Markus Bühner

Einführung in die Test- und Fragebogenkonstruktion

2., aktualisierte und erweiterte Auflage

ein Imprint von Pearson Education
München • Boston • San Francisco • Harlow, England
Don Mills, Ontario • Sydney • Mexico City
Madrid • Amsterdam

Testtheoretische Grundlagen

2

ÜBERBLICK

2.1 Gegenstand einer Testtheorie

Über welche theoretischen Grundlagen sollte ich verfügen?

Nach Rost (2004, S. 17 ff.) gibt es in der sozialwissenschaftlichen Methodenlehre zwei Begriffe von Testtheorie. Der erste bezeichnet eine Theorie über statistische Schlüsse, die man aufgrund von Stichprobendaten bezüglich bestimmter Eigenschaften der Population (oder Grundgesamtheit) zieht. Ein solcher statistischer Schluss wird als Test bezeichnet, weil man damit eine **Hypothese** testet oder jemanden einer Prüfung unterzieht. Dieser Begriff von Testtheorie wird im Folgenden nicht verwendet.

Der zweite Begriff bezeichnet die **Theorie über „psychologische Tests“**. Testtheorien in diesem Sinne sind die Klassische und die Probabilistische Testtheorie. Psychologische Tests erfassen psychische Eigenschaften, Fähigkeiten oder Merkmale bzw. Zustände von Personen, die abstrakt auch als psychologische Konstrukte bezeichnet werden können. Unter Tests im weiteren Sinne versteht man auch Fragebögen, standardisierte Interviews und standardisierte Beobachtungen. Tests in einem engeren Sinne bezeichnen Verfahren, die durch die getestete Person nicht willentlich in eine gewünschte Richtung verfälscht werden können. Diese Definition ist insofern nicht unproblematisch, da Personen immer in der Lage sind, Testergebnisse willentlich zu beeinflussen. Dies gilt vor allem für Fragebögen, Interviews oder Beobachtungen, z.B. durch sozial erwünschtes Verhalten oder sozial erwünschte Beantwortung. Ergebnisse von vermeintlich objektiven Tests, wie Intelligenztests oder allgemeinen Leistungstests, sind vom Probanden über die Motivation beeinflussbar. Daher muss eine Definition diesen Fakten Rechnung tragen. Gegenstand der Testtheorie sind also „objektive“ Leistungstests sowie Daten von Fragebögen, Beobachtungen und Interviews. „Objektiv“ bedeutet hier im Zusammenhang mit Leistungstests lediglich, dass Antwortalternativen eindeutig als richtig oder falsch ausgewiesen werden können. Die Frage, ob Testverfälschung durch Personen möglich ist, wird aus dieser Definition ausgeklammert.

An dieser Stelle sei auch auf die enge Verknüpfung der Testtheorie mit der Diagnostik hingewiesen, da testtheoretische Grundlagen und deren Methoden als Hilfsmittel der diagnostischen Begutachtung dienen.

> Testtheorien befassen sich entweder mit dem Zusammenhang von Testverhalten und dem zu erfassenden psychischen Merkmal (Rost, 2004, S. 21) oder mit der Frage, in welche Bestandteile sich Messwerte aufgliedern.

Theorien und Testverhalten Theorien sollen nach Rost (2004, S. 29 ff.) vor allem erklären und nicht nur beschreiben. Wichtige Punkte sind die Beschreibung und, wenn möglich, Erklärung von Wirklichkeit durch Theorien sowie die Prognose zukünftigen Verhaltens. Abstrakt ausgedrückt erklärt ein Testmodell systematische Zusammenhänge zwischen den Antworten oder Reaktionen der Personen bezüglich der verschiedenen Aufgaben oder Fragen dadurch, dass latente Personenvariablen (latent = verborgen, nicht sichtbar) eingeführt werden. Aufgaben oder Fragen werden im Rahmen der Testtheorie auch unter dem Begriff **Item** zusammengefasst. Man bezeichnet Items auch als **beobachtbare** oder **manifeste Variablen**. Zwischen den **Itemantworten** wird es in aller Regel bestimmte systematische **Zusammenhänge** (auch als „Korrelation“ oder „Kontingenz“ bezeichnet) geben. Das heißt, Personen, die bei

dem Item „Ich bin traurig“ die Antwortalternative „trifft zu“ ankreuzen, werden auch überzufällig oft bei einem anderen Item mit ähnlichem Iteminhalt („Ich fühle mich niedergeschlagen“) entweder dieselbe oder eine ähnliche Einstufung vornehmen.

Um diese systematischen Zusammenhänge (Korrelation oder Kontingenz) zwischen Items (manifesten Variablen) zu erklären, werden eine oder mehrere **latente Variablen** konstruiert. Auf diese latenten Variablen kann dann das Antwortverhalten der Testitems zurückgeführt werden. Latente Variablen werden auch als Konstrukte bezeichnet, weil sie im Rahmen der Theorienbildung konstruiert worden sind. Eines der bekanntesten psychologischen Konstrukte ist „Intelligenz“. Im Gegensatz zum Begriff des „Konstruktes“, das keine speziellen mess- oder testtheoretischen Eigenschaften impliziert, ist mit der „latenten Variable“ in der Regel gemeint, dass es sich um genau eine Variable handelt. Es wird also angenommen, dass eine (zunächst unbekannte) **latente Variable** für das Zustandekommen der Antworten bei bestimmten Items „verantwortlich“ ist und daher deren beobachtbare Zusammenhänge „produziert“. Wenn diese „Erklärung“ richtig ist, so müssten die Zusammenhänge zwischen den Items „verschwinden“ (Nullkorrelation), wenn man die latente Variable „ausschaltet“, also beispielsweise konstant hält, was auch als „auspartialisieren“ bezeichnet wird. Genau das wird in den meisten Testmodellen vorausgesetzt: *Wenn ein bestimmtes Testmodell mit latenten Variablen gelten soll, dürfen die Itemantworten bei „festgehaltener“ latenter Variable untereinander keine Zusammenhänge mehr aufweisen.* Man bezeichnet dies als **lokale Unabhängigkeit**, worauf in *Kapitel 7* „Probabilistische Testtheorie“ noch näher eingegangen wird. In *Abbildung 2.1* wird der beschriebene Gedankengang nochmals veranschaulicht.

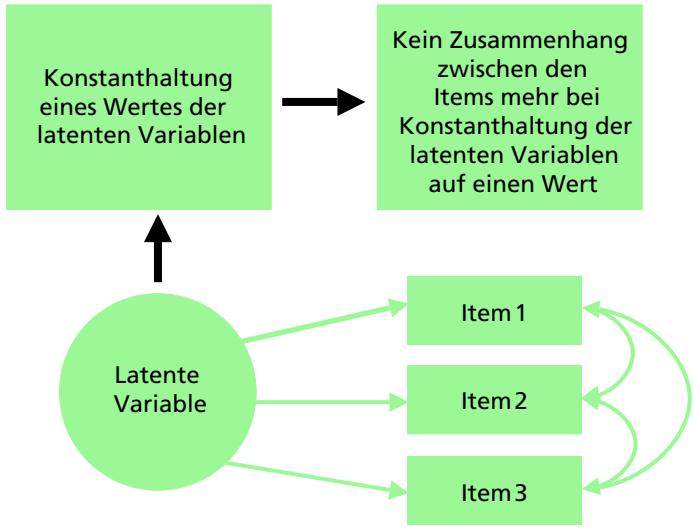

Abbildung 2.1: Lokale Unabhängigkeit

Eine andere Weise, sich mit Tests zu befassen, besteht darin, sich über die Bestandteile eines Messwerts Gedanken zu machen. Alltägliche Beispiele weisen uns immer wieder darauf hin, dass keine Messung wirklich perfekt ist. So sind beispielsweise beim Einbau einer Küche immer wieder Nacharbeiten nötig, Platten müssen verkürzt oder Leisten verlängert werden, da die Maße für einige Teile ungenau gemessen wur-

den. Im Falle von Konstrukten ist der Messfehler, der einem bei der Messung unterlaufen kann, noch viel größer als beim Messen mit einem Metermaß. Die so genannte Klassische Testtheorie beschäftigt sich mit Messungen und deren Ungenauigkeit. Im Gegensatz dazu wird in der Probabilistischen Testtheorie neben anderen Annahmen direkt die Annahme der Eindimensionalität geprüft. Mit anderen Worten heißt das, es wird getestet, ob die Antworten zu verschiedenen Items wirklich auf nur eine Eigenschaft oder Fähigkeit zurückgeführt werden können.

Z U S A M M E N F A S S U N G

Gegenstand der Testtheorie sind „objektive" Leistungstests sowie Daten von Fragebögen, Beobachtungen und Interviews. Sie dienen dazu, Verhalten zu beschreiben und zu erklären.

Man unterscheidet zwischen der **Klassischen** und der **Probabilistischen Testtheorie**.

- Die **Probabilistische Testtheorie** beschäftigt sich mit der Frage, wie das Testverhalten einer oder mehrerer Personen von einem zu erfassenden psychischen Merkmal abhängt.
- Die **Klassische Testtheorie** befasst sich mit den unterschiedlichen Bestandteilen von Messwerten (wahrer Wert + Messfehler).

Systematische Zusammenhänge zwischen den beobachtbaren (manifesten) Items eines Tests werden durch so genannte **latente Variablen** erklärt, auf die wiederum das Antwortverhalten einer oder mehrerer Personen zurückgeführt wird. Die Items sind dabei beobachtbare Indikatoren dieser latenten Variablen.

Lokale Unabhängigkeit: Hält man die zu messende Eigenschaft oder Fähigkeit auf einer Stufe der latenten Variable konstant (z.B. Betrachtung von Personen mit gleichem beobachteten IQ), dürfen die Items untereinander keinen Zusammenhang mehr aufweisen. Davon abzugrenzen ist die lokale stochastische Unabhängigkeit, auf die in *Kapitel 7* eingegangen wird.

Z U S A M M E N F A S S U N G

2.2 Eine Test-Definition

Was versteht man unter einem psychometrischen Test?

Für die Testkonstruktion sollte eine Theorie über den Messgegenstand bzw. genaue Überlegungen zur Erfassung des angestrebten Messgegenstandes vorliegen. Das interessierende Merkmal sollte definiert und spezifiziert sein. Beispielsweise könnte man definieren, was unter dem Begriff „Konzentration" verstanden werden soll und wodurch er gekennzeichnet ist. Dabei sollte das zu erfassende Merkmal möglichst hoch mit einem objektiven Indikator des betreffenden Merkmales korrelieren (ein Aspekt, den man als Validität bezeichnet). Bleibt man beim Beispiel der Konzentration, sollte Konzentrationsfähigkeit mit Fahrtauglichkeit bedeutsam zusammenhängen. Leider fehlt psychologischen Testverfahren oftmals eine theoretische Fundierung.

Definition „psychometrischer Test“ (Lienert & Raatz 1998, S. 1)

Ein psychometrischer Test ist ein wissenschaftliches **Routineverfahren** zur Untersuchung eines oder mehrerer empirisch abgrenzbarer **Persönlichkeitsmerkmale** mit dem Ziel einer möglichst **quantitativen Aussage** über den relativen Grad der **individuellen Merkmalsausprägung**. Rost (2004) erweitert diese Definition mit dem Hinweis, dass es nicht immer um eine quantitative Aussage geht, sondern das Ziel eines Tests auch eine **qualitative Aussage** sein kann (z.B. Zuordnung von Personen zu bestimmten Kategorien).

2.3 Kennzeichen psychometrischer Tests

Woran erkenne ich einen psychometrischen Test?

Psychometrische Tests haben den Anspruch, **normiert**, **objektiv**, **strukturiert** und **zulänglich** sowie nach der **Klassischen** oder der **Probabilistischen Testtheorie** konstruiert zu sein.

Im Folgenden sollen die oben genannten einzelnen Anforderungen kurz beschrieben werden.

Ein psychometrischer Test ist in der Regel **normiert**. Normen sind für die Testauswertung unverzichtbar. So reicht es zur Bestimmung der Intelligenz nicht aus, einfach die Summe der richtig gelösten Items zu ermitteln, um den IQ zu bestimmen. Erst die Einordnung der individuellen Leistung in eine Referenzgruppe erlaubt einen Rückschluss auf den IQ. Man möchte also die Leistung eines Probanden im Vergleich zu anderen Probanden erfahren. Dabei ist von Interesse, ob die Leistung als unterdurchschnittlich, durchschnittlich oder überdurchschnittlich gegenüber einer Normgruppe bezeichnet werden kann.

Ein psychometrischer Test sollte bei der Durchführung und der Auswertung **objektiv** sein. Das heißt, die Beurteilung der Leistung einer Person variiert nicht aufgrund der Person des Testleiters. Dazu müssen die Durchführungsbedingungen beschrieben sein. Auswertung und Interpretation des Tests sollen standardisiert sein, damit jeder Untersucher die gleiche Testleistung für ein und denselben Probanden ermittelt und diese auch gleich interpretiert.

Es werden zwei Arten von **Strukturiertheit** unterschieden: Itemstrukturiertheit und Antwortstrukturiertheit. Itemstrukturiertheit heißt, der Test gibt eine klare und eindeutige Aufgabenstellung vor. Unter Antwortstrukturiertheit versteht man, dass die Antworten/Antwortmöglichkeiten vorgegeben sind oder/und nur eine Antwort richtig ist bzw. die richtige Antwort genau festgelegt werden kann. Erst dieses Vorgehen ermöglicht eine „objektive“ Auswertung. Bei Persönlichkeitstests gibt es in der Regel keine richtige Antwort. Das heißt, ein Proband gibt in einem Persönlichkeitstest beispielsweise an, dass er sehr „offen für neue Erfahrungen“ sei, indem er „starke Zustimmung“ auf einer vorgegebenen vierstufigen Antwortskala von „starker Ableh-

nung“ bis hin zu „starker Zustimmung“ ankreuzt. Diese Antwort ist natürlich auch objektiv auswertbar, indem man die in diesem Falle höchste Ausprägung mit der höchsten Punktzahl bewertet. Dies impliziert nicht, dass die Antwort auch wirklich in dem Sinne objektiv ist, dass sie der Proband nicht verfälscht bzw. andere Personen übereinstimmend der Meinung sind, dass die Person in starkem Maße offen für neue Erfahrungen ist.

Unter **Zulänglichkeit** (nach Lienert & Raatz, 1998) versteht man den Grad dessen, was an „Gemeinsamkeit“ durch einen Test und einem oder mehreren Außenkriterien erfasst wird. Dieser Aspekt bezieht sich auf ein ganz bestimmtes Gütemerkmal (Validität), auf das in diesem Kapitel noch eingegangen wird. Für eine erste Näherung ist damit gemeint, ob der Test auch zur Vorhersage eines bestimmten Kriteriums (z.B. Intelligenztest auf Schulerfolg) herangezogen werden kann.

Schließlich basiert jeder psychometrische Test auf der **Klassischen** oder **Probabilistischen Testtheorie**.

2.4 Klassische Testtheorie

Was muss ich über die klassische Testtheorie wissen?

Bevor nun die Grundannahmen der Klassischen Testtheorie dargestellt werden, möchten wir anhand eines Beispiels aus dem Sport einige Überlegungen anstellen, die zu einem besseren Verständnis für die Nützlichkeit einer Testtheorie führen sollen.

Beispiel 2.1

Nehmen wir als Beispiel einen Hochsprungwettbewerb. Hochspringer haben für jede Höhe, die ihnen vorgegeben wird, drei Sprungversuche. Dabei wird es mit zunehmender Höhe schwieriger, diese zu überspringen. Das heißt, den Hochspringern werden unterschiedlich schwere Höhen (Items) vorgegeben, um ihre Leistungsfähigkeit zu messen. Es gewinnt der Hochspringer, der das schwierigste Item löst und damit am höchsten springt. Allerdings wird ein Hochspringer nicht in jedem Wettkampf die gleiche Höhe erzielen. Bei einem Wettkampf wird er vielleicht 2.10 m überspringen und bei einem anderen 2.15 m. **Seine erzielten Leistungen werden also in einem bestimmten Bereich schwanken.**

Wenn wir annehmen, dass die Wettkampfbedingungen konstant sind, sollten die Einzelleistungen eines Hochspringers einer bestimmten individuellen Verteilung (z.B. glockenförmige Normalverteilung, siehe dazu Bortz, 1999, S. 79) folgen. Das heißt, Leistungen, die am ehesten seiner Fähigkeit entsprechen („wahre“ Leistungsfähigkeit) werden bei wiederholten Sprüngen häufiger vorkommen, extrem schlechte oder gute Leistungen seltener. Unter Konstanz der Wettkampfbedingungen versteht man, dass sich diese Bedingungen während des Wettkampfs nicht durch plötzlich auftretenden Regen, stärkeren Gegenwind oder Ähnliches verändern. Die Konstanz mag innerhalb eines Wettkampfs manchmal noch realisierbar sein, in zeitlich aufeinander folgenden Wettkämpfen ist jedoch die Bedingungskonstanz selten gegeben. Die erzielten Höhen werden sich aber selbst unter perfekter Konstanz der Wettkampfbedingungen unterscheiden, da auch Faktoren, die in der Person liegen, variieren können, wie zum Beispiel Müdigkeit oder Motivation.

Das heißt, selbst unter optimalen Bedingungen ist es sehr unwahrscheinlich, dass Personen immer die gleiche Leistung erzielen. Es ist also sinnvoll, mehrere Messungen vorzunehmen, um die Leistung eines Hochspringers zu ermitteln. In der Regel hat jeder Hochspringer für jede Höhe drei Versuche. Die Leistung wird aber nicht nur über einen Wettkampf zusammengefasst, sondern auch über verschiedene Wettkämpfe.

Angenommen, man würde die Hochsprungleistung nicht durch verschiedene Durchgänge mit unterschiedlichen Höhen messen, sondern nach einem einzigen Sprung ermitteln. Zieht man nur diesen Einzelwert zur Beurteilung der Leistungsfähigkeit eines Hochspringers heran, so könnte dies zu Fehlschlüssen führen. In einem solchen Fall ist es möglich, dass ein Hochspringer mit einer „tatsächlich" niedrigeren Leistungsfähigkeit einen Hochspringer mit einer „tatsächlich" höheren Leistungsfähigkeit in einem Wettkampf besiegt. Dieses Ergebnis kann man damit erklären, dass dem Hochspringer mit geringerer Leistungsfähigkeit ein extrem guter Sprung (Item mit hoher Schwierigkeit) gelungen ist und dem Hochspringer mit der höheren Leistungsfähigkeit nur ein extrem schlechter Sprung (Item mit geringer Schwierigkeit). Die Wahrscheinlichkeit für ein solches Wettkampfergebnis ist gering, da man annehmen muss, dass ein Hochspringer mit einer hohen Fähigkeit auch mit einer höheren Wahrscheinlichkeit eine bessere Höhe erzielt als ein Hochspringer mit einer niedrigen Fähigkeit. Dieses Beispiel zeigt, dass es sinnvoll ist, mehrere Sprünge (Items) mit unterschiedlichen Höhen (Schwierigkeiten) heranzuziehen, um die „wahre" Leistungsfähigkeit eines Springers zu beurteilen.

Folgende Schlüsse lassen sich aus dem beschriebenen Beispiel ziehen:

- Auch wenn die Bedingungen für jeden Sprung (konstante Wettkampfbedingungen) gleich sind, werden die Leistungen eines Hochspringers variieren. Grund dafür sind nicht kontrollierbare Einflüsse. Daher kann die Kenntnis eines einzelnen Wertes zu falschen Schlussfolgerungen führen.
- Es ist notwendig, die Leistung einer Person über mehrere Messgelegenheiten zu erheben. Diese Messgelegenheiten können mehrere Sprungversuche (Items) oder aber auch mehrere Wettkämpfe (Testwiederholungen) sein.
- Ob sich Personen in ihrer (Hochsprung-)Fähigkeit unterscheiden, erkennt man, indem man ihnen Items (Sprungversuche) unterschiedlicher Schwierigkeit (unterschiedliche Höhen) vorgibt: beispielsweise 2.10 Meter und 2.15 Meter.
- Eine gute Schätzung für die Leistungsfähigkeit von Personen erhält man, wenn man die Leistung von Personen über unterschiedlich schwere Items mittelt, die Summenleistung bestimmt oder Treffer zählt.

2.4.1 Grundannahmen der Klassischen Testtheorie

Was sind die Grundannahmen der Klassischen Testtheorie?

Die Klassische Testtheorie ist gegenwärtig die Grundlage der meisten psychologischen Testverfahren. Nach Rost (1999, S. 140) basieren 95 Prozent aller Tests auf der Klassischen Testtheorie. Damit ist es schon aus rein pragmatischen Gesichtspunkten notwendig, sich mit dieser Theorie auseinander zu setzen. Klassisch heißt sie deshalb,

weil sie die erste Theorie war, die zur Konstruktion von psychologischen Tests herangezogen wurde. Im Laufe der Zeit kam die Probabilistische Testtheorie hinzu, mit der es gelungen ist, Schwächen der Klassischen Testtheorie zu überwinden.

Ein großer Vorteil der Klassischen Testtheorie liegt in ihrer einfachen Anwendbarkeit (Henard, 2000). Zudem haben sich Tests, die nach der Klassischen Testtheorie konzipiert wurden, bewährt. Darin liegt wahrscheinlich auch ihr Erfolg bis heute. Im Folgenden werden die Grundannahmen der Klassischen Testtheorie einfach und ohne formale Herleitung oder Einbettung geschildert. Sehr umfassend ist die Klassische Testtheorie bei Steyer und Eid (2001) dargestellt. Der fortgeschrittene Leser sei zur Vertiefung der Klassischen Testtheorie ausdrücklich auf dieses Lehrbuch verwiesen.

Die Klassische Testtheorie trägt dem Umstand Rechnung, dass Testergebnisse einzelner Personen mit dem gleichen Test zwischen verschiedenen Messzeitpunkten variieren. Steyer und Eid (2001, S. 102) nennen dafür unterschiedliche Gründe. Übertragen auf unser obiges Beispiel kann zum einen die Fähigkeitsausprägung „hoch zu springen“ durch ein besonderes Aufbautraining (**Übungs- und Transfereffekte**) verbessert werden. Zum anderen ist es möglich, dass die Messung der Fähigkeit durch **unsystematische äußere Einflüsse,** wie Wind und Regen, oder **unsystematische innere Einflüsse,** wie Müdigkeit oder mangelnde Motivation, zufällig schwanken (Messung ist fehlerbehaftet). Diese Einflussfaktoren treten meist in Kombination auf. Der Fehlerbegriff in der Klassischen Testtheorie berücksichtigt jedoch nur **unsystematische Fehler**. Darüber hinaus erfolgen keine Annahmen darüber, wie Items beantwortet werden oder wie eine Testleistung zustande kommt (Fischer, 1974, S. 124), sondern nur, aus welchen **Komponenten** Messwerte bestehen. Die Klassische Testtheorie ist eine reine **Messfehlertheorie**.

Novick (1966) geht davon aus, dass (1) die getestete Person zufällig aus einer Population entnommen wurde und (2) das Testergebnis einer Person zufallsabhängig variiert (z.B. aufgrund von Müdigkeit). (3) Eine Person kann zu verschiedenen Zeitpunkten getestet werden und erhält dabei jeweils unterschiedliche Werte. Aus der Verteilung dieser verschiedenen Werte wird zufällig ein Wert beobachtet. Mit welcher Wahrscheinlichkeit die Werte einer Person beobachtet werden, wird durch die intraindividuelle Verteilung der beobachteten Werte einer Person festgelegt. Diese Verteilung ist eine hypothetische Verteilung, die sich ergeben würde, wenn man eine Person unendlich oft unter denselben Bedinungen testen würde. Nimmt man dann für die beobachteten Werte eine Normalverteilung an, haben die Werte um den Mittelwert der Verteilung die höchste Wahrscheinlichkeit, beobachtet zu werden, und extreme Werte eine geringere Wahrscheinlichkeit. Dabei ist der **wahre Wert** definiert als der Mittelwert (Erwartungswert) dieser hypothetischen intraindividuellen Verteilung. Diese Aussage wird auch als Existenzaxiom bezeichnet (Moosbrugger & Hartig, 2003). Der Mittelwert der beobachteten Werte wird demnach kaum ein ganzzahliger Wert sein. Das heißt, dass es relativ unwahrscheinlich ist, den wahren Wert zu beobachten. Nehmen wir als Beispiel einen IQ-Wert. Eine Person wird immer einen ganzzahligen IQ-Wert in einem Test erzielen (z.B. 110), auch wenn der wahre Wert vielleicht nicht ganzzahlig ist (z.B. 109.5). Somit ist nach Krauth (1995) bei der Veranschaulichung des wahren Werts Vorsicht geboten (Möglichkeiten, den wahren Wert zu schätzen, werden in *Abschnitt 4.8* dargestellt). Steyer und Eid (2001) lehnen den Begriff „wahrer Wert“ ab und verstehen den Mittelwert über unendlich viele Testwiederholungen als Verhaltenstendenz einer Person in einer konkreten Situation. Zusätzlich sei darauf hingewiesen, dass (4) die beobachteten Messwerte in einem begrenzten Bereich

streuen (das heißt, endliche Varianzen besitzen; vgl. Steyer & Eid, 2001, S. 104). Unter diesen Voraussetzungen wird die folgende Grundannahme der Klassischen Testtheorie formuliert, die auch als Verknüpfungsaxiom bezeichnet wird (Moosbrugger & Hartig, 2003): Der **beobachtete Messwert (*X*)** einer Person in einem Test setzt sich aus dem **konstanten wahren Wert (*T*)** einer Person und einem **Messfehler (*E*)** zusammen:

$$X = T + E$$

Der Messfehler (E) setzt sich aus der Differenz zwischen beobachtetem Testwert (X) einer Person und wahrem Wert (T) einer Person zusammen und repräsentiert alle unkontrollierten und unsystematischen Störeinflüsse (vgl. Amelang & Zielinski, 2002, S. 34):

$$E = X - T$$

Aus dieser Festsetzung bzw. Definition von wahrem Wert und Messfehler ergeben sich Folgerungen für die **Eigenschaften des Messfehlers** und des **Zusammenhangs zwischen Messfehler und wahrem Wert**. Die beiden Folgerungen oder Ableitungen können nach Steyer und Eid (2001, S. 103) in der empirischen Anwendung nicht falsch sein und sind aus logischen Gründen wahr.

Die erste Folgerung besagt, dass der **Mittelwert (*M*)** des Messfehlers (E)[1] über unendlich viele Messungen einer **Person (*I*)** bzw. einer Messung einer beliebigen **Population oder Teilpopulation (*P*)** null ist. Dies muss so sein, da der Mittelwert über alle beobachteten Werte per Definition dem wahren Wert entspricht. Daher muss der gemittelte Messfehler über unendlich viele Messungen „0“ entsprechen:

$$M(E)_I = 0 \tag{1a}$$

und

$$M(E)P = 0 \tag{1b}$$

Die zweite Folgerung besagt, dass kein **Zusammenhang (*r*)** zwischen dem Messfehler (E) und dem wahren Wert (T) einer Person oder in einer Population oder Teilpopulation besteht:

$$r(E, T) = 0 \tag{2}$$

Die erste Annahme (1a) bedeutet beispielsweise, dass, wenn man einen Hochspringer unendlich viele Sprünge unter identischen Bedingungen machen lässt, sich der Messfehler ausmittelt und null ergibt. Dies kennzeichnet nach Rost (2004, S. 36) auch einen Messfehler im Gegensatz zu einem systematischen Fehler oder „Bias“ (z.B. bei unendlich vielen Sprüngen herrscht immer Rückenwind), der sich nicht ausmitteln würde. Unter diesen Voraussetzungen entspricht der Mittelwert einer Person über verschiedene Messungen dem wahren Wert einer Person. Auch wenn man unendlich viele Hochspringer unter identischen Bedingungen nur einmal springen lässt (Annahme 1b), mittelt sich der Messfehler aus und ergibt null (vgl. Amelang & Zie-

1 Es handelt sich bei dem Mittelwert des Messfehlers und der wahren Werte um Erwartungswerte. Unter einem Erwartungswert versteht man den Mittelwert einer theoretischen (nicht empirischen) Verteilung.

linski, 2002). Das heißt, der Mittelwert entspricht der wahren Leistungsfähigkeit der Population. Die zweite Annahme besagt, dass kein Zusammenhang zwischen dem Messfehler (E) und dem wahren Wert (T) einer Person, einer Population oder Teilpopulation besteht. Der Messfehler nimmt also mit abnehmender oder zunehmender wahrer Fähigkeit der Hochspringer weder ab noch zu.

Weiterhin wurden folgende zusätzliche Annahmen formuliert: Der Messfehler (E_A) eines Tests A (z.B. Gedächtnistest) weist keinen Zusammenhang (r) mit dem Messfehler (E_B) eines anderen Tests B (z.B. Leistungsmotivationstest) auf:

$$r\,(E_A,\, E_B) = 0 \tag{3}$$

Diese Annahme gilt nach Kristof (1983, S. 547) nur dann, wenn beide Messvorgänge (Tests) experimentell unabhängig sind. Das heißt, für einen beliebigen Probanden darf der beobachtete Punktwert im Test A nicht den Punktwert in Test B beeinflussen. Steyer und Eid (2001, S. 104) bemerken, dass die Annahme eines nicht vorhandenen Zusammenhangs von Messfehlern in der Praxis falsch sein kann. Die Annahme sei für mathematische Ableitungen zwar bequem, aber nicht zwingend notwendig.

Eine weitere Annahme ist, dass die Messfehler (E_A) eines Tests A keinen Zusammenhang (r) mit dem tatsächlichen Wert (T_B) aus einem Test B aufweisen:

$$r\,(E_A,\, T_B) = 0 \tag{4}$$

Nimmt man die Definition $X = T + E$ ernst, ergibt sich daraus die Implikation, dass die Klassische Testtheorie nur für solche Messwerte definiert ist, für die die Berechnung von Differenzen sinnvoll ist. Dies ist nicht für alle Messungen der Fall. Für kategoriale Daten ergeben Differenzen beispielsweise keinen Sinn. Man könnte sich Häufigkeiten vorstellen, die in die Kategorien „Äpfel“ und „Birnen“ fallen. Niemand würde auf die Idee kommen, Birnen von Äpfeln abzuziehen (5 Äpfel – 2 Birnen = ?). Auch für ordinale Messwerte ist die Klassische Testtheorie nicht definiert. Ein Beispiel für ordinale Daten ist eine Platzierung beim Hochsprung. Würde man nur wissen, welchen Rang ein Hochspringer belegt hat, könnte der Erstplatzierte klar gewonnen haben oder auch nur denkbar knapp. Man kann es anhand der Information nicht entscheiden. Erst mit der Intervallskala (siehe *Abschnitt 3.3*), ist es sinnvoll, nach Differenzen zu fragen. Zieht man die Höhen des Erstplatzierten von denen des Zweitplatzierten ab, erhält man eine Ahnung davon, wie knapp oder wie souverän ein Hochspringer einen Wettkampf gewonnen hat. In der Regel vertraut man darauf, dass bei der Anzahl richtiger Lösungen oder der Summe von Itemantworten generell Differenzen sinnvoll interpretierbar sind. Dies wird auch als „Per fiat“-(„Es möge sein“-)Messung bezeichnet.

Durch die aufgeführte Definition und die dargestellten Annahmen lässt sich ein Kernkonzept der Klassischen Testtheorie herleiten, die **Reliabilität** (r_{tt}) oder Messgenauigkeit eines Tests (siehe *Kapitel 4*). Das Konzept der Reliabilität ist nicht nur auf Testverfahren beschränkt, sondern auch auf andere Methoden anwendbar, z.B. Interviews oder Beobachtungsverfahren. Um das Konzept der Reliabilität besser verstehen zu können, soll zunächst das Konzept der Varianz eingeführt werden.

Was heißt Varianz?

Konstruiert man einen Test, möchte man natürlich, dass sich Probanden, die diesen Test bearbeiten, in ihrer Testleistung unterscheiden. Testet man also mit einem Leistungsmotivationsfragebogen 100 Personen, werden diese bei geeigneter Itemauswahl und wenn sie sich tatsächlich in ihrer Leistungsmotivation unterscheiden, unterschiedliche Werte in diesem Test erzielen. Man kann nun den Mittelwert (M) der 100 Personen in diesem Test bestimmen und nachsehen, wie stark die Testwerte der einzelnen Probanden von diesem abweichen. Das heißt, es wird berechnet, wie weit oder eng sich die Personenmesswerte um den Mittelwert gruppieren. Summiert man die quadrierten Abweichungen über alle Probanden auf und teilt sie durch die Anzahl der Probanden, erhält man die Varianz eines Tests (S^2). Zieht man die Wurzel aus der Varianz, resultiert daraus die Standardabweichung. Die Unterschiede in den Messwerten der Personen sollen nun systematisch sein und auf ihre tatsächliche bzw. wahre Leistungsfähigkeit zurückzuführen sein. Dies ist aber – wie bereits erwähnt – eine sehr idealistische Sichtweise. Vielmehr sind die Unterschiede der Probanden sowohl durch ihre tatsächliche Leistungsfähigkeit als auch durch unsystematische Fehler bedingt.

Drückt man die Unterschiedlichkeit der Personen in einem Test durch die Varianz der beobachteten Messwerte (S_X^2) aus, so lässt sich diese Varianz in die Varianz der wahren Werte (S_T^2) und in die Fehlervarianz (S_E^2) aufteilen. Unter Verwendung dieser Varianzanteile erhält man folgende Formel für die Schätzung der Reliabilität eines Tests (r_{tt}):

$$r_{tt} = \frac{S_T^2}{S_T^2 + S_E^2}$$

$$r_{tt} = \frac{S_T^2}{S_X^2}$$

Aus den bis hierhin angeführten Annahmen der Klassischen Testtheorie ist damit die Reliabilität als Varianzverhältnis zwischen wahren und beobachteten Werten (wahrer Wert + Fehler) definiert und spiegelt den Anteil an wahrer Varianz der Personenmesswerte wider (Fischer, 1974, S. 37; Steyer & Eid, 2001, S. 104). Wie man leicht nachvollziehen kann, ist die Varianz der wahren Werte nicht bestimmbar. Daher müssen Wege gefunden werden, dieses Varianzverhältnis zu schätzen. Es geht, wie schon im oberen Abschnitt erwähnt, um die Reproduzierbarkeit von Antworten und Leistungen unter identischen Bedingungen. Ein Test misst ein Merkmal dann genau, wenn der Testwert von Personen bei wiederholter Messung im Idealfall immer gleich ausfällt, und/oder wenn die Rangreihe der Personen zwischen erster und zweiter Messung gleich ausfällt. Das heißt, dass die Person mit dem höchsten Wert bei der ersten Messung auch bei der zweiten Messung den höchsten Wert erzielt. Dies gilt dann sukzessive für alle weiteren Personen. Da aber in der Praxis meist nur eine Messung üblich ist, behilft man sich zur Bestimmung der Messgenauigkeit mit anderen Methoden. Eine Möglichkeit besteht darin, einen Test in zwei oder beliebig viele Testteile oder sogar in Einzelitems aufzuteilen. Eine hohe Messgenauigkeit liegt dann vor, wenn eine Person die verschiedenen Fragen oder Items in diesen verschiedenen Testteilen oder Einzelitems immer in der gleichen Art und Weise (konsistent) beantwortet oder löst.

Wie kann man diese Überlegungen nun in Schätzungen der Reliabilität umsetzen?

Zum einen kann man einer Person mehrere Fragen oder Aufgaben (Items) vorgeben, die das gleiche Konstrukt (Fähigkeit oder Eigenschaft) messen. Um die Reliabilität zu bestimmen, wird dann beispielsweise der mittlere Zusammenhang zwischen allen Items einer Skala bestimmt und darüber hinaus berücksichtigt, dass mehrere Messungen zu einer genaueren Schätzung des wahren Wertes führen. Man könnte auch zwei Tests mit unterschiedlichen Items konstruieren, die das gleiche Konstrukt messen, und dann den Zusammenhang zwischen beiden Tests ermitteln. Das ist aus verschiedenen Gründen sehr schwer. Daher kann man, wie bereits erwähnt, alternativ einen Test in zwei Hälften aufteilen und zwischen beiden Hälften den Zusammenhang bestimmen. Eine letzte Möglichkeit besteht darin, einen Test zweimal in einem gewissen Abstand vorzugeben und den Zusammenhang zwischen erster und zweiter Testung zu bestimmen. Auf die hier vorgeschlagenen Berechnungsmöglichkeiten der Reliabilität und ihre Schwierigkeiten soll in *Kapitel 4* „Reliabilität“ noch genauer eingegangen werden.

Z U S A M M E N F A S S U N G

Testergebnisse einzelner Personen im gleichen Test können zwischen verschiedenen Testzeitpunkten variieren. Das Ergebnis kann z.B. durch **systematische Einflüsse**, wie **Übungs- und Transfereffekte**, verbessert oder durch **unsystematische Einflüsse**, wie Müdigkeit und externe Störungen (z.B. Rückenwind), verschlechtert oder verbessert werden.

In den beobachteten Wert gehen diese systematischen und unsystematischen Einflüsse ein. Der **beobachtete Messwert (*X*)** einer Person in einem Test setzt sich zusammen aus einem **konstanten wahren Wert (*T*)** einer Person und einem **Messfehler (*E*), der sich in der klassischen Testtheorie aber nur auf unsystematische Einflüsse bezieht**. Aus dieser Definition ergeben sich zwei Folgerungen: (1) Der **wahre Wert** ist dabei als Mittelwert über unendlich viele „beobachtete Testwerte“ einer Person unter gleichen Bedingungen definiert. (2) Der Messfehler (E) hingegen enthält alle unkontrollierten und unsystematischen Störeinflüsse. Der Mittelwert des Messfehlers (M_E) über unendlich viele Messungen einer Person (I) oder einer Messung einer beliebigen Population oder Teilpopulation (P) ist null. Folgende zusätzliche Annahmen werden im Rahmen der Klassischen Testtheorie gemacht: (1) Es besteht kein Zusammenhang zwischen dem Messfehler (E) und dem wahren Wert (T) einer Person, einer Population oder Teilpopulation. (2) Der Messfehler (E_A) eines Tests A weist keinen Zusammenhang mit dem Messfehler (E_B) eines anderen Tests B auf. Schließlich (3) hängt der Messfehler (E_A) eines Tests A nicht mit dem wahren Wert (T_B) eines Tests B zusammen. Die Klassische Testtheorie ist nur für Messwerte definiert, die mindestens Intervallskalenniveau erreichen. Liegt **Intervallskalenniveau** vor, lassen sich Differenzen zwischen den Messwerten sinnvoll interpretieren.

Unter **Reliabilität** versteht man die Messgenauigkeit eines Tests. Man kann sie auf folgende Art und Weise schätzen: (1) Berechnung des mittleren Zusammenhangs aller Items einer Skala unter Berücksichtigung der Testlänge (innere Konsistenz). (2) Ermittlung des Zusammenhangs zwischen zwei parallelen Tests mit unterschiedlichen Items, die das gleiche Konstrukt messen (Paralleltestreliabilität). (3) Vorgabe desselben Tests in zeitlichem Abstand und Bestimmung des Zusammenhangs zwischen erster und zweiter Testung (Retest-Reliabilität).

Berechnet man den **Mittelwert (*M*)** der Testwerte aller Personen in einem Test, kann man zusätzlich bestimmen, inwieweit sich die einzelnen Testwerte der Personen von diesem unterscheiden, d.h. wie weit oder wie eng sich die Messwerte um den Mittelwert gruppieren. Die Summe der quadrierten Abweichungen über alle Personen wird dabei als **Varianz (S^2)** bezeichnet. Zieht man aus der Varianz die Wurzel, erhält man die **Standardabweichung (*S*)** eines Items oder Testwertes. Die Reliabilität eines Tests ist definiert als Varianzverhältnis zwischen wahren und beobachteten Werten:

$$r_{tt} = \frac{S_T^2}{S_T^2 + S_E^2}$$

Z U S A M M E N F A S S U N G

2.4.2 Kritische Anmerkungen zur Klassischen Testtheorie

Was muss ich über Schwächen der Klassischen Testtheorie wissen?

Die Klassische Testtheorie stellt keine Verbindung zwischen einer Fähigkeit, einem Merkmal oder einer Eigenschaft und der Itembeantwortung her. Die Klassische Testtheorie ist eine reine **Messfehlertheorie** und beschäftigt sich nur mit den Komponenten eines beobachteten beliebigen **Messwerts**. Borsboom und Mellenbergh (2002) geben ein anschauliches Beispiel dafür, dass (wahre) Werte im Rahmen der Klassischen Testtheorie nicht für Fähigkeiten stehen müssen. Nehmen wir einen Test mit folgenden Items: „Ich wäre gerne ein militärischer Befehlshaber“ und „Ich bin über 1.80 m groß“. Nun wird zu beiden Itemantworten konsistent eine Zufallszahl addiert. Danach wird der Summenwert mit der Anzahl der Buchstaben des Vornamens der Person multipliziert. Das ergibt dann den Messwert. Über verschiedene Wiederholungsmessungen hat jede Person einen „wahren“ Wert in diesem Test. Möglicherweise hängen sogar die Testwerte aufeinander folgender Messungen in hohem Maße zusammen. Allerdings sind die Testwerte nicht als Fähigkeit oder Eigenschaft zu interpretieren.

Wie bereits geschildert, sind einige Annahmen der Klassischen Testtheorie nicht überprüfbar, sondern ergeben sich logisch aus der Festsetzung des beobachteten Wertes als wahrer Wert plus Messfehler. Auch wenn die Klassische Testtheorie mathematisch durchaus befriedigend formuliert ist (vgl. Fischer, 1974), sind manche der Annahmen in der psychologischen Praxis nur schwer haltbar (vgl. Fischer, 1974, S. 26). Wie Fischer (1974, S. 28) richtig bemerkt, können nicht alle Einflüsse auf das Testergebnis als Zufallseinflüsse abgetan werden. Übungs- und Transfereffekte wirken sich unter Umständen systematisch auf die Testleistung aus und verändern die wahre Leistungsfähigkeit einer Person. In letzter Konsequenz ist damit sowohl die Annahme eines fehlenden Zusammenhangs zwischen wahrem Wert und Messfehler zu bezweifeln als auch die Konstanz des wahren Wertes über verschiedene Messwiederholungen. Stumpf (1996, S. 415) merkt dazu an, dass bei einem Persönlichkeitstest neben Persönlichkeitseigenschaften – je nach Situation – auch soziale Erwünschtheit eine Rolle spielen könnte. Damit werden solche systematischen Varianzanteile dem wahren Wert zugerechnet. Der wahre Wert setzt sich also aus dem Persönlichkeitsmerkmal plus sozialer Erwünschtheit zusammen. Aus diesen Beispielen wird ersichtlich, dass innerhalb der Klassischen Testtheorie keine Annahmen hinsichtlich des Zustandekommens der Leistungen erfolgen, und unter Verletzung der Annahme von Eindimensionalität (Test oder Skala misst nur ein Konstrukt) sowohl der wahre Wert einer

Person als auch die Messgenauigkeit eines Tests über- oder in manchen Fällen auch unterschätzt werden. Man muss sich also behelfen. Die konfirmatorische Faktorenanalyse (siehe *Kapitel 6*) ist beispielsweise eine Methode, mit der man prüfen kann, ob Eindimensionalität vorliegt oder nicht.

Fischer (1974, S. 144) verweist zurecht darauf, dass gerade bei extrem hohen und niedrigen Fähigkeitsausprägungen Leistungen ungenauer als im mittleren Bereich gemessen werden können (vgl. auch Bortz & Döring, 2002, S. 555 f., „Regression zur Mitte"). Dieselbe Messgenauigkeit in allen Eigenschafts- bzw. Fähigkeitsbereichen stellt aber z.B. für die psychologische Einzelfalldiagnostik (siehe *Kapitel 4.8*) eine wichtige Voraussetzung dar.

Weitere Annahmen der Klassischen Testtheorie sind zum Teil nicht zwingend oder auch widerlegbar, wie zum Beispiel die Annahme des fehlenden Zusammenhanges zwischen verschiedenen Fehlerwerten (vgl. Steyer & Eid, 2001, S. 104). Die Autoren weisen darauf hin, dass eine solche Annahme zwar bequem, aber keinesfalls zwingend ist. Treten keine korrelierten Fehler zwischen Messwerten auf, liegt Eindimensionalität (Ausnahmen siehe *Kapitel 7*) vor. Hier wird deutlich, dass die Klassische Testtheorie streng genommen Eindimensionalität annimmt, diese Annahme aber an keiner Stelle überprüft. Eine Methode, um solche Verletzungen aufzuzeigen, bietet die Verwendung von konfirmatorischen Faktorenanalysen (siehe *Kapitel 6*).

Ein weiteres Problem mit großem Gewicht ist sicherlich, dass die Testwerte der Klassischen Testtheorie stichprobenabhängig sind. Das heißt, für Abiturienten mag ein Intelligenztest andere Testkennwerte (z.B. Schwierigkeiten) besitzen als für Hauptschüler und Realschüler. Zu berücksichtigen ist auch die Tatsache, dass die Werte einer Person in verschiedenen Tests, die nach der Klassischen Testtheorie konstruiert wurden und dasselbe Konstrukt messen sollen, nicht direkt vergleichbar sind. Das heißt, ein Summenwert von 20 gelösten Items kann in einem Test eine gute und in einem anderen Test eine schlechte Leistung bedeuten. Auch innerhalb eines Tests kann dieselbe Leistung unterschiedlich interpretiert werden. Vergleicht man einen Probanden zum Beispiel mit einer Normgruppe von Hauptschülern, ergibt sich vielleicht ein IQ von 115. Mit derselben Anzahl an gelösten Items würde im Vergleich mit Gymnasiasten jedoch nur ein IQ von 100 festgestellt werden. Mit anderen Worten bedeutet das, dass sich je nach Test und Referenzgruppe völlig andere Bedeutungen der individuellen Leistung ergeben können. Hier weist die Probabilistische Testtheorie einen großen Vorzug auf, denn im Rahmen einzelner probabilistischer Modelle ist es möglich, stichprobenunabhängige Item- und Personenkennwerte zu ermitteln. Allerdings ist die Itemkonstruktion für einen solchen Test aufwändig (sorgfältige Itemkonstruktion ist dabei grundsätzlich von Vorteil), und häufig genügen die Items nicht den strengen Modellkriterien. Bei Testverfahren, die nach der Klassischen Testtheorie konzipiert sind, behilft man sich damit, dass man Gütekriterien für alle Teilstichproben zur Verfügung stellt. Allerdings wird dies nicht immer von Testautoren bis zur letzten Konsequenz verfolgt, was in der Praxis einen großen Mangel darstellt.

Die oben genannten Punkte zu Unzulänglichkeiten der Klassischen Testtheorie, die noch problemlos erweitert werden könnten (vgl. Amelang und Zielinski 2002, S. 62 f.), sind grundsätzlich berechtigt und schwerwiegend. Insgesamt könnte nun der Eindruck entstehen, dass die Klassische Testtheorie so unzulänglich ist, dass sie in der Praxis nicht eingesetzt werden sollte. Dennoch ist die Klassische Testtheorie nach Stumpf (1996, S. 416) nicht unbrauchbar, sondern hat sich in der Praxis bewährt (vgl.

auch Amelang & Zielinski, 2002). Dies mag daran liegen, dass die Brauchbarkeit eines Tests vor allem von einer inhaltlich begründeten Konstruktion der Items und der Skalen abhängt. Im Folgenden soll kurz und überblicksartig auf wenige Grundüberlegungen des Rasch-Modells eingangen werden und danach auf die Haupt- und Nebengütekriterien psychologischer Tests.

Kurzer Ausblick auf die Probabilistische Testtheorie

Die Grundlagen der Probabilistischen Testtheorie werden in *Kapitel 7* beschrieben. Daher wird an dieser Stelle nur ein ganz kurzer Ausblick gegeben. In der Probabilistischen Testtheorie geht es im Gegensatz zur Klassischen Testtheorie darum, wie Antworten auf Items zustande kommen. Aus diesem Grund werden Antwortmuster untersucht. Die beobachteten Antwortmuster müssen einem bestimmten Modell folgen. Dieses Modell sagt voraus, dass mit steigender Personenfähigkeit die **Wahrscheinlichkeit** einer Itemlösung zunimmt. Die **Lösungswahrscheinlichkeit** für ein bestimmtes Item hängt (1) von der **Fähigkeit** oder **Eigenschaftsausprägung** einer Person sowie (2) der **Schwierigkeit** eines Items ab. Diese **Beziehung** zwischen Personenfähigkeit und Itemlösungswahrscheinlichkeit ist **probabilistisch**. Das heißt, auch eine Person mit geringer Fähigkeit im Vergleich zur Schwierigkeit eines Items hat eine, wenn auch geringe, Wahrscheinlichkeit, ein solches Item zu lösen. Im Rahmen der Probabilistischen Testtheorie kann ein Modelltest durchgeführt werden. Die folgenden Ausführungen beziehen sich auf das Rasch-Modell, einem Modell aus der Familie der Probabilistischen Testmodelle. Wird das Modell durch den Modelltest nicht abgelehnt, sagt der **Summenwert** der Itemantworten auch wirklich etwas über den **Ausprägungsgrad einer Person** auf der latenten Variable (Fähigkeit) aus. Dann ist der Summenwert auch eine **erschöpfende Statistik** der Personenfähigkeit. Erschöpfend heißt, der Summenwert einer Person liefert alle Informationen über die Fähigkeitsausprägung der Person. Demnach muss das Antwortmuster der Person nicht mehr Item für Item betrachtet werden. Ein Item ist dann ein guter Indikator für eine latente Variable, wenn die Leistung in diesem Item komplett auf die Fähigkeitsausprägung auf der latenten Variable zurückzuführen ist und nicht auf andere Fähigkeiten. Dies ist eine höchst wünschenswerte Annahme für die Testkonstruktion, da sie eine sehr präzise Definition von Itemhomogenität darstellt (vgl. Stelzl, 1993). Formalisiert wird diese Eigenschaft durch die **lokale stochastische Unabhängigkeit**. Wenn das Rasch-Modell durch den Modelltest nicht verworfen wird, liegt auch diese Eigenschaft vor. Das Rasch-Modell implementiert damit eine echte Messtheorie in die Psychologie. Weitere Ausführungen zum Rasch-Modell finden sich in *Kapitel 7*.

2.5 Haupt- und Nebengütekriterien

Was macht einen guten psychometrischen Test aus?

Es gibt verschiedene anerkannte Kriterien, nach denen die Güte eines Tests beurteilt werden kann. Nur wenn die nachfolgend beschriebenen Gütekriterien vollständig im Testhandbuch aufgeführt sind, kann ein Test in seiner Güte beurteilt werden. Hierbei unterteilt man Haupt- und Nebengütekriterien (Lienert & Raatz, 1998; Amelang & Zielinski, 2002). Zu den Hauptgütekriterien gehören Objektivität, Reliabilität und Validität. Alle diese Begriffe lassen sich weiter differenzieren. Im Hinblick auf die

Testauswahl und Testbeurteilung ist ein sicherer Umgang mit diesen Begriffen unumgänglich.

Objektivität
- Durchführung
- Auswertung
- Interpretation

Reliabilität
- Konsistenz
- Retest-Reliabilität
- Paralleltestreliabilität

Validität
- Inhaltsvalidität
- Konstruktvalidität
- Kriteriumsvalidität

2.5.1 Hauptgütekriterien

Was sind die wichtigsten Indikatoren für einen guten psychometrischen Test?

Objektivität Unter Objektivität versteht man den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind. Man unterscheidet drei Arten von Objektivität:

- *Durchführungsobjektivität*
  Die Durchführung eines Tests darf nicht von Untersuchung zu Untersuchung variieren. Dazu muss genau definiert sein, wie und unter welchen Bedingungen ein Test oder Fragebogen durchzuführen ist. Zeitbegrenzung oder Hilfestellungen bei der Beantwortung der Fragen müssen vorgegeben werden. Die größte Sorgfalt sollte auf die Instruktion verwendet werden, denn dadurch können Rückfragen an den Untersucher minimiert werden. Dieser läuft wiederum nicht Gefahr, den Personen unterschiedliche Hilfestellungen zu geben. Häufig muss auch festgelegt werden, welche Ausschlusskriterien zu berücksichtigen sind. Nehmen wir beispielsweise den Test d2 (Brickenkamp, 2002). Bei diesem Test ist die Aufgabe der Versuchsperson, den Buchstaben „d“ mit zwei Strichen (oben oder unten zwei Striche oder oben und unten jeweils einen Strich) unter den Buchstaben „p“ und „d“ mit unterschiedlicher Stricheanzahl zu markieren. Hält sich nun der Testleiter nicht an die genaue Testinstruktion, beeinträchtigt dies unter Umständen das Testergebnis, und die wahre Leistungsfähigkeit des Probanden wird über- oder unterschätzt. Dies ist beispielsweise dann wahrscheinlich, wenn der Testleiter bei einer Testung folgende Instruktion gibt: „Es kommt darauf an, sorgfältig und schnell zu arbeiten“ und in der anderen Testung: „Es kommt darauf an, nur schnell zu arbeiten“.

- *Auswertungsobjektivität*
  Jeder Auswerter muss die gleichen Punkt- oder Leistungswerte eines Probanden ermitteln. Dazu sind genaue Auswertungsvorschriften nötig. Hilfreich sind Schablonen und Auswertungsblätter, die die für die Auswertung relevanten Daten ent-

halten. Aber auch Schablonen garantieren nicht immer eine ausreichende Auswertungsobjektivität. So kann das Auflegen von Schablonen selbst wieder fehlerträchtig sein. Bleiben wir bei dem Beispiel des Tests d2. Hier müssen über 600 Zeichen mit Hilfe der Schablonen überprüft werden. Würden zwei Auswerter unabhängig voneinander denselben Test auswerten, ist nicht gesagt, dass jeder Auswerter auch alle falsch durchgestrichenen Zeichen entdeckt. Die Auswertungsobjektivität sollte überprüft werden. Dazu sind eine Reihe von Indizes geeignet (Cohen´s Kappa, Scott´s Pi oder die Intraklassenkorrelation; vgl. Wirtz & Caspar, 2002). Die lapidare Anmerkung, dass die Auswertungsobjektivität durch das Bereitstellen von Schablonen zur Auswertung vorliegt, ist nicht ausreichend, um eine hohe Auswertungsobjektiviät zu belegen. Die Auswertungsobjektivität hängt nicht zuletzt von der Art und Weise, wie gefragt wird, ab. Wird die Frage offen, ohne festgelegte Antwortmöglichkeiten, gestellt, muss sehr exakt definiert werden, was als „richtig" zu bewerten ist. Bei manchen Tests ist es notwendig, Regeln zu definieren, ab wann ein Ergebnis zu werten ist und ab wann nicht. Auch dies lässt sich am Beispiel von Test d2 zeigen. In Test d2 lässt sich ein so genanntes Ü-Syndrom diagnostizieren. Ein Ü-Syndrom liegt dann vor, wenn die Mengenleistung (Gesamtzahl bearbeiteter Zeichen, GZ) über einem Prozentrang von PR = 90 liegt und die Sorgfaltsleistung (Fehlerprozentwert, F%) unter PR = 10. Ist dies der Fall, liegt der Verdacht nahe, dass die Testleistung instruktionswidrig zustande gekommen ist. In einem solchen Fall sollte das Ergebnis mit Vorsicht interpretiert werden. In manchen Fällen kann es vorkommen, dass ein Ergebnis gar nicht interpretiert werden kann.

- *Interpretationsobjektivität*

  Jeder Auswerter sollte möglichst zur gleichen Beurteilung oder Interpretation der Testergebnisse kommen. Interpretationsobjektivität schließt ausreichend große Normstichproben und ausreichend geprüfte Gütekriterien mit ein, so dass man davon ausgehen kann, dass jede Person mit dem gleichen Maßstab beurteilt wird. Allerdings ist dies alleine nicht ausreichend. Häufig fehlen standardisierte Interpretationen. Manchmal wird dies damit entschuldigt, dass man den Testleiter nicht an ein bestimmtes Schema binden möchte. Als Begründung wird herangezogen, dass der Test weitaus mehr Interpretationsmöglichkeiten bietet als durch eine standardisierte Interpretation zur Verfügung gestellt werden kann. Standardisierte Interpretationsmöglichkeiten sollten jedoch für jeden Test vorliegen. Einzelne Interpretationsbeispiele reichen nicht aus.

Reliabilität Unter Reliabilität versteht man den Grad der Genauigkeit, mit dem ein Test ein bestimmtes Merkmal misst, unabhängig davon, ob er dieses Merkmal auch zu messen beansprucht. Dabei werden drei Reliabilitätsarten unterschieden (siehe auch *Kapitel 4*):

- *Innere Konsistenz/Halbierungsreliabilität*

  Unter Halbierungsreliabilität versteht man Folgendes: Der Test wird in möglichst gleiche Testhälften unterteilt und diese werden miteinander korreliert. Dabei wird als Korrekturfaktor die Testlänge berücksichtigt. Unter innerer Konsistenz wird Folgendes verstanden: Jedes einzelne Item wird als eigenständiger Testteil angesehen, und die Messgenauigkeit stellt den mittleren Zusammenhang unter Berücksichtigung der Testlänge dar.

- *Retest-Reliabilität (oder Stabilität)*
  Der Test wird zu zwei verschiedenen Testzeitpunkten durchgeführt und dann die Korrelation zwischen den Testleistungen ermittelt. Bei der Retest-Reliabilität ist zu beachten, dass die Korrelationen in Abhängigkeit vom Zeitintervall zwischen den beiden Testungen variieren können. Beispielsweise können sich während dieses Zeitintervalls negative Lebensereignisse (Tod eines Angehörigen) auf die Persönlichkeit kurzfristig oder auch längerfristig auswirken. Erinnerungseffekte oder Übungseffekte könnten beispielsweise die zweite Intelligenzmessung „verfälschen".

- *Paralleltestreliabilität*
  Es wird die Korrelation zwischen zwei Tests, die dieselbe Eigenschaft oder Fähigkeit mittels verschiedenen Items („Itemzwillingen") erfassen sollen, berechnet.

Validität Unter Validität versteht man im eigentlichen Sinne das Ausmaß, in dem ein Test das misst, was er zu messen vorgibt. Nach Bryant (2000) unterscheidet man grundsätzlich drei Validitätsarten: **Inhaltsvalidität**, **Kriteriumsvalidität** und **Konstruktvalidität**. Murphy und Davidshofer (2001) weisen allerdings darauf hin, dass eigentlich nur die Inhaltsvalidität der obigen Definition entspricht. Der Inhalt des Tests bestimmt schließlich, was er misst. Wie im nächsten Abschnitt noch dargestellt wird, ist es jedoch sehr schwierig, die Inhaltsvalidität eines Tests zu bestimmen. Diese ist statistisch nicht prüfbar. Daher nutzt man üblicherweise die Messwerte einer Stichprobe in einem Test, um dessen Validität indirekt zu ermitteln (Kriteriums- und Konstruktvalidität). Streng genommen bestimmt man mit der Kriteriums- und Konstruktvalidität nicht die Validität des Tests im eigentlichen Sinne, sondern die Validität der abgeleiteten Aussagen, welche mit Hilfe der Testkennwerte getroffen werden (z.B. ein Intelligenztest misst die Intelligenzstruktur oder ein Persönlichkeitstest sagt Verhalten vorher).

- *Inhaltsvalidität*
  Von **Inhaltsvalidität** spricht man, wenn ein Test oder ein Testitem das zu messende Merkmal auch wirklich bzw. hinreichend genau erfasst. Man geht dabei nach Michel und Conrad (1982) von einem Repräsentationsschluss aus. Das heißt, dass die Testitems eine repräsentative Itemmenge aus dem „Universum" von Items bilden, die das interessierende Merkmal abbilden. Nach Michel und Conrad (1982, S. 57) wird die Inhaltsvalidität in der Regel nicht numerisch anhand eines Kennwertes, sondern „aufgrund logischer und fachlicher Überlegungen" bestimmt und „mit oder ohne Einschränkung akzeptiert oder verworfen". Die Autoren verweisen darauf, dass auch die Begriffe **logische Validität** oder **Augenscheinvalidität** (was eigentlich kein wissenschaftliches Konzept darstellt) mit der Inhaltsvalidität eng verbunden sind. Während die logische Validität in etwa der Inhaltsvalidität entspricht, wird unter Augenscheinvalidität verstanden, dass selbst ein Laie unmittelbar den Zusammenhang zwischen Testaufgaben und gemessenem Verhalten erkennt. Es ist in der Praxis sehr schwierig, zu beurteilen, ob ein Test eine repräsentative Itemmenge enthält. Es gibt sehr viele Verhaltensweisen, die eine Fähigkeit kennzeichnen; bei breiten Fähigkeiten nahezu unendlich viele. Ob eine Auswahl in einem solchen Falle repräsentativ ist, kann nur schwer entschieden werden. Nehmen wir an, wir konstruieren einen Test, der messen soll, wie gut eine Sekretärin ist. Was müsste ein guter inhaltsvalider Test erfassen? Umgang mit Softwareprogrammen, Tippgeschwindigkeit, Termine koordinieren und wahrscheinlich vieles mehr.

Ein gutes Beispiel, das Probleme bei der Interpretation der Inhaltsvalidität aufzeigt, liefern Murphy und Davidshofer (2001, S. 152): Man nehme eine Arbeitsprobe als Mittel zur Personalauswahl. Arbeitsproben beinhalten das, was wirklich später Teil der Arbeit ist. Solche Arbeitsproben werden hoch standardisiert dargeboten und sorgfältig beobachtet. Man könnte nun versucht sein, zu sagen, dass diese Art von Aufgabe hoch inhaltsvalide sei. Allerdings ist zu bedenken, dass im späteren Berufsalltag auch andere Aufgaben die Leistung bestimmen, die nicht in dieser Art und Weise beobachtet und bewertet wurden. Das heißt, hier wird eher die maximale spezifische Leistung in einer Konkurrenzsituation erfasst. Murphy und Davidshofer (2001, S. 152) bemerken, dass die typische Leistung nicht hoch mit der maximalen Leistung zusammenhängt. Allerdings wird es in der Regel die Erfassung der typischen Arbeitsleistung sein, auf welche die Arbeitsprobe abzielt.

Die Autoren geben ein hilfreiches Vorgehen an, wie man Inhaltsvalidität erfassen kann (S. 150): (1) Beschreibung der Inhaltsebene des Konstruktes (Fähigkeit, Eigenschaft); (2) Festlegung, welcher Inhaltsbereich durch welches Item erfasst wird; (3) Vergleich der Teststruktur mit der Struktur des Konstrukts. Da viele Konstrukte nur vage formuliert sind, ist gerade der erste Schritt nicht einfach. Man behilft sich hier mit Arbeitsdefinitionen oder man betrachtet nur Teilausschnitte eines Konstruktes.

Gerade die Testentwicklung krankt an der mäßigen Inhaltvalidität der Tests. Tests sind häufig das Ergebnis eines statistischen Homogenisierungsprozesses, der mit theoretischer Fundierung nichts mehr zu tun hat. Mangelnde Überlegungen am Anfang des Konstruktionsprozesses führen schon in der Entwicklungsphase zu unzureichenden Verfahren.

An dieser Stelle sei ein Beispiel zur Veranschaulichung vorgestellt. Der Testkonstrukteur Homo Genität möchte einen Fragebogen konstruieren, der Emotionale Intelligenz erfassen soll. Im ersten Schritt überlegt er sich selbst Fragen, die Emotionale Intelligenz erfassen sollen, z.B.: „Ich kann mich gut in andere hineinversetzen." Zu seinem Glück fallen ihm sehr viele solcher Fragen ein. Er vermutet aber, dass Emotionale Intelligenz mehr Facetten hat, sucht nach Verfahren, die es bereits auf dem Markt gibt, und formuliert Items, die er dort vorfindet, einfach um. Er versucht auch, jede Skala, die er in einem anderen Fragebogen vorfindet, in die Struktur seines Tests zu integrieren. Er glaubt, dadurch würde sein Test inhaltsvalider. Da er sich bei der Formulierung der Items nicht immer sicher ist, beinhalten manche Skalen Items doppelt oder mehrfach mit nur unmerklich anderer Formulierung. Fallen ihm zu einer Facette besonders viele Fragen ein, nimmt er alle diese Fragen mit auf. Er denkt, dass viele Fragen ja nicht schaden können, und rausschmeißen kann man Fragen immer noch. Da ihm zu manchen Facetten aber kaum Fragen einfallen, verwendet er für solche Skalen nur wenige Items. Damit die Skalen aber trotzdem eine hohe Messgenauigkeit erzielen, verwendet er einfach das Item, von dem er glaubt, es sei das beste, mit unterschiedlichen Formulierungen mehrfach: „Ich weiß, was andere Menschen fühlen;" „Gefühle anderer Menschen erkenne ich;" „Was andere Menschen fühlen, finde ich schnell heraus." Homo Genität generiert so einen Itemsatz von 600 Items und 18 Skalen, den er einer ersten Stichprobe zur Bearbeitung gibt. Danach wird der Test einer Faktorenanalyse unterzogen, und nur Items, die inhaltlich gut zum Faktor passen, werden in den Fragebogen aufgenommen. Eigentlich war angedacht, den Test für die Bewerberauslese von Auszubildenden in der Industrie zu konzipieren, da aber Stichproben in der Praxis nicht gut zu

bekommen sind, wurden 80 Studenten vor der Mensa befragt. Jedem Student wurde auch eine Rückmeldung versprochen, damit er den Fragebogen gewissenhaft ausfüllt und nicht sozial erwünscht antwortet. Es wurden dann per Voreinstellung 18 unkorrelierte Faktoren in der Faktorenanalyse extrahiert. Manche Skalen beinhalteten am Ende nur vier oder fünf Items, andere 20 Items. Da 20 Fragen für eine Facette viele Fragen sind und der Test ja auch ökonomisch sein soll, führt Home Genität so lange Itemanalysen durch, bis maximal nur vier bis zehn Fragen pro Facette übrig sind. Mit dem nun gekürzten und optimierten Fragebogen besucht Homo Genität jetzt zwei große Vorlesungen (etwa 500 Studenten), um den Test zu normieren und die Struktur des Fragebogens zu bestätigen. Nachdem er gute Erfahrungen bei der Rückmeldung gemacht hat, verspricht er auch den Teilnehmern der Vorlesung eine Rückmeldung. Weil er aber den Studenten nicht 100-prozentig vertraut, lässt er einen Fragebogen zur sozialen Erwünschtheit mitlaufen. Als er festellt, dass es keine Korrelationen mit diesem Fragebogen gibt, ist er sich sicher, dass der Fragebogen nicht verfälschbar und somit für die Bewerberauslese geeignet ist. Ich hoffe, Sie werden nach der Lektüre dieses Buches erkennen, welche Fehler Homo Genität begangen hat, und selbst ein angemesseneres Vorgehen wählen.

- *Kriteriumsvalidität*

  Es handelt sich hier um den Zusammenhang der Testleistung mit einem oder mehreren Kriterien (z.B. Schulnote), mit denen der Test aufgrund seines Messanspruches korrelieren sollte. Man bezeichnet dies auch als Korrelationsschluss, das heißt, die Prüfung der Kriteriumsvalidität basiert auf Zusammenhängen zwischen Testkennwerten und Kriterien. Man unterscheidt folgende Arten von Kriteriumsvalidität:

  – *Vorhersagevalidität* (prognostische Validität, prädiktive Validität). Es werden Zusammenhänge (Korrelationen) mit zeitlich später erhobenen Kriterien ermittelt. Beispielsweise wird die Intelligenztestleistung vor Beginn der Lehre ermittelt und mit der Abschlussnote der Ausbildung korreliert.
  – *Übereinstimmungsvalidität* (konkurrente Validität). Korrelationen mit zeitlich (fast) gleich erhobenen Kriterien. Beispielsweise könnte die Konzentrationstestleistung vor Beginn einer Klausur ermittelt und dann die Korrelation mit der Klausurnote berechnet werden.
  – *Retrospektive Validität.* Es werden Zusammenhänge (Korrelationen) mit zeitlich vorher ermittelten Kriterien berechnet. Beispielsweise wird die Intelligenztestleistung während des Studiums erhoben und mit den Schulnoten des zurückliegenden Abiturs korreliert.
  – *Inkrementelle Validität.* Sie bezeichnet den Beitrag eines Tests zur Verbesserung der Vorhersage eines Kriteriums über einen anderen Test hinaus. Durch Intelligenztests lässt sich beispielsweise besonders gut Berufserfolg vorhersagen. Jede andere diagnostische Methode muss sich nun daran messen lassen, ob sie über die Intelligenz hinaus noch etwas zur Vorhersage von Berufserfolg beitragen kann. Eine der Methoden, die das leisten kann, ist beispielsweise das strukturierte Interview (Schmidt & Hunter, 1998). Zur Feststellung der inkrementellen Validität werden hierarchische Regressionsanalysen verwendet.

- *Konstruktvalidität*

  Mit der **Konstruktvalidität** soll abgeleitet werden, dass der Test auch die Eigenschaft oder Fähigkeit misst, die er messen soll. Viele Autoren fassen unter der Konstruktvalidität alle Validitätsarten (z.B. Kriteriumsvalidität, Inhaltsvalidität, konvergente und diskriminante Validität) zusammen. In diesem Sinne sagt die Konstruktvalidität etwas darüber aus, wie angemessen ein Test das erfasst, was er zu messen beansprucht.

  Fasst man Konstruktvalidität enger, fallen darunter lediglich konvergente, diskriminante und faktorielle Validität. Für diese existieren im Gegensatz zur Inhaltsvalidität konkrete Strategien zur Quantifizierung. Ein häufig gewählter Ansatz besteht darin, a priori konkrete Erwartungen über den Zusammenhang des vorliegenden Tests mit konstruktverwandten (konvergenten) und konstruktfremden (diskriminanten) Tests zu formulieren. Der Nachteil dieses Ansatzes besteht nicht selten darin, dass ein Test mit einem oder mehreren anderen Tests verglichen wird, dessen/deren Inhaltsvalidität selbst unzureichend ist. Innerhalb dieses Ansatzes kann zwischen konvergenter Validität und diskriminanter/divergenter Validität unterschieden werden:

  – *Konvergente Validität*

    Es werden Korrelationen mit Tests gleicher oder ähnlicher Gültigkeitsbereiche ermittelt, z.B. die Korrelation eines neu entwickelten Intelligenztests, wie dem I-S-T 2000 R (Amthauer, Brocke, Liepmann & Beauducel, 2001), mit einem bereits etablierten Verfahren, z.B. dem HAWIE-R (Tewes, 1991). Man erwartet hier hohe Zusammenhänge.

  – *Diskriminante/divergente Validität*

    Es werden Korrelationen mit Tests anderer Gültigkeitsbereiche ermittelt. Beispielsweise wird die Korrelation eines Konzentrationstests mit einem Arbeitsgedächtnistest ermittelt. Der Konzentrationstest soll nämlich nicht die Arbeitsgedächtnisleistung erfassen, sondern möglichst rein das Konstrukt „Konzentration". Man erwartet hier niedrigere Zusammenhänge. Es ist sinnvoll, an dieser Stelle nicht nur Leistungen heranzuziehen, die offensichtlich etwas anderes messen (z.B. Kreativität), sondern auch Leistungen, die einem verwandten Konstrukt zugeordnet werden können (z.B. Gedächtnis, Aufmerksamkeit). In diesem Falle möchte man sichergehen, dass man eben gerade nicht dieses verwandte Konstrukt erfasst.

Es gibt verschiedene Methoden, um konvergente und diskriminante Validität zu bestimmen. Die drei häufigsten seien hier kurz aufgeführt. An erster Stelle sind Korrelationen zu nennen. In Testhandbüchern werden häufig Korrelationen mit konstruktnahen und konstruktfremden Verfahren angegeben.

Ebenfalls sehr häufig werden die Zusammenhänge zwischen verschiedenen Tests mit Faktorenanalysen untersucht. Die so genannte *faktorielle Validität* dient zum einen dazu, homogene konstruktnahe Inhaltsbereiche zusammenzufassen, und zum anderen, diese von konstruktfremden Bereichen zu trennen (siehe *Kapitel 5*). Systematischer und aufwändiger ist die Strategie, die konvergente und diskriminante Validität eines Tests mit dem so genannten **Multitrait-Multimethoden-Ansatz** (Campbell & Fiske, 1959) zu ermitteln (siehe für einen ausgezeichneten Überblick Schermelleh-Engel & Schweizer, 2003, S. 103 ff.). Bei dieser Methode werden verschiedene Matrizen (Korrelationsmatrizen) gebildet: Monotrait-Monomethoden-Matrix, Monotrait-Heteromethoden-Matrix, Heterotrait-Monomethoden-Matrix und die Heterotrait-Heteromethoden-

Matrix. Dem Ansatz liegt die Überlegung zugrunde, dass Kennwerte einer **Fähigkeit** (z.B. Intelligenz), die mit den gleichen **Methoden** (z.B. Tests) erfasst werden, am höchsten miteinander zusammenhängen (Monotrait-Monomethoden-Matrix). Im Vergleich dazu sollte der Zusammenhang zwischen Kennwerten einer Fähigkeit (z.B. Intelligenz) geringer ausfallen, wenn diese mit unterschiedlichen Methoden (Intelligenztest, Verhaltensbeobachtung) erfasst werden (Monotrait-Heteromethoden-Matrix). Ein wiederum geringerer Zusammenhang ist zwischen den Kennwerten unterschiedlicher Fähigkeiten (z.B. Intelligenz und Konzentration) zu erwarten, die mit der gleichen Methode (Tests) erhoben wurden (Heterotrait-Monomethoden-Matrix). Am geringsten sollten die Zusammenhänge zwischen Kennwerten unterschiedlicher Fähigkeiten (z.B. Intelligenz und Konzentration) ausfallen, wenn sie mit unterschiedlichen Methoden (z.B. Test und Verhaltensbeobachtung) erfasst werden (Heterotrait-Heteromethoden-Matrix). Mit diesem Verfahren wird auf eine wichtige Varianzquelle hingewiesen, nämlich die Methodenvarianz. Die Varianz der Testwerte geht also nicht nur auf eine Fähigkeit oder Eigenschaft zurück, sondern ist auch auf Methodenvarianz zurückzuführen. Beispielsweise korrelieren ein zeitbegrenzter Intelligenztest und ein zeitbegrenzter Konzentrationstests miteinander, da beide Tests die gleiche Methode zur Messung unterschiedlicher Fähigkeiten verwenden (vgl. Wilhelm & Schulze, 2002).

Nach Messick (1995) sollte bei der Betrachtung eines Tests oder der Interpretation der daraus abgeleiteten Testkennwerte stets beachtet werden, dass bei der Erfassung von Konstrukten in der Regel zwei Aspekte die Konstruktvalidität beeinträchtigen können: Die Konstruktvalidität eines Tests kann durch die **Unterrepräsentation des Konstrukts** im Test beeinflusst sein. Dies bedeutet, dass eine Messung zu „eng" erfolgt und die wichtigsten Aspekte des Konstrukts im Test nicht enthalten sind. Als ein zweiter Einflussfaktor auf die Konstruktvalidität lässt sich die **Konstrukt-irrelevante Varianz** bezeichnen. In diesem Falle erfolgt die Messung zu „breit", was bedeutet, dass zusätzlich zur eigentlichen Messintention eines Tests noch andere Aspekte erfasst werden. So könnte ein Test in einem gewissen Ausmaß zusätzlich Aspekte anderer Konstrukte erfassen oder aber Varianzaspekte enthalten, die sich auf die verwendete Methode selbst zurückführen lassen. Es lassen sich zwei Formen der Konstrukt-irrelevanten Varianz unterscheiden: Im Falle der **Konstrukt-irrelevanten Schwierigkeit** werden Testaufgaben für Individuen oder Gruppen durch Aspekte, die außerhalb des eigentlichen Fokus des Konstrukts liegen, erschwert. Nehmen wir an, eine Person soll einen Konzentrationstest bearbeiten, der ein schnelles Durchstreichen von Zeichen verlangt. Die Person hat aber eine leichte Lähmung der dominanten Hand. Diese Person wird in diesem Test schlechter abschneiden als eine gesunde Person, was aber nicht an der Konzentrationsfähigkeit liegt, sondern an der Verlangsamung, welche durch die leichte Lähmung bedingt ist. Im Falle der **Konstrukt-irrelevanten Leichtigkeit** sind die Aufgaben des Tests durch Konstrukt-irrelevante Aspekte leichter zu lösen. Als ein Beispiel lässt sich hier die Vertrautheit mit dem Testmaterial nennen. So könnte z.B. einer Person, die häufig Planspiele am Computer spielt, die Bearbeitung von computergestützten Problemlöse-Szenarien leichter fallen, da sie mit der Situation, komplexe Denkaufgaben am Computer zu lösen, vertraut ist.

Als Fazit ist zur Konstruktvalidität anzumerken, dass Konstruktvalidierung ein sorgfältig geplanter Prozess ist. Das heißt, es müssen Überlegungen angestellt werden, welches Verfahren einen ähnlichen Messanspruch hat und welches Verfahren aus theoretischen oder praktischen Erwägungen abgegrenzt werden muss. Darüber hinaus sind bei der weiteren Fassung des Begriffs „Konstruktvalidität" Angaben zur Inhalts-

validität zu machen. Zusätzlich muss im Vorhinein festgelegt werden, welches verhaltensrelevante Kriterium erfasst werden soll. Erst wenn alle Korrelationen als Gesamtpaket den a priori gestellten Erwartungen entsprechen bzw. die Inhaltsvalidität plausibel nachvollzogen werden kann, ist von einer gesicherten Konstruktvalidiät auszugehen. Es handelt sich daher nicht um einen vagen Prozess, bei dem lediglich Korrelationen bereitgestellt werden. Vielmehr sollte sich durch die Konstruktvalidierung ein a priori aufgestelltes Korrelationsmuster zeigen. Allerdings sei an dieser Stelle auch angemerkt, dass es keine Richtlinien oder festgelegten Grenzen für die Höhe der Korrelationen im Rahmen der Konstruktvalidität gibt. Das heißt, die Aussage, dass Konstruktvalidität vorliegt, sollte jeder Testanwender selbst durch einen kritischen Blick auf die ermittelten Korrelationen prüfen.

Gründe für mangelnde Validität

Es lassen sich fünf Hauptgründe für eine mangelnde Validität angeben. Dabei ist hier insbesondere die Kriteriumsvalidität gemeint. Die Ausführungen beziehen sich auf Korrelationen (siehe *Kapitel 8*). Korrelationen zwischen einem Kriterium (z.B. Berufserfolg, gemessen durch Vorgesetztenbeurteilungen) und einem Test bzw. Prädiktor (z.B. Intelligenztestleistung) können aus folgenden Gründen gemindert sein:

- *Methodenfaktoren*
  Damit ist gemeint, dass eine Korrelation niedrig ausfallen kann, wenn verschiedene Methoden verwendet wurden, um Prädiktor und Kriterium zu erfassen. Beispielsweise wurde als Prädiktor die Intelligenztestleistung gewählt und als Kriterium eine Vorgesetztenbeurteilung mit Hilfe eines standardisierten Fragebogens. In diesem Fall wurde das Kriterium mit einer anderen Methode (Fremdbeurteilung) als der Prädiktor (Leistungstest) erhoben. Dies kann sich mindernd auf die Korrelation zwischen Kriterium und Prädiktor auswirken.

- *Kriteriumskontamination und -defizienz*
  Unter Kriteriumskontamination versteht man, dass das Kriterium eigentlich etwas anderes erfasst als beabsichtigt. Beispielsweise wird als Kriterium für Berufserfolg der Umsatz der Mitarbeiter herangezogen. Der Umsatz wird jedoch auch durch die Größe des Verkaufsgebiets mitbestimmt, ist also kontaminiert.

  Unter Kriteriumsdefizienz versteht man, dass das Kriterium wichtige Aspekte nicht beinhaltet. Das Kriterium „Umsatz“ ist defizient, da es keine Information über die Kundenzufriedenheit enthält. Beispielsweise könnte ja ein hoher Umsatz auch dadurch zustande gekommen sein, dass Mitarbeiter Kunden unwahre Versprechungen über das Produkt machen. Infolgedessen steigt zwar kurzfristig der Umsatz, die Kundenzufriedenheit sinkt jedoch.

- *Mangelnde Symmetrie zwischen Prädiktor und Kriterium*
  Unter mangelnder Symmetrie wird verstanden, dass zwei zu korrelierende Tests eine Eigenschaft oder Fähigkeit unterschiedlich breit messen. Nehmen wir als Beispiel die Big-Five. Wird beispielsweise ein kleiner Verhaltensausschnitt, z.B. Geselligkeit, mit einem großem Verhaltensausschnitt, wie etwa Gewissenhaftigkeit, korreliert, kann keine maximale Korrelation auftreten. Geselligkeit ist lediglich ein Unterbereich eines generelleren Persönlichkeitsmerkmals (Extraversion). Gewissenhaftigkeit hingegen ist ebenfalls ein generelleres Persönlichkeitsmerkmal. In

diesem Fall wäre eine höhere Korrelation zwischen Extraversion und Gewissenhaftigkeit zu erwarten, da beide dieselbe Generaltiätsebene erfassen. Solche Überlegungen werden gerne vernachlässigt, sind jedoch von großer Bedeutung. Weiterführende Informationen finden sich bei Wittmann (2002).

- *Streuungsrestriktion*
  Wann eine Streuungsrestriktion vorliegt, wird im Folgenden beschrieben. Nehmen wir an, eine untersuchte Personengruppe ist bezüglich eines bestimmten Merkmals vorausgewählt. Beispielsweise werden zum Psychologiestudium nur Personen mit einem Abitur von 1.4 oder besser zugelassen. Nun wird die Abiturnote als Prädiktor zur Vorhersage der Vordiplomsnoten herangezogen. Die Psychologiestudenten unterscheiden sich durch diese Vorselektion aber nur geringfügig in ihren Abiturnoten. Das heißt, die Varianz der Abiturnote ist eingeschränkt. Eine hohe Korrelation setzt aber voraus, dass die zu korrelierenden Merkmale auch genügend große Varianzen aufweisen. Ist dies nicht der Fall, kann sich dies mindernd auf die Korrelation auswirken. Eine Untersuchung würde möglicherweise zu folgendem Schluss kommen: Bei Psychologiestudenten weisen die Abiturnote und die Vordiplomsnoten keinen starken Zusammenhang auf. Dies wäre aber eine Fehleinschätzung, da die Korrelation aufgrund der eingeschränkten Varianz gemindert wird. Methoden, um mit einer solchen Streuungsrestriktion umzugehen, werden in *Abschnitt 8.9* dargestellt.

- *Mangelnde Reliabilität im Kriterium oder Prädiktor*
  Eine geringe Reliabilität bedeutet, dass eine Eigenschaft oder Fähigkeit nur mit einer geringen Genauigkeit erfasst wird. Das heißt, die Messung beinhaltet auch einen großen Anteil an unsystematischer Varianz, die auch als Fehlervarianz bezeichnet wird. Korreliert man nun Prädiktor und Kriterium, dann wird die Höhe des Zusammenhangs durch die systematische gemeinsame Varianz bestimmt. Fällt diese gering aus, kann auch der Zusammenhang zwischen Prädiktor und Kriterium nicht maximal werden. In einem solchen Fall ist es möglich, die Korrelation mittels einer Minderungskorrektur aufzuwerten (siehe *Abschnitt 4.3*). Eine minderungskorrigierte Korrelation gibt an, wie hoch der Zusammenhang zwischen Prädiktor und Kriterium wäre, wenn beide perfekt, das heißt ohne Messfehler, erfasst werden könnten. An dieser Stelle sei bereits angemerkt, dass es unrealistisch ist, ohne Messfehler zu messen. Minderungskorrigierte Korrelationen spiegeln also einen Zusammenhang wider, der in der Praxis nie erzielt werden kann. Sollen jedoch beispielsweise Korrelationen zwischen Testwerten in unterschiedlichen Stichproben verglichen werden, kann die Korrektur des Zusammenhangs durch die unterschiedlichen Reliabilitäten aufschlussreich sein. Möglicherweise sind vorgefundene Korrelationsunterschiede nicht inhaltlich begründet, sondern messfehlerbedingt.

Zusammenhang zwischen Objektivität, Reliabilität und Validität

Die Hauptgütekriterien (Objektivität, Reliabilität, Validität) stehen in einem bestimmten Abhängigkeitsverhältnis. Ein Test, der nicht objektiv ist, kann mit großer Wahrscheinlichkeit keine optimale Reliabilität erreichen. Gelangen unterschiedliche Auswerter zu unterschiedlichen Testergebnissen einer Person, kann dies daran liegen, dass die Durchführungsbedingungen variieren oder die Art der Auswertung variiert und damit nicht eindeutig festgelegt ist. Es entstehen Fehler bei der Ermittlung und Interpretation der Ergebnisse, und das Testergebnis ist verzerrt. Diese Fehler beeinflussen

die Reliabilität eines Tests (wie genau der Test ein Merkmal erfasst). Ist die Reliabilität gering, kann die Vorhersage auf ein Kriterium (Validität), zum Beispiel „Berufserfolg", nicht hoch sein, da der Test nur sehr ungenau den Wert eines Probanden schätzt. Der Zusammenhang zwischen Testwerten und einem Kriterium hängt rechnerisch von der Reliabilität ab. Das heißt, misst man zwei Merkmale nur sehr ungenau, kann der systematische wahre Zusammenhang nur unzureichend geschätzt werden. Die Korrelation zwischen zwei Merkmalen kann maximal den folgenden Wert annehmen:

$$r_{\max} = \sqrt{r_{tt1} \cdot r_{tt2}}$$

Dabei ist:

- r_{max} = maximale Korrelation zwischen zwei Variablen oder Tests
- r_{tt1} = Reliabilität der ersten Variable oder des ersten Tests
- r_{tt2} = Reliabilität der zweiten Variable oder des zweiten Tests

2.5.2 Nebengütekriterien

Was sind weitere Indikatoren für einen guten psychometrischen Test?

Man unterscheidet vier Nebengütekriterien:

- Normierung
- Vergleichbarkeit
- Ökonomie
- Nützlichkeit

Diese sind neben den Hauptgütekriterien wichtig für die Beurteilung der Güte eines Tests.

Normierung Über einen Test liegen Angaben (Normen) vor, die als Bezugssystem für die Einordnung des individuellen Testergebnisses dienen können. Die Testnormierung ermöglicht es, die Frage zu beantworten, ob eine Person unterdurchschnittlich, durchschnittlich oder überdurchschnittlich im Vergleich zu anderen Personen abgeschnitten hat. Testnormen sollten aktuell sein (nicht älter als acht Jahre, siehe DIN 33430) und vor allem für verschiedene Personengruppen vorliegen. Als Daumenregel für die Mindestgröße von Normstichproben kann die Zahl 300 Probanden gelten. Für die Normstichproben sollten genaue Angaben zu folgenden Punkten vorliegen:

- Repräsentativität der Stichprobe
  - z.B. repräsentativ für Deutschland, repräsentativ für alle hessischen Gymnasiasten, ...
- Anwerbung der Stichprobe
  - z.B. Zeitung, Testinstitut, Werbung durch Hilfskräfte, ...
- Bedingungen, unter denenen die Stichprobe getestet wurde
  - Wurde in einer Bewerbungssituation oder in einem klinischen Setting getestet?
  - Fand eine Rückmeldung der Ergebnisse statt?
  - Wurde die Teilnahme bezahlt?

- Wurde eine Einzel- oder Gruppentestung durchgeführt?
- Wurde die Testung unter Aufsicht in einem Labor oder zu Hause durchgeführt?
- Zu welcher Tageszeit fanden die Untersuchungen statt?
- Wie lange dauerte der Test?
- An welcher Position wurde der Test durchgeführt, falls mehrere Verfahren eingesetzt wurden?
- ...

■ Zusammensetzung der Normstichprobe
- z.B. Alter, Geschlecht, Bildung, ...

Vergleichbarkeit Ein Test ist dann vergleichbar, wenn ein oder mehrere Parallelformen oder Tests mit gleichen Gültigkeitsbereichen vorhanden sind. Letzteres meint, dass eine Person in zwei Tests, die Ähnliches messen sollen, auch ähnliche Ergebnisse erzielen sollte. Parallelformen sind besonders vorteilhaft, wenn ein Test mehrmals an einer Person durchgeführt wird oder wenn mehrere Personen in einer Gruppensituation getestet werden, damit „Abschreiben" verhindert wird.

Ökonomie Ein Test ist dann ökonomisch, wenn er (1) eine kurze Durchführungszeit beansprucht und wenig Material verbraucht, wenn er (2) einfach zu handhaben ist, wenn er (3) als Gruppentest durchführbar ist und wenn er (4) schnell und bequem auszuwerten ist.

Nützlichkeit Ein Test gilt dann als nützlich, wenn er ein Persönlichkeitsmerkmal oder eine Verhaltensweise misst oder vorhersagt, für dessen oder deren Untersuchung ein praktisches Bedürfnis besteht. Leider besteht für viele Tests kein praktisches Bedürfnis. Sie sind zum Teil redundant. Manche Fragebogen messen möglicherweise nichts anderes als bereits bekannte Persönlichkeitsmerkmale. Bevor ein Test entwickelt wird, muss also geprüft werden, ob nicht bereits ein Verfahren vorliegt, das denselben Messanspruch hat. Liegt ein solches Verfahren vor, sollte begründet werden, welche Vorteile das neue Verfahren gegenüber dem bereits bestehenden Verfahren aufweist.

Zusammenfassend wird in *Abbildung 2.2* dargestellt, welche Haupt- und Nebengütekriterien zur Beurteilung von Testverfahren herangezogen werden können.

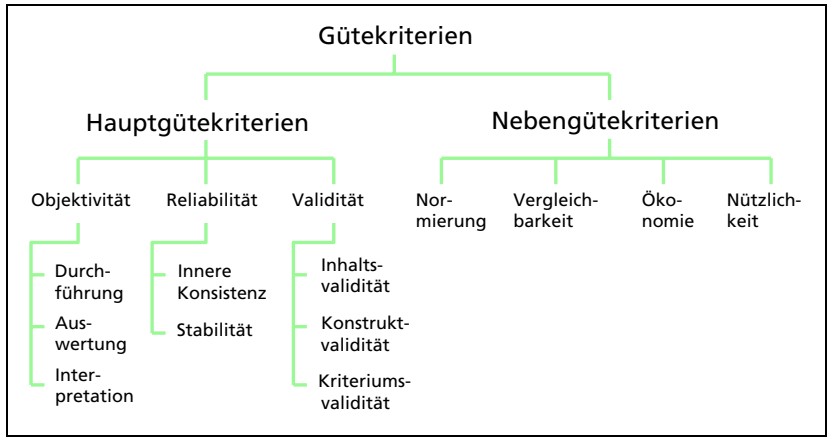

Abbildung 2.2: Zusammenfassung der Haupt- und Nebengütekriterien