Leseprobe aus: Lienert, Testaufbau und Testanalyse, ISBN 978-3-621-27424-1

http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-407-25538-9

Kapitel 1

GRUNDSÄTZLICHES ÜBER DEN TEST

1.1 Wesen und Bedeutung des Tests

1.1.1 Zur Begriffsbestimmung des Wortes „Test"

Das Wort „*Test*" stammt bekanntlich aus dem englischen Sprachgebrauch und bedeutet soviel wie Probe. Dieser Begriff ist inzwischen in den deutschen Sprachschatz eingegangen, wobei sich für den Plural die englische Form „Tests" durchgesetzt hat.

Das Wort „Test" hat eine *mehrfache* Bedeutung. Man versteht darunter:

a) Ein Verfahren zur Untersuchung eines Persönlichkeitsmerkmals.
b) Den Vorgang der Durchführung der Untersuchung.
c) Die Gesamtheit der zur Durchführung notwendigen Materialien.
d) Jede Untersuchung, sofern sie Stichprobencharakter hat.
e) Gewisse mathematisch-statistische Prüfverfahren (z. B. χ^2-Test).
f) Kurze, außerplanmäßige „Zettelarbeiten" im Schulunterricht.

Unter diesen Bedeutungen ist die erste die weitaus wichtigste; sie soll in der folgenden Definition in ihren in diesem Zusammenhang wesentlichen Punkten festgelegt werden:

Definition: Ein 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.[1])

Die Bedeutung des Wortes „Test" ist also hier in mannigfacher Weise *eingeschränkt*: nicht jede beliebige zu diagnostischen Zwecken angestellte Untersuchung kann als Test gelten, sondern nur eine solche, die erstens wissenschaftlich begründet ist; zweitens routinemäßig – also unter Standardbedingungen mehr oder weniger handwerksmäßig – durchführbar ist; drittens eine relative Positionsbestimmung des untersuchten Individuums innerhalb einer Gruppe von Individuen oder im Bezug auf ein bestimmtes Kriterium, z. B. ein Lehrziel, ermöglicht; und viertens bestimmte empirisch – also verhaltens- und erlebnisanalytisch, phänomenologisch und nicht etwa rein begrifflich – abgrenzbare Eigenschaften, Verhaltensdispositionen, Fähigkeiten, Fertigkeiten oder Kenntnisse prüft.

Bedingt und nicht ausdrücklich wird vorausgesetzt, daß das untersuchte Merkmal eindimensional und metrisch ist, daß es also entlang einer *Einheitskala* meßbar ist, was keineswegs für alle in der Persönlichkeitsbeschreibung üblichen

---

[1]) Die Definition lehnt sich unter Hervorhebung der wesentlichen Merkmale eng an: WARREN (1934).

Begriffe gesichert zu sein scheint. Diese Einschränkung kommt in der Formulierung „möglichst quantitative Aussage" zum Ausdruck.

Die vorgenannte Definition des Wortes „Test" erhebt naturgemäß keinen Anspruch auf Allgemeingültigkeit, sondern dient lediglich als mehr oder weniger operationale Umschreibung für den hier zu erörternden Tatbestand.

1.1.2 Die geschichtlichen Stadien der Testentwicklung

Die ersten Versuche einer quantitativen Erfassung von Persönlichkeitsmerkmalen galten den *Intelligenzdefekten* und gingen demgemäß von psychiatrischer Seite aus. RIEGER hat 1885 als einer der ersten einen „Entwurf zu einer allgemein anwendbaren Methode der Intelligenzprüfung" ausgearbeitet. Darin werden Anweisungen erteilt, wie man verschiedene psychische Funktionen, wie Perzeption, Apperzeption, Gedächtnis, Kombinationsfähigkeit usw. untersuchen kann. Die Wahl der konkreten Aufgabe blieb weitgehend dem Untersucher überlassen.

Ein Jahrzehnt später formuliert ZIEHEN in den „Prinzipien und Methoden der Intelligenzprüfung" 1897 seine Aufgaben bereits im Sinne der heutigen Tests. ZIEHEN ahnt das Kriterium der Trennschärfe von Aufgaben voraus, wenn er darauf dringt, daß man nicht – wie dies noch RIEGER tat – naiv entscheiden könne, welche Antwort normal und welche anormal sei, sondern daß man prüfen müsse, ob und wie oft unzureichende Antworten von Normalen gegeben würden. Erst dadurch bekäme man einen Einblick in die normale Schwankungsbreite.

Ein wenig später beginnen die individuellen Differenzen auch für die *experimentelle Psychologie* interessant zu werden. In seinem Buch „Mental Tests and Measurements" von 1890 schlägt der WUNDT-Schüler McKEEN CATTELL zwei Prüfreihen vor, die sich vorwiegend auf *elementare* psychische und psychophysische Funktionen wie Schwellen- und Reaktionszeitbestimmungen beschränken (die Untersuchung komplexer Phänomene, wie z. B. Intelligenzleistungen, wurde von der Leipziger Schule nicht durchgeführt). CATTELL befürwortet in seiner Arbeit bereits eine Vereinheitlichung, ein einheitliches System von Prüftests, damit die Ergebnisse verschiedener Untersucher vergleichbar würden. KRAEPELIN, MÜNSTERBERG, MOEDE und vor allem EBBINGHAUS und MEUMANN greifen die Anregungen auf und wenden sich in ihren Testvorschlägen mehr den *höheren* psychischen Funktionen, wie Auffassen, Einordnen, Lernen und Gedächtnis zu. Immer noch aber sind die Aufgabenbereiche des Tests als Prüf- und als Forschungsexperiment – Begriffe, die später von STERN geprägt wurden – nicht klar geschieden.

Mit der Jahrhundertwende geht die von HYLLA (1927) so benannte *Vorperiode der Testentwicklung* zu Ende. Die nachfolgende Zeit ist durch ein zweifaches Bemühen gekennzeichnet: sie will nicht mehr psychische Einzelfunktionen, sondern „die Intelligenz" erfassen, und weiter will sie diese nach Maß und Zahl bewerten.

In diese Zeit fällt der Einbruch *mathematisch-statistischer Methoden*, als deren Begründer und langjähriger Proponent im Bereich der psychologischen Forschung

Ch. SPEARMAN gelten darf. Er war es, der in „General Intelligence Objectively Determined and Measured" im Jahre 1905 die theoretischen Grundlagen der modernen Testentwicklung geschaffen hat.

Um dieselbe Zeit erhält die Testpraxis einen großartigen Aufschwung durch die Veröffentlichung der „Testserie zur Auslese Minderbegabter" durch BINET und SIMON, die in ihrer Revision von 1908 bereits eine Art *Aufgabenanalyse* bei der Ausschaltung ungeeigneter Aufgaben anwandten. BOBERTAG hat seit 1911 die BINET-Methode den deutschen Verhältnissen erfolgreich angepaßt.

Ebenfalls seit 1911 hat W. STERN als bedeutendster deutscher Testforscher drei wichtige Schriften veröffentlicht: die erste als eine Zusammenschau der bisherigen Ergebnisse unter theoretischen und systematischen Gesichtspunkten: „Die differentielle Psychologie und ihre methodischen Grundlagen" (1911), die zweite unter vorwiegend praktischem Blickpunkt: „Die Intelligenzprüfung an Kindern und Jugendlichen" (1912) mit der *klassischen* Definition des *Intelligenzquotienten* ($IQ = 100 \cdot JA/LA$) die dritte Schrift 1920 ist eine Zusammenfassung und Neubearbeitung der Schrift von 1912.

In Amerika hat sich die Standford-Universität unter L. M. TERMAN seit 1916 der BINET'schen Methode angenommen und verschiedene Revisionen durchgeführt, von denen der Stanford-BINET-Test bis auf den heutigen Tag verwendet wird. Die deutsche Bearbeitung stammt von LÜCKERT (1965). TERMAN ist auch die *psychometrische Verallgemeinerung* der klassischen Definition des STERN'schen IQ ($IQ = 15 \cdot (X - \bar{X}) + 100$) zu verdanken (s. Kap. 12).[1])

Ein neuer Umbruch in der Testentwicklung beginnt mit der Einführung des *Gruppentests* durch die amerikanische Armee im 1. Weltkrieg (TERMAN 1916). Dadurch wurde die Testmethode erst ökonomisch. Zugleich aber warf der Gruppentest neue Probleme auf, die sehr viel zur Förderung der modernen Testanalyse beitrugen und früh zur Entwicklung von Parallelformen führten. In der Monographie „The Reliability and Validity of Tests" hat L. L. THURSTONE dann 1931 die Leitsätze zur Entwicklung von standardisierten Gruppentests aufgestellt. Mit seinem Schüler GULLIKSEN (1950) ist die „*Klassische Testtheorie*" entstanden (siehe Kap. 10).

Seit dem 1. Weltkrieg wurden zunehmend auch andere als kognitige Persönlichkeitsbezüge durch objektive Tests untersucht. Der *Persönlichkeitsfragebogen* wurde durch das „Personal Data Sheet" von WOODWORTH (1920) begründet. Diesem Fragebogen folgten andere über Interessen und Einstellungen.

Die die objektiven Tests ergänzenden *projektiven Techniken* gehen bekanntlich auf RORSCHACH's Formdeuteversuch (1921) zurück; sie strecken hinsichtlich ihrer wissenschaftlichen Analyse allerdings bis heute noch in den Anfängen, von wenigen Ausnahmen, beispielsweise FISCHER's (1972) probabilistischen Zugang, abgesehen.

Ähnliches gilt für die in den 20er Jahren von HARTSHORNE und MAY propagierten *Situationstests*. Sie bieten eine Fülle diagnostischer Möglichkeiten auf der Grundlage von Ausdrucks- und Verhaltensbeobachtung.

---

[1]) Die Zahl 15 kommt folgendermaßen zustande: Die empirische Standardabweichung des „klassischen" IQ in verschiedenen Intelligenztests bei Kindern und Jugendlichen liegt zwischen 14 und 16. Um den „psychometrischen" IQ, der für Kinder und für Erwachsene gelten soll, diesem klassischen IQ anzugleichen, hat man hier die Streuung auf 15 festgelegt.

Kriterienorientierte Tests werden seit Anfang der 60er Jahre diskutiert. Sie messen, ob ein Individuum durch eine „Behandlung" im weitesten Sinne (z. B. Therapie, Unterricht) ein vorher festgelegtes Kriterium erreicht hat oder nicht, und wenn ja, wie gut. Diese Verfahren können in der Therapie (z. B. SCHOTT 1973), insbesondere aber als *„lehrzielorientierte Tests"* in der Schule (KLAUER 1972) eingesetzt werden. Ihre testtheoretischen Grundlagen stellen z. B. FRICKE (1974) und KLAUER (1987), ihre Konstruktionsprinzipien HERBIG (1976), das zugrundeliegende Konzept der *Kontentvalidität* KLAUER (1984) dar. Eine allgemeinere Anwendung von kriterienorientierten Tests stößt allerdings z. Z. noch auf eine Reihe von theoretischen und praktischen Schwierigkeiten.

Seit der Einführung des Army-α-Tests durch die US-Armee 1917 hat sich der Gruppentest in seiner ihm ureigensten Form als *„Papier- und Bleistift"-Test* in allen Ländern und Sprachen ein weites Feld erobert. Die *Reliabilität* dieser Tests (s. Kap. 10) ist in der Regel außerordentlich gut, ihre *Validität* (s. Kap. 11) jedoch oft begrenzt. Die neuere Testkonstruktion versucht auf verschiedene Weisen, dieser Schwierigkeiten Herr zu werden. So gewinnt z. B. das Konzept der *Inhaltsvalidität* eine immer größere Bedeutung (KLAUER 1984). Bei der Konstruktvalidierung von Tests, insbesondere von Intelligenztests, werden zunehmend komplexere Methoden, wie z. B. die Faktorenanalyse, eingesetzt. Das gesamte Validitätskonzept (s. Kap. 11) wird unter Zugrundelegung entscheidungstheoretischer Ansätze und Kosten–Nutzen-Überlegungen neu formuliert (CRONBACH und GLESER 1957). In diesem Zusammenhang wird auch die *„Testfairness"*[1]) untersucht (z. B. MÖBUS, 1983). Schließlich schränkt man den Geltungsbereich (s. Kap. 3) eines Tests ein, um seine Validität zu erhöhen.

Diesen Ausweg wählen zunehmend viele Praktiker, z. B. Lehrer in der Schule oder Psychologen in Betrieben oder Kliniken. Sie entwickeln sog. *„informelle Tests"* nur für ihre eigenen Zwecke. Diese Tests haben zwar einen stark eingeschränkten Geltungsbereich und sind meist nicht so reliabel wie käufliche Gruppentests. Sie besitzen aber oft eine höhere Validität, da sie exakt auf ein spezifisches Kriterium, z. B. einen ganz bestimmten Arbeitsplatz oder einen speziellen Lehrplan, hin entwickelt worden sind. Informelle Tests gewinnen in der Schule zunehmend an Bedeutung (WENDELER, 1969, RAPP, 1975, INGENKAMP, 1985).

Bei konventionellen Tests werden allen Pbn alle Aufgaben vorgelegt. Dieses Vorgehen ist insbesondere bei Pbn mit extremen Merkmalsausprägungen unökonomisch (warum soll ein besonders guter Pb auch alle leichten Aufgaben bearbeiten?), für die schlechteren Pbn im schwereren Testteil oft frustrierend, für die besseren Pbn in leichterem Testeil langweilig. Dieses Problem läßt sich durch das sog. *„Adaptive Testen"* lösen, bei dem jeder Pb andere Items entsprechend seinem Leistungsvermögen erhält (z. B. HORNKE 1977). Diese Technik erfordert jedoch in der Regel für die jeweilige Aufgabenauswahl, die Vorgabe und die Auswertung den Einsatz von PC's, was inzwischen auch bei uns kein Problem mehr sein dürfte.

---

[1]) Ein Test ist dann fair, wenn gleiche Testwerte bei Angehörigen verschiedener Gruppen zu den gleichen Entscheidungen führen.

Einen Überblick über die letzten 50 Jahre in der Geschichte der psychologischen Diagnostik gibt ANASTASI (1985). Eine lesenswerte Darstellung der Geschichte der verschiedenen Schulen, der Trends und der Strategien der Diagnostik findet man in dem Lehrbuch von JÄGER (1988).

Die testtheoretische Grundlage der Testkonstruktion bildet die sog. *Klassische Testtheorie*, die von GULLIKSEN (1950) entwickelt und u. a. später durch *entscheidungstheoretische Ansätze* ergänzt wurde. Von großer theoretischer, aber z. Z. noch von geringerer praktischer Bedeutung sind die *probabilistischen* Meß- bzw. Testtheorien, insbesondere die Modelle von RASCH, die seit der Mitte der 60er Jahre diskutiert und angewendet werden (z. B. LORD u. NOVICK 1968, FISCHER 1968, 1974, KUBINGER in Jäger 1988), neuerdings auch das „klassische latent-additive Testmodell" (KLA-Modell) von MOOSBRUGGER und MÜLLER (1981). Die Klassische Testtheorie, sowie die daraus abgeleiteten Prinzipien der Testkonstruktion sind in den letzten Jahren bei uns in Handbuchartikeln, Monographien und in einzelnen Kapiteln von diagnostischen Lehrbüchern mehr oder weniger ausführlich dargestellt worden. Hier sind über die bereits genannten Autoren hinaus z. B. noch MICHEL (1964), VUKOVICH (1964), MAGNUSSEN (1969), GUTJAHR (1971), HORST (1971), DIETERICH (1973), SARRIS und LIENERT (1974), GRUBITZSCH und REXILIUS (1978), LEICHNER (1979), KRANZ (1979), WOTTAWA (1980) SCHELTEN (1980) oder MOOSBRUGGER (1988) zu nennen.

Klassische Testtheorie und *Probabilistische Testtheorien* bestehen heute nebeneinander, sie ergänzen sich (z. B. KLAPPROTH, 1981). Die meisten Tests und Fragebogenverfahren werden immer noch „klassisch" konstruiert, obwohl die theoretischen Schwächen dieses Ansatzes seit langem bekannt sind (FISCHER 1968, 1974). Probabilistische Ansätze werden nur zur Überprüfung der Eindimensionalität oder beim Aufbau eines Aufgabenpools herangezogen. Beim Adaptiven Testen sind sie unabdingbar.

Die große Zahl von Veröffentlichungen zur Testtheorie und Testkonstruktion darf aber nicht darüber hinwegtäuschen, daß in der diagnostischen Praxis leider immer noch eine große Zahl von Verfahren „zusammengebastelt" werden. Hier sind nach wie vor Warnung, Aufklärung und Anleitung notwendig.

1.1.3 Die Bestandteile eines Tests

a) Die **Materialbestandteile** eines Tests, wie sie bei standardisierten Tests vorliegen, setzen sich zusammen aus:

- dem *Testmanual*, in dem Beschreibung, Entstehung und wissenschaftliche Grundlage des Tests dargestellt, sowie Durchführungs- und Interpretationsanweisungen samt Normen enthalten sind,
- dem *Testmaterial*, das Testhefte, Antwortbogen, Material zur Bearbeitung und Verarbeitung umfaßt,
- den *Auswertungshilfen*, wie Kontrollblätter, Lochfolien, Schablonen, Meßgeräte usw.

b) Die **Durchführungsbestandteile** eines Tests lassen sich methodisch und zugleich testchronologisch wie folgt gliedern:

- Die *Testanweisung* instruiert den Testleiter über Bedingungen, Durchführung und Auswertung des Tests und den Probanden (Pb) darüber, was von ihm als Testleistung verlangt wird.
- Die *Testdurchführung* kann in einer motorischen Reaktion (z. B. Tapping), einer sensorischen Apperzeption (z. B. Tachistoskop-Auffassung), einer manuellen oder geistigen Routinetätigkeit (z. B. Perlen-Reihen, Rechnen nach PAULI), einer konkreten oder abstrakten Problemlösung (z. B. SCHULZE-Pumpe aufbauen, Denkaufgaben lösen) oder in einer Stellungnahme zu einer Verhaltens-, Erlebnis- oder Einstellungsfrage bestehen. Die Durchführung kann sich auf eine einzige oder auch auf viele Aufgaben beziehen. Die Aufgaben ihrerseits können einen einfachen oder komplexen Inhalt haben, ihre Schwierigkeit kann annähernd gleich oder auch sehr verschieden sein.
- Die *Testauswertung* kann intuitiv, regelmäßig, schematisch oder automatisch erfolgen und besteht darin, daß das Ergebnis der Testdurchführung in einem oder mehreren Punktwerten gekennzeichnet wird.

1.1.4 Die Aufgaben des Tests

Ein Test im Sinne der von uns gegebenen Definition hat folgende Aufgaben zu erfüllen:

a) Als Hilfsmittel im Dienste der Querschnittsdiagnose.

- Kennzeichnung der Stellung des Einzelindividuums innerhalb einer Gruppe vergleichbarer Individuen hinsichtlich einer bestimmten *Leistung* oder eines bestimmten *Persönlichkeitsmerkmals*.
- Feststellung von Unterschieden hinsichtlich der *Gradausprägung* der Leistung oder des Persönlichkeitsmerkmals,
  zwischen verschiedenen Individuen,
  zwischen verschiedenen Gruppen.
- Entscheidung über *Erfüllung* oder *Nichterfüllung* einer Bedingung, über Vorhandensein oder Nichtvorhandensein eines geforderten Merkmals oder über den Grad einer Merkmalsausprägung (Auslese).
- Feststellung individueller Merkmalskombinationen (Persönlichkeits- oder Leistungsprofil).

b) Als Hilfsmittel im Dienste der Längsschnittdiagnose.

Feststellung von *Merkmalsveränderungen* innerhalb definierter Zeitspannen (Verlaufsprofil),
  bei Einzelindividuen,
  bei Gruppen, besonders Altersgruppen.

c) Als Hilfsmittel der Forschung[1]).

- Im Dienste der *Persönlichkeitsforschung*.

---

[1]) In dieser Anwendung dient der Test als Hilfsmittel zur Datenerhebung z. B. bei Experimenten, ist aber selbst kein Experiment (KLAUER, 1973). Denn es fehlt die Bedingungsvariation.

a) Feststellung „Lokaler" Merkmalszusammenhänge (Eigenschaftstrauben, Persönlichkeitstypen) durch Typenanalyse[1]).
b) Analyse „globaler" Merkmalszusammenhänge (Cluster- und Faktorentechniken).

– Im Dienste der *allgemeinpsychologischen Forschung.*
Feststellung von Merkmalsveränderungen unter planmäßig variierten Bedingungen,
bei Einzelindividuen (Versuchsreihe),
bei Gruppen (Planversuch).

Die Aufgaben des primär diagnostischen Zielen dienenden Tests reichen also weit in das Gebiet der wissenschaftlichen Forschung hinein, weshalb ein Test als Meßinstrument für psychologische Merkmale gewissen, in der Folge behandelten, Forderungen genügen muß.

1.2 Die Gütekriterien eines Tests

Ein guter Test soll als *Hauptgütekriterien* folgende drei Forderungen erfüllen:
– er soll objektiv,
– er soll reliabel,
– er soll valide sein.[2])

Daran schließen sich vier *Nebengütekriterien* als bedingte Forderungen:
– er soll normiert,
– er soll vergleichbar,
– er soll ökonomisch,
– er soll nützlich sein.

1.2.1 Die Objektivität eines Tests

Unter *Objektivität* eines Tests verstehen wir den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind. Ein Test wäre demnach vollkommen objektiv, wenn verschiedene Untersucher bei denselben Pbn zu gleichen Ergebnissen gelangten. Man spricht deshalb auch von „interpersoneller Übereinstimmung" der Untersucher[3]).

---

[1]) Eine detailliert ausgearbeitete Methode der Typenanalyse ist die Konfigurationsfrequenzanalyse (KFA) (LIENERT, 1968). Sie operationalisiert Typen (Syndrome) als überzufällig oft auftretende Kombinationen von Merkmalen (Symptomen). Eine zusammenfassende Darstellung gibt KRAUTH (1993). Weitere Arbeiten zur KFA sind bei LIENERT (1988), LAUTSCH und v. WEBER (1990) und von EYE (1990) zusammengestellt.

[2]) Reliabilität und Validität können als Sonderfälle experimenteller Präzision verstanden werden (LIENERT und ORLIK, 1965).

[3]) Der Begriff der Objektivität wird in der anglo-amerikanischen und deutschen Literatur recht unterschiedlich gefaßt. Über zahlreiche mögliche Definitionen orientiert SCHEIER (1958).

Als Maß für die interpersonelle Übereinstimmung oder Objektivität verschiedener Untersucher könnte der durchschnittliche Korrelationskoeffizient zwischen den durch verschiedene Untersucher an einer Stichprobe von Pbn erhobenen Testbefunden gelten. Je nachdem, in welcher Phase der Testdurchführung allfällige Nicht-Übereinstimmungen auftreten bzw. ausgelöst werden, unterscheidet man verschiedene *Aspekte* der Objektivität.

a) Die sog. **Durchführungsobjektivität** betrifft den Grad der Unabhängigkeit der Testergebnisse von zufälligen oder systematischen Verhaltensvariationen des Untersuchers während der Testdurchführung, die ihrerseits zu Verhaltensvariationen des Pbn führen und dessen Ergebnis beeinflussen. Soll die Durchführungsobjektivität maximal hoch werden, dann muß die Instruktion an den Untersucher (schriftlich) so genau wie möglich festgelegt und die Untersuchungssituation so weit wie möglich standardisiert werden. Das läuft in aller Regel darauf hinaus, die soziale Interaktion zwischen Untersucher und Pb auf ein unumgängliches Minimum zu reduzieren (bekanntlich ist dies bei projektiven Tests nicht im gleichen Maße wie bei Leistungstests möglich).

b) Die sog. **Auswertungsobjektivität** betrifft die numerische oder kategoriale Auswertung des registrierten Testverhaltens nach vorgegebenen Regeln. Bei Leistungstests und Fragebogen, in welchen die *Schlüsselrichtung* der Aufgabenbeantwortung („Richtig" oder „Falsch") festliegt, ist die Auswertungsobjektivität praktisch vollkommen verwirklicht; das gilt zumindest für Leistungstests mit Auswahlbeantwortung, in welchen der Pb lediglich diejenige Antwort von mehreren vorgegebenen Antworten, die ihm als richtig oder auf ihn zutreffend erscheint, anzukreuzen hat. Etwas geringer ist die Auswertungsobjektivität bei Leistungstests oder Fragebogen mit *freier* Aufgaben- oder Fragebeantwortung, da hier der Untersucher bzw. Auswerter entscheiden muß, ob eine frei gegebene Antwort in die Schlüsselrichtung weist oder nicht. Am relativ niedrigsten ist die Auswertungsobjektivität im allgemeinen bei projektiven Tests, z. B. bei der Beantwortung eines Rorschach-Tests, wo die kategoriale Zuordnung der Antworten („Signierung") durch verschiedene Auswerter i. a. nur unvollkommen übereinstimmt.

c) Die sog. **Interpretationsobjektivität** betrifft den Grad der Unabhängigkeit der Interpretation des Testergebnisses von der Person des interpretierenden Psychologen, der nicht mit dem Untersucher oder Auswerter identisch zu sein braucht. Die Interpretationsobjektivität ist dann gegeben, wenn aus den gleichen Auswertungsergebnissen (verschiedener Pbn) *gleiche* Schlüsse gezogen werden. Die Interpretationsobjektivität ist vollkommen und zugleich trivial, wenn es sich um normierte Leistungstests oder Fragebogen handelt, in welchen die Auswertung einen numerischen Wert liefert, der die Position des Pb entlang der Testskala festgelegt; sie ist nicht trivial, wenn – wie bei projektiven Tests – die Auswertung nur als notwendige und nicht zugleich auch als hinreichende Voraussetzung für eine Aussage oder eine Entscheidung über den Pb gelten kann. Hier wird die Interpretation umso objektiver sein, je genauer die Angaben des Testautors in der Handanweisung den Interpreten lenken und in seinen Freiheitsgraden einengen.

Je nach den verschiedenen Aspekten der Objektivität muß sie ggfs. in entsprechender Weise erfaßt werden: So wäre die Durchführungsobjektivität zu bestimmen, indem eine Stichprobe von Pbn von zwei (oder mehreren) Unter-

suchern getestet und die aufgrund einer absolut objektiven Auswertung resultierenden Ergebnispaare korreliert werden. Die Auswertungsobjektivität wäre dadurch zu prüfen, daß man die Antworten einer Stichprobe von Pbn zwei (oder mehreren) Auswertern zur Rohwertbestimmung übergibt und die unabhängig erhaltenen Rohwertpaare korreliert. Die Interpretationsobjektivität schließlich wäre zu bestimmen, indem man dem Interpreten einschlägige Interpretationskategorien zur Auswahl stellt und prüft, ob zwei (oder mehrere) Interpreten gleich zuordnen. Als Objektivitätsmaß wäre hier der Kontingenzkoeffizient einer aus den Interpretationskategorien gebildeten Kontingenztafel oder Kappa (LIENERT 1978) zu ermitteln.

1.2.2 Die Reliabilität eines Tests

Unter der *Reliabilität* oder Zuverlässigkeit eines Tests versteht man den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal mißt, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht (welche Frage ein Problem der Validität ist).

Ein Test wäre demnach vollkommen reliabel, wenn die mit seiner Hilfe erzielten Ergebnisse den Pb genau, d. h. fehlerfrei beschreiben bzw. auf der Testskala lokalisieren. Diese Genauigkeit betrifft lediglich den beobachteten Meßwert und nicht auch seinen Interpretationswert, also nicht die Frage, ob er auch das mißt, was er messen soll.

Der Grad der Reliabilität wird durch einen *Reliabilitätskoeffizienten* bestimmt, der angibt, in welchem Maße unter gleichen Bedingungen gewonnene Meßwerte über ein und denselben Pbn übereinstimmen, in welchem Maße also das Testergebnis reproduzierbar ist.

Wie bei der Objektivität, so müssen auch bei der Reliabilität eines Tests verschiedene *Aspekte* unterschieden werden. Jeder dieser Aspekte schließt (wie bei der Objektivität) einen anderen operationalen Zugang zu seiner Bestimmung mit ein. Damit ist gesagt, daß die Reliabilität eines Tests per se nicht existiert, sondern nur verschiedene ihrer methodischen Zugänge[1]). Diese Zugänge bzw. die ihnen entsprechenden Begriffe der Reliabilität sind:

a) Die **Paralleltest-Reliabilität**: Sie wird in der Weise bestimmt, daß einer Stichprobe von Pbn zwei miteinander streng vergleichbare Tests (Paralleltests) vorgelegt und deren Ergebnisse korreliert werden *(Paralleltest-Methode)*.

b) Die **Retest-Reliabilität** wird mittels der Testwiederholungsmethode bestimmt: Man gibt ein und denselben Test einer Stichprobe von Pbn zweimal vor und ermittelt die Korrelation der beiden Ergebnisreihen *(Retest-Methode)*.

c) Die **innere Konsistenz** eines Tests[2])

– Nach der Methode der **Testhalbierung**. Die Halbierungsreliabilität oder -konsistenz gewinnt man durch folgendes Vorgehen: Ein Test wird

---

[1]) Besonders CRONBACH (1947) hat hierauf in einem lesenswerten Aufsatz hingewiesen.

[2]) Den theminologischen Empfehlungen der APA (1986) folgend, werden hier die Methoden der Testhalbierung und der Konsistenzanalyse der „inneren Konsistenz" eines Test zugeordnet.

einer Stichprobe von Pbn vorgegeben, und zwar nur ein einziges Mal. Dann werden die Elemente (Aufgaben, Fragen, Feststellungen) des Tests in zwei gleichwertige Hälften geteilt („gesplittet") und das Testergebnis eines einzelnen Pb für jede Testhälfte gesondert ermittelt (Split-half-Methode).
Schließlich werden die Testergebnisse der beiden Hälften korreliert und der (für den halbierten Test geltende) Reliabilitätskoeffizient so aufgewertet, daß er für den ganzen Test Geltung beanspruchen kann (Halbierungs- oder Split-Half-Konsistenzkoeffizient).

– Nach der Methode der **Konsistenzanalyse**. Hier handelt es sich darum, die Elemente eines Tests als multipel halbierte Testteile aufzufassen, und die Reliabilität über bestimmte Kennwerte dieser Testelemente (Aufgabenschwierigkeits- und Trennschärfenstatistiken) auf indirektem Wege zu ermitteln und nicht wie oben durch Korrelation von Testwertepaaren. Wir werden auf diese Methode später genauer eingehen.
Wendet man auf ein und denselben Test alle vier genannten Methoden der Reliabilitätsbestimmung an, so erhält man in der Regel numerisch mehr oder weniger unterschiedliche Reliabilitätskoeffizienten, wobei freilich – wie wir noch sehen werden – bestimmte Gesetzmäßigkeiten zwischen diesen Koeffizienten bestehen. Immerhin ist allen Methoden gemeinsam, daß sie Schätzwerte für den Anteil liefern, in dem ein einzelner Testwert fehlerbehaftet ist oder sein kann. Die numerischen Unterschiede resultieren daraus, daß jede der vier Methoden mehr oder weniger spezifische *Meßfehlerarten* und *-anteile* berücksichtigt.

1.2.3 Die Validität eines Tests

Die *Validität* oder Gültigkeit eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die)er messen oder vorhersagen soll, tatsächlich mißt oder vorhersagt.

Ein Test ist demnach vollkommen valide, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluß auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmals zulassen, wenn also der individuelle Testpunktwert eines Pb diesen auf der Merkmalsskala eindeutig lokalisiert.

Auch bei der Validität können verschiedene *Aspekte* unterschieden werden, worüber ausführlich in Kap. 11 diskutiert wird. Hier seien drei grundsätzliche Unterscheidungen vorangestellt.

a) **Inhaltliche Validität** (Content validity). Der Test bzw. seine Elemente sind so beschaffen, daß sie das zu erfassende Persönlichkeitsmerkmal oder die in Frage stehende Verhaltensweise repräsentieren, mit anderen Worten: Der Test selbst stellt das optimale Kriterium für das Persönlichkeitsmerkmal oder die Verhaltensweise dar. Beispielsweise würde ein Schulkenntnistest, etwa in Geographie, dann inhaltlich valide sein, wenn seine Aufgaben inhaltlich eine repräsentative Auswahl aus dem Unterrichtsstoff (Curriculum) darstellen. Ebenso ist es evident, daß eine Schreibprobe inhaltlich valide ist für die Erfassung der Schnelligkeit und Genauigkeit, mit der eine Sekretärin Maschine schreiben kann.

Inhaltliche Validität wird einem Test in der Regel durch ein Rating von Experten als „Konsens von Kundigen" zugebilligt.

b) **Konstruktvalidität.** Aufgrund theoretischer – sachlogischer und begrifflicher – Erwägungen und anhand von sich daran anschließenden empirischen Untersuchungen wird entschieden, ob ein Test ein bestimmtes Konstrukt zu erfassen vermag. Die Konstruktvalidität zielt direkt ab auf die psychologische Analyse der einem Test zugrunde liegenden Eigenschaften und Fähigkeiten, also auf Beschreibungsmerkmale, die nicht in eindeutiger Weise operational erfaßbar, sondern theoretischen Charakter haben, wobei freilich eine empirische Basis gegeben ist. Ein Test etwa, der den individuellen Ausprägungsgrad von Angst messen soll, hätte dann eine hinreichende Konstruktvalidität, wenn nachgewiesen wurde, daß das vom Test erfaßten Merkmal in genügender Übereinstimmung mit dem theoretischen Konstrukt „Angst" steht.

c) Die **kriterienbezogene** Validität (criterion validity). Während sich im Fall der beiden oben genannten Validitätsaspekte im allgemeinen keine Maßzahl für den Grad der Validität eines Tests ermitteln und angeben läßt, ist dies bei der kriterienbezogenen Validität dadurch möglich, daß man die Testergebnisse einer Stichprobe von Pb mit einem sog. *Außenkriterium* korreliert, das vom Test unabhängig erhoben wird. Man kann dann entweder direkt von der Testleistung auf die Kriterienleistung schließen, oder man kann das Kriterium als einen ausreichend validen Repräsentanten für das Persönlichkeitsmerkmal, das es zu erfassen gilt, ansehen und so indirekt Aussagen über dieses Merkmal machen.

Betrachten wir nur die kriterienbezogene Validität eines Tests bzw. dessen *Validitätskoeffizienten*, so hängt des letzteren Höhe im wesentlichen von drei Faktoren oder Einflußgrößen ab; und zwar

- von dem Grad dessen, was an „Gemeinsamkeit" durch den Test und das Kriterium erfaßt und oft als „Zulänglichkeit" des Tests bezeichnet wird, wobei von der Reliabilität des Tests und des Kriteriums abgesehen wird.[1])
- von der Reliabilität des Tests und
- von der Reliabilität des Kriteriums.

Je größer die Gemeinsamkeit des von Test und Kriterium erfaßten Merkmalsanteils, umso größer ist die kriteriumsbezogene Validität eines Tests. Andererseits ist diese umso geringer, je geringer die Reliabilitätskoeffizienten ausfallen bzw. zu veranschlagen sind.

1.2.4 Die Normierung eines Tests

Unter den Güte-Kriterien der *Normierung* versteht man, daß über einen Test Angaben vorliegen sollen, die als Bezugssystem für die Einordnung des individuellen Testergebnisses dienen können. Danach werden die Ergebnisse verschiedener Tests vergleichbar. Bei eindimensionalen Tests läuft die Normierung schlicht darauf hinaus, daß man die *Verteilung* der Testwerte einer Population

[1]) Ein Koeffizient zur Kennzeichnung der „Zulänglichkeit" ist nicht gebräuchlich.

von Pbn zu ermitteln trachtet, diese Verteilung nötigenfalls normaltransformiert und sie mit bestimmten Verteilungsparametern (z. B. bei der Intelligenzquotientenskala mit einem Durchschnitt von 100 und einer Standardabweichung von 15) ausstattet. Zu jedem sog. Testrohwert gehört dann ein bestimmter Teststandardwert, der die Position eines Pbn eindeutig fixiert und unabhängig vom Schwierigkeitsgrad und von der Aufgabenzahl des Tests ist.

Die Normierung kann für die Gesamtpopulation eines Areals *(Gesamtnormen)*, für die Population einer bestimmten sozialen Gruppe *(Gruppennormen)* oder für mehrere solche Populationen erfolgen. Es ist einleuchtend, daß ein Test, der zwar die Hauptgütekriterien erfüllt, aber nicht normiert ist, keine oder nur sehr geringe diagnostische Brauchbarkeit besitzt, wiewohl er als Forschungsinstrument (wo es in aller Regel nur um den Vergleich von Gruppen geht) durchaus geeignet erscheint.

1.2.5 Die Vergleichbarkeit eines Tests

Ein Test ist dann *vergleichbar*, wenn:

a) ein oder mehrere Paralleltestformen vorhanden sind,
b) validitätsähnliche Tests verfügbar sind.

Die Parallelform eines Tests gestattet gewissermaßen einen Vergleich des Tests mit sich selbst. Sie ermöglicht eine intra-individuelle Reliabilitätskontrolle, indem man einen bestimmten Pb mit beiden Testformen untersucht und die Ergebnisse vergleicht.

Validitätsgleiche oder -ähnliche Tests prüfen dasselbe oder ein nahe verwandtes Persönlichkeitsmerkmal. Wenn die Korrelation zwischen zwei validitätsähnlichen Tests bekannt ist, so ist eine intraindividuelle Validitätskontrolle möglich, indem man den gleichen Pb mit diesen beiden Tests untersucht und die Ergebnisse vergleicht.

1.2.6 Die Ökonomie eines Tests

Ein Test ist dann *ökonomisch*, wenn er:

- eine kurze Durchführungszeit beansprucht,
- wenig Material verbraucht,
- einfach zu handhaben,
- als Gruppentest durchführbar,
- schnell und bequem auszuwerten ist.

Ein Test ist demgemäß sehr ökonomisch, wenn er alle oder die wichtigsten dieser Bedingungen erfüllt, und er ist minder ökonomisch oder sogar unökonomisch, wenn er nur einen Teil oder keine der genannten Voraussetzungen erfüllt.

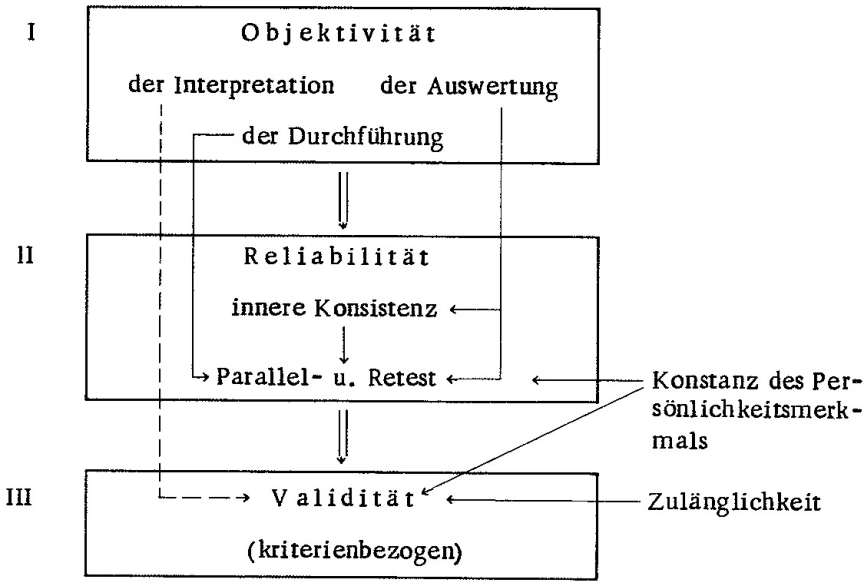
Wie für alle übrigen Nebenkriterien, so gibt es auch für die Ökonomie eines Tests keinen zahlenmäßigen Kennwert.

1.2.7 Die Nützlichkeit eines Tests

Ein Test ist dann *nützlich*, wenn er ein Persönlichkeitsmerkmal oder eine Verhaltensweise mißt oder vorhersagt, für dessen Untersuchung ein praktisches Bedürfnis besteht. Ein Test hat demgemäß eine hohe Nützlichkeit, wenn er in seiner Funktion durch keinen anderen Test vertreten werden kann, und er hat eine geringe Nützlichkeit, wenn er ein Persönlichkeitsmerkmal prüft, das mit einer Reihe anderer Tests ebenso gut untersucht werden könnte.

1.2.8 Die Wechselbeziehungen zwischen den Gütekriterien

Ein vereinfachendes Schema der wechselseitigen Abhängigkeit der ersten beiden Gütekriterien und der kriterienbezogenen Validität würde wie folgt aussehen:



Aus dem obigen Schema lassen sich folgende Regeln[1]) für die Beziehungen zwischen den Gütekriterien ableiten:

a) Die Parallel- oder Retest-Reliabilität eines Test kann nicht höher sein als seine Konsistenz oder seine Objektivität; ferner ist ein Test – kriterienbezogen – nicht valider als er reliabel ist.

b) Ein Test mit einer hohen kriterienbezogenen Validität muß notwendigerweise auch hohe Objektivität, Konsistenz und Zulänglichkeit besitzen. Die Feststellung einer hohen kriterienbezogenen Validität entbindet somit in gewissem Maße von der Überprüfung der übrigen Gütekriterien.

[1]) Diese Beziehungen gelten in der Theorie, in der Praxis sind jedoch Ausnahmen möglich.