen Fehler bei der Schätzung eines Werts oder Kriteriumsmaßes (Reynolds, 1982, p.199). Der Bezug auf einen zu schätzenden Wert oder ein Kriterium ist wesentlich für die Definition von *bias*; Mittelwertsunterschiede zwischen verschiedenen Gruppen alleine stellen nicht notwendigerweise Unfairneß dar (Reynolds, 1982; Angoff, 1993). In diesem Punkt unterscheidet sich das psychometrische Konzept der Fairneß wesentlich von landläufigen Begriffen der Chancengleichheit.

Bias kann auf der Ebene von Testwerten (s. besonders Jensen, 1980; Reynolds, 1982) und auf der Ebene einzelner Items (s. besonders Holland & Wainer, 1993) auftreten.

Die Diskussion von *bias* auf der Ebene von Testwerten orientiert sich an den verschiedenen Aspekten des Validitätskonzepts, insbesondere an den Begriffen der prognostischen und Konstruktvalidität. In bezug auf die prognostische Validität versteht man unter *bias* einen systematischen Fehler in der Vorhersage von Werten in einem Kriteriumsmaß. In diesem Zusammenhang gilt ein Test als voreingenommen, wenn die Schlußfolgerungen, die aus dem Testergebnis gezogen werden, nicht für jede Gruppe von Probanden dem kleinstmöglichen Zufallsfehler unterliegen oder wenn die Schlußfolgerungen oder Vorhersagen anhand des Testergebnisses für eine Gruppe von Probanden einem konstanten Fehler unterliegen (Cleary, 1968; Cleary, Humphreys, Kendrick & Wesman, 1975; Reynolds, 1982). Ein typischer Fall von *bias* in diesem Sinne liegt etwa dann vor, wenn in einem Test Männer im Durchschnitt besser abschneiden als Frauen, aber die Leistungen der Geschlechtsgruppen im Kriterium nicht differieren. Verwendet man in diesem Fall für Männer und Frauen dieselbe Regressionsgleichung zur Vorhersage des Kriteriums, so wird das Abschneiden der Männer im Kriteriumsmaß zu günstig prognostiziert. Eine typische Analyse zur Frage von Test-*Bias* in diesem Sinn beschreiben Stumpf und Nauels (1988).

Test-*Bias* wird natürlich dann praktisch relevant, wenn an die Ergebnisse im Test Entscheidungen geknüpft werden. Würde der Test im oben genannten Beispiel zur Auslese von Bewerbern verwendet, so würden Frauen benachteiligt. Es ist jedoch wichtig, zwischen Test- und Selektions-*Bias* zu unterscheiden. Test-*Bias* ist eine Eigenschaft des Tests, wie die zu günstige Vorhersage der Kriteriumsleistungen von Männern. Selektions-*Bias* bezieht sich auf die Entscheidungen, die aus den Testergebnissen abgeleitet werden, wie die überproportional häufige Ablehnung von Frauen. Nicht selten treten beide Formen des *bias* zusammen auf. Es ist aber im Prinzip möglich, trotz Test-*Bias* Selektionsfairneß zu erreichen, wenn man den im Test benachteiligten Gruppen einen Bonus gewährt. Selektionsfairneß oder Selektions-*Bias* ist nicht allein eine psychometrische Frage, sondern in erster Linie eine sozialpolitische. Es kommt daher nicht von ungefähr, daß in der Literatur sehr viele verschiedene Modelle zur Sicherung von Selektionsfairneß vorgeschlagen wurden (s. z.B. Petersen & Novick, 1976; Bartussek, 1982; Möbus, 1983). Die Unterschiedlichkeit dieser Modelle beruht nicht zuletzt darauf, daß sie von verschiedenen moralischen Grundpositionen ausgehen (Hunter & Schmidt, 1976).

In bezug auf die Konstruktvalidität liegt *bias* vor, wenn ein Test in einer Gruppe von Personen andere latente Merkmale oder Konstrukte mißt als in einer anderen oder wenn der Test in zwei Gruppen dasselbe Merkmal mit unterschiedlichem Grad an Genauigkeit mißt. Es gibt zahlreiche Methoden zur Aufdeckung von Konstrukt-*Bias* (vgl. die Übersicht bei Reynolds, 1982).

Auch bei der Beurteilung von Item-*Bias* wird das zu untersuchende psychometrische Instrument, in diesem Fall die Testaufgabe, in Beziehung zu einem Wert oder Kriterium gesetzt. Idealerweise handelt es sich bei diesem Bezugswert um ein unvoreingenommenes Maß des zu erfassenden Konstrukts, aber in der Praxis ist man meist auf den Gesamtscore im jeweiligen Test als Schätzung der Ausprägung des latenten Merkmals angewiesen. Üblicherweise vergleicht man das Abschneiden von

zwei Gruppen von Personen im Gesamttest und in den einzelnen Items. Es gibt zahlreiche statistische Verfahren zur Aufdeckung von *bias* (Holland & Wainer, 1993), von denen eine Reihe allerdings nicht auf dem Modell der klassischen Testtheorie aufbauen.

Von den Ansätzen, die im Rahmen des klassischen Modells angewendet werden, können hier nur die Mantel-Haenszel-Methode (Holland & Thayer, 1988; Dorans & Holland, 1993) und die Standardisierungsmethode (Dorans & Kulick, 1986; Dorans & Holland, 1993) erwähnt werden. Beide Methoden vergleichen das Abschneiden zweier Gruppen in einem Item, tragen dabei aber Unterschieden, die im Testgesamtwert bestehen, Rechnung. Sie betrachten somit Unterschiede im jeweiligen Item, die nicht durch die Leistungen der Gruppen im Gesamttest erklärt werden können. Analysiert man mit diesen Methoden einen Test, so werden im allgemeinen einige Aufgaben identifiziert, in denen die Gruppen mehr voneinander abweichen, als der Gesamttest es vermuten läßt. Ob dabei wirklich eine Benachteiligung einer Gruppe vorliegt, ist vielfach nicht ausgemacht oder von dem Zweck abhängig, zu dem der Test eingesetzt wird.

Aus diesen Gründen hat es sich in der Fairneßdiskussion eingebürgert, nicht generell von Item-*Bias* zu sprechen, sondern von *differential item functioning* (DIF), wenn eine Aufgabe etwa im Sinne des Mantel-Haenszel- oder des Standardisierungsverfahrens auffällig ist. Ob es sich um Item-*Bias* handelt, ist darüber hinaus noch von logischen und inhaltlichen Überlegungen abhängig (Shepard, 1982).

3 Weiterführende Literatur

Eine ausführliche Behandlung der Grundannahmen der klassischen Testtheorie findet der Leser bei Novick (1966) und Lord und Novick (1968). Von den zahlreichen lehrbuchartigen Darstellungen des Themas sei hier die Arbeit von Allen und Yen (1979) genannt. Die wichtigste Abhandlung zum Konzept der Reliabilität ist nach wie vor das Kapitel von Stanley (1971), wenngleich seither auf diesem Gebiet etliche Neuerungen eingeführt wurden (vgl. die oben zitierte Literatur). Neuere Arbeiten zum Validitätsbegriff findet der Leser in dem von Wainer und Braun (1988) herausgegebenen Sammelband. Die Einarbeitung in das MTMM-Modell erleichtert die bei Schmitt und Stults (1986) sowie Grayson und Marsh (1994) zitierte Literatur. Wichtige Abhandlungen zum Fairneßkonzept sind bei Jensen (1980), Berk (1982) sowie Holland und Wainer (1993) zusammengestellt. Eine Darstellung der wichtigsten psychometrischen Methoden zur Testkonstruktion und -evaluation geben beispielsweise Lienert und Raatz (1994).

Literaturverzeichnis

Alexander, R. A., Carson, K. P., Alliger, G. M. & Barrett, G. V. (1984). Correction for restriction of range when both X and Y are truncated. *Applied Psychological Measurement, 8*, 231–241.

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory.* Belmont: Wadsworth.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (1985). *Standards for educational and psychological testing.* Washington, D.C.: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, D.C.: American Council on Education.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19–32). Hillsdale: Erlbaum.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale: Erlbaum.

Bartussek, D. (1982). Modelle der Testfairneß und Selektionsfairneß. *Trierer Psychologische Berichte, 9*, Heft 2.

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Blixt, S. & Shama, D. (1986). An empirical investigation of the standard error of measurement at different ability levels. *Educational and Psychological Measurement, 46*, 545–550.

Braun, H. I. & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Brinberg, D. J. & McGrath, J. E. (1982). A network of validity concepts within the research process. In D. J. Brinberg & L. H. Kidder (Eds.), *Forms of validity in research.* (pp. 5–21). San Francisco: Jossey-Bass.

Bührer, M. & Kohr, H. U. (1984). *ASIA - Automatische Skalenbildung für itemanalytische Anwendungen.* München: Sozialwissenschaftliches Institut der Bundeswehr.

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

Cattell, J. McKeen (1890). Mental tests and measurements. *Mind, 15*, 373–380.

Cattell, R. B. (1958). What is „objective" in „objective personality tests ?" *Journal of Counseling Psychology, 5*, 285–289.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A. & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30*, 15–41.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443–507). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1988a). Internal consistency of tests: Analyses old and new. *Psychometrika, 53*, 63–70.

Cronbach, L. J. (1988b). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Erlbaum.

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

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale: Erlbaum.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Edmundson, E. W., Koch, W. B. & Silverman, S. (1993). A facet analysis approach to content and construct validity. *Educational and Psychological Measurement, 53*, 351–368.

Fechner, G. T. (1860). *Elemente der Psychophysik.* Leipzig: Breitkopf & Härtel.

Feger, B. (1984). Die Generierung von Testitems zu Lehrtexten. *Diagnostica, 30*, 24–46.

Feldt, L. S., Steffen, M. & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9*, 351–361.

Feldt, L. S., Woodruff, D. J. & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93–103.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen.* Bern: Huber.

Fleishman, J. & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47*, 925–939.

Galton, F. (1883). *Inquiries into human faculty and its development.* London: MacMillan.

Grayson, D. & Marsh, H. W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis: Multitrait-multimethod models. *Psychometrika, 59*, 121–134.

Green, S. B., Lissitz, R. W. & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 73*, 827–838.

Guion, R. M. (1965). *Personnel testing.* New York: McGraw Hill.

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

Häcker, H. (1982). Objektive Tests zur Messung der Persönlichkeit. In K. J. Groffmann & L. Michel (Hrsg.), *Persönlichkeitsdiagnostik* (= Enzyklopädie der Psychologie, Themenbereich B, Serie II, Band 3, S. 132–185). Göttingen: Hogrefe.

Hanges, P. J., Rentsch, J. R., Yusko, K. P. & Alexander, R. A. (1991). Determining the appropriate correction when the type of range restriction is unknown: Developing a sample-based procedure. *Educational and Psychological Measurement, 51*, 329–340.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Heise, D. (1969). Separating reliability and stability in test-retest correlations. *American Sociological Review, 34*, 93–101.

Holland, P. W. & Rubin, D. B. (1982). *Test equating.* New York: Academic Press.

Holland, P. W. & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale: Erlbaum.

Holland, P. W. & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale: Erlbaum.

Hunter, J. E. & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin, 83*, 1053–1071.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz & R. D. Suppes (Eds.), *Contemporary developments in mathematical psychology*, Vol. II (pp. 1–54). San Francisco: Freeman.

Kaufman, A. S. (1972). Restriction of range: Questions and answers. *Test Service Bulletin, 59*, 2–12.

Klauer, K. J. (1978). Perspektiven der pädagogischen Diagnostik. In K. J. Klauer (Hrsg.), *Handbuch der pädagogischen Diagnostik* (Band 1, S. 3–14) . Düsseldorf: Schwann.

Klauer, K. J. (1983). Kriteriumsorientierte Tests. In H. Feger & J. Bredenkamp (Hrsg.), *Messen und Testen* (= Enzyklopädie der Psychologie, Themenbereich B, Serie I, Band 3, S. 693–726). Göttingen: Hogrefe.

Klauer, K. J. (1984). Kontentvalidität. *Diagnostica, 30*, 1–23.

Klieme, E. (1989). *Mathematisches Problemlösen als Testleistung.* Frankfurt: Lang.

Kristof, W. (1983). Klassische Testtheorie und Testkonstruktion. In H. Feger & J. Bredenkamp (Hrsg.), *Messen und Testen* (= Enzyklopädie der Psychologie, Themenbereich B, Serie I, Band 3, S. 544–603). Göttingen: Hogrefe.

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lienert, G. A. (1969). *Testaufbau und Testanalyse.* Weinheim: Beltz.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Beltz.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Erlbaum.

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

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251–280.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod matrices: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335–361.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Erlbaum.

Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K. J. Groffmann & L. Michel (Hrsg.) *Grundlagen psychologischer Diagnostik* (= Enzyklopädie der Psychologie, Themenbereich B, Serie 2, Band 1, S. 1–129). Göttingen: Hogrefe.

Möbus, C. (1983). Die praktische Bedeutung der Testfairneß als zusätzliches Kriterium zu Reliabilität und Validität. In R. Horn, K. Ingenkamp & R. S. Jäger (Hrsg.), *Tests und Trends. 3. Jahrbuch der pädagogischen Diagnostik* (S. 155–203). Weinheim: Beltz.

Novick, M. R. (1966). The axioms and principle results of classical test theory. *Journal of Mathematical Psychology, 3*, 1–18.

Petersen, N. S. & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29.

Piedmont, R. L. & Hyland, M. E. (1993). Inter-item correlation frequency analysis: A method for evaluating scale dimensionality. *Educational and Psychological Measurement, 53*, 369–378.

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

Rasmussen, J. L. (1988). Evaluation of small-sample statistics that test whether variables measure the same trait. *Applied Psychological Measurement, 12*, 177–187.

Reuterberg, S. E. & Gustafsson, J. E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement, 52*, 795–811.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199–227). Baltimore: Johns Hopkins University Press.

Roid, G. & Haladyna, T. (1982). *A technology for test item writing.* New York: Academic Press.

Schott, F. & Wieberg, H. J. W. (1984). Regelgeleitete Itemkonstruktion. *Diagnostica, 30*, 47–66.

Serlin, R. C. & Kaiser, H. F. (1978). A method for increasing the reliability of a short multiple-choice test. *Educational and Psychological Measurement, 38*, 337–340.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore: Johns Hopkins University Press.

Schmitt, N. & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1–22.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 356–442). Washington, D.C.: American Council on Education.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N. & Beloch-Till, H. (1985). Deutsche Personality Research Form (PRF). Göttingen: Hogrefe.

Stumpf, H. & Nauels, H. U. (1988). Untersuchungen zur prognostischen Validität des TMS. In G. Trost (Hrsg.), *Test für medizinische Studiengänge. 12. Arbeitsbericht* (S. 94–217). Bonn: Institut für Test- und Begabungsforschung.

Überla, K. (1968). *Faktorenanalyse.* Berlin: Springer.

Ulrich, R. (1985). Die Beziehung zwischen Testlänge und Validität für nicht-parallele Aufgaben: Verschiedene Methoden der Validitätsmaximierung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 6*, 32–45.

Wainer, H. & Braun, H. I. (Eds.) (1988). *Test validity.* Hillsdale: Erlbaum.

Wiggins, J. S. (1973). *Personality and prediction.* Reading: Addison-Wesley.

Woodruff, D. J. & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika, 51*, 393–413.

Wundt, W. (1862). *Beiträge zur Theorie der Sinneswahrnehmung.* Leipzig: Winter.