Psychologische Testverfahren

Psychologische Testverfahren

Helmut Lukesch

1. Einleitung

Unter psychologischer Diagnostik werden „alle Methoden (verstanden), welche zur Messung bzw. Beschreibung inter- und intraindividueller Unterschiede verwendet werden" (Dorsch, 1982, S.528). Die Vielfalt der verwendeten Verfahren läßt sich dabei in vier Grobbereiche gliedern, und zwar in (1) Gesprächsmethoden (Anamnese, Exploration, Interview), (2) Beobachtungsverfahren, die nach Mees (1977, S.1f.) wieder in Verfahren der isomorphen und der reduktiven Deskription unterteilt werden können, wobei im letzteren Falle zwischen Zeichensystemen (hier werden nur einzelne relevante Verhaltensweisen in ihrer Auftretenshäufigkeit festgehalten) und Kategoriensystemen (hier werden alle auftretenden Verhaltensweisen einer Beobachtungskategorie zugewiesen) unterschieden wird (Cranach & Frenz, 1969, S.272), (3) Beurteilungsmethoden (Fremd- und Selbstbeurteilungsverfahren) sowie in (4) Testmethoden. Obzwar die situationalen Bedingungen, unter denen die jeweiligen diagnostisch relevanten Daten erhoben werden, höchst unterschiedlich sein können (z.B. hinsichtlich der Reaktivität vs. Spontaneität, der Aktualität oder der Voraussetzungen hinsichtlich des zu erfassenden Verhaltens (vgl. Webb et al., 1966), gilt für alle psychologischen Diagnoseverfahren der gleiche Kanon an Gütekriterien, dessen zumindest graduelle Erfüllung in nachvollziehbarer Weise gewährleistet sein muß, bevor das Verfahren verantwortlich angewendet werden kann. Bei psychologischen Tests ist die Einhaltung dieser Gütekriterien im allgemeinen eher gegeben als bei anderen Verfahren.

Um den Unterschied zwischen Forschung (Erkenntnis allgemeiner Zusammenhänge) und Diagnostik (Klassifikation eines Einzelfalls) besser zu verstehen, kann man sich das allgemeine Schema für wissenschaftliche Erklärungen in Erinnerung rufen (Stegmüller, 1969). Nach dem Hempel-Oppenheimschen Schema werden aus allgemeinen Gesetzen (G_1 bis G_n) und gegebenen Rand- oder Antezedensbedingungen (A_1 bis A_n) Ereignisse deduktiv erschlossen oder eben erklärt. Im Fall deterministischer Gesetze folgt das Ereignis mit Notwendigkeit, im Fall probabilistischer Gesetze ist der Schluß, daß das Ereignis eintritt, nur mit einer gewissen Wahrscheinlichkeit wahr.

$G_1, G_2 \ldots G_n$
$A_1, A_2 \ldots A_n$

E_i

Bereits die forscherische Gewinnung bzw. Prüfung solcher allgemeiner Gesetze setzt diagnostische Verfahren voraus (i.S. der Operationalisierung von Indikatoren für „Intelligenz", „Begabung", „Lernbereitschaft" etc.). Im Falle der psychologischen Diagnostik wird aber nach dem Einzelereignis E_i bzw. seiner genaueren Beschreibung gefragt. In anderen Fällen ist das Einzelereignis E_i gegeben (z.B. „Schulversagen", „berufliche Bewährung"), wird allenfalls z.B. durch Testverfahren nochmals erhärtet, und die Suche richtet sich nach den Rand- oder Antezedensbedingungen $A_1, A_2 \ldots A_n$, um dieses Ereignis entsprechend erklären zu können. Auch diese Antezedensbedingungen müssen wiederum diagnostisch abgeklärt werden. Im psychologischen Kontext wird zumeist nicht bei einer bloßen Konstatierung von Ergebnissen stehengeblieben, sondern aufgrund der Testdaten werden weitere Maßnahmen im Sinne einer anderen Plazierung und Selektion bzw. Förderung und Modifikation getroffen (d.h. diagnostische Erkenntnisbemühungen stehen in einem breiteren Anwendungskontext, begründen also Entscheidungen; vgl. Rollett, 1976, S.139).

2. Tests in historischer und systematischer Sicht

Ein Test selbst kann im allgemeinsten Sinn als eine Verhaltensstichprobe aufgefaßt werden, aufgrund der ein bestimmtes anderes Verhalten erwartet wird (Ekman, 1955). Solche Verhaltensstichproben gibt es schon seit sehr langer Zeit (Hofstätter, 1971; Heiss, 1964, S.5). Bei den frühen Methoden (z.B. Hexenprobe, Gottesurteile, Pubertätsriten zur Aufnahme in die Erwachsenengesellschaft) ist es für heutige Menschen nur schwer nachvollziehbar, wie

es zu der Zuordnung zwischen Verhaltensstichprobe und dem von dieser erwarteten Verhaltensfolge kam. Erst im 19. Jahrhundert entstand jene Art von Testpsychologie, deren methodische Prinzipien auch heute noch akzeptiert werden.

Da am Anfang der wissenschaftlichen Psychologie (im Leipziger Labor Wilhelm Wundts um 1879) das Experiment steht, muß auch die eigentliche Testpsychologie von den Intentionen der Gründer jener Methodik verstanden werden. Individuelle Unterschiede im Verhalten wurden zuerst als „Fehlerquellen" bei der Erforschung allgemeiner Gesetzlichkeiten angesehen, bis schließlich James McKeen Cattell (1890), ein Schüler Wundts, diese Unterschiede zu seinem originären Forschungsinteresse machte; er gilt denn auch als Begründer der Testpsychologie (von ihm stammt z.B. der Begriff „mental test"). Allerdings war nicht nur die experimentelle Psychologie für die Entstehung der Testpsychologie verantwortlich, sondern wesentliche Anregungen kamen auch aus dem Bereich der Psychiatrie (z.B. Emil Kraepelin, 1895). Ferner ist aufgrund von Fragestellungen in der Genetik diese Entwicklung vorangetrieben worden; Francis Galton (1870), ein Cousin von Charles Darwin, wollte z.B. mit Testmethoden nachweisen, daß für die Erblichkeit psychischer Eigenschaften dieselben Gesetzmäßigkeiten gelten wie für die Vererbung körperlicher Merkmale. Letztendlich waren es auch schulbezogene Anwendungsfragen, die zum Entstehen der Testpsychologie beigetragen haben. Zum Beispiel war Herrmann Ebbinghaus (1897) vom Breslauer Magistrat beauftragt worden zu prüfen, inwieweit der Vormittagsunterricht die Schüler belaste; dabei erfand er als Zufallsergebnis in Form eines Lückentextes den ersten Intelligenztest. Ebenfalls aus anwendungsorientierten Fragen hat sich das Intelligenzdiagnostikum von Alfred Binet entwickelt (Unterscheidung von normalen und schwachsinnigen Kindern, um sie dann in der jeweils für sie geeigneten Schulform unterzubringen; Binet & Henri, 1898).

Gegen Ende des vorigen Jahrhunderts beginnt eine Springflut der Testproduktion. Einmal entwickelt sich die Methode der Intelligenzprüfung in breiterer Form, zum anderen wird die Methode der Testprüfung auch auf andere Gebiete ausgeweitet. So sind etwa die ersten Anfänge der sogenannten „Psychotechnik" darauf ausgerichtet, spezielle Begabungen, Berufs- und Arbeitseignung zu bestimmen. Großen Auftrieb erhält diese Entwicklung durch den Ersten Weltkrieg. Amerika hat 1917 in großem Stil begonnen, Soldaten mit Testuntersuchungen auszuwählen. Mit den „Group Examinations Alpha and Beta" wurden damals etwa 2 Millionen Rekruten untersucht (Eysenck, 1962). In der Folge der anfänglichen Testeuphorie entwickelte sich eine Unzahl von Verfahren, die allerdings nur bedingt diagnostischen Gütekriterien entsprachen. Schätzungen der Zahl vorhandener Tests sind abhängig davon, was alles als „Test" gezählt wird. Bei einer eher großzügigen Interpretation des Testbegriffes werden für die USA z.Zt. etwa 2.500 aktuelle Testverfahren gezählt (Sweetland & Keyser, 1986). Im deutschen Sprachbereich sind vor allem die Dokumentationen von Brickenkamp (1975; 1983) zu erwähnen, in denen 219 bzw. weitere 164 Tests, die über Verlage zu beziehen sind, beschrieben sind. Auch das Skalenhandbuch der ZUMA (1983) enthält für Anwender wichtige Informationen über 120 zumeist unveröffentlichte Skalen. Ein universalistischer Anspruch lag der Entwicklung der Datenbank über psychologische und pädagogische Testverfahren (PSYTKOM) zugrunde; diese enthält z.Zt. etwa 2.300 Verfahren und ist nach den üblichen Kriterien über die ZPID nutzbar (Lukesch, 1992).

Was bei einer schärferen begrifflichen Fassung unter einem Test verstanden werden soll, definiert Lienert (1967, S.7) wie folgt: *„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."* Die einzelnen Bestimmungsstücke, die in dieser Definition enthalten sind, können nach Selg und Bauer (1971, S.66ff.) noch genauer erläutert erläutert werden:

1. *wissenschaftlich*: Tests müssen den diagnostischen Gütekriterien genügen (s. u.), d. h. nicht jede beliebige zu diagnostischen Zwecken angestellte Untersuchung kann als Test gelten. Die Wissenschaftlichkeit des Verfahrens wird durch die Erfüllung der Gütekriterien gesichert.

2. *Routineverfahren*: Gemeint ist damit, daß ganz bestimmte Maßnahmen und Testauswirkungen wiederkehren, z.B. eine festgelegte Instruktion. Dadurch versucht man, dem Experiment ähnliche Standardbedingungen zu schaffen. Es soll damit auch gewährleistet sein, daß der Test mehr oder weniger handwerksmäßig durchgeführt werden kann, z.B. von einer Hilfskraft.

3. Relativer Grad der individuellen Merkmalsausprägung: Dies bedeutet, daß eine relative Positionsbestimmung des untersuchten Individuums innerhalb einer Gruppe von Individuen möglich sein muß; man muß z.B. angeben können, wie ein Proband im Verhältnis zum Durchschnittswert seiner Gruppe liegt. Ergänzend hierzu wird im Rahmen einer kriteriumsorientierten Messung (z.B. im kli-

nischen oder pädagogischen Kontext) der Abstand zu einem vorgegebenen Ziel zu erfassen versucht (Klauer, 1987).

4. *Empirisch abgrenzbare Persönlichkeitsmerkmale*: Es werden hierunter solche verstanden, die verhaltens- oder erlebnisanalytisch oder phänomenologisch abgrenzbar sind. Ein Test kann nicht vage definierte Eigenschaften (z.B. Gemütstiefe) erfassen, sondern nur Merkmale, die sich als beobachtbar und objektiv beschreibbar erwiesen haben.

5. *Möglichst quantitative Aussage*: Die Quantifizierung von Merkmalen erlaubt logisch aufgebaute und prägnante Aussagen, wobei die Vermittlung an den Adressaten eines Testergebnisses wieder beträchtliche Ansprüche an die sprachliche Umsetzung dieser quantitativen Ergebnisse stellen kann (Hartmann, 1970).

6. *Untersuchung eines oder mehrerer Persönlichkeitsmerkmale*: Damit ist gesagt, daß mit einem Test nie „alle" Merkmale einer Person untersucht werden. Selg und Bauer (a.a.O.) bezeichnen Tests, die einen solchen Anspruch erheben, als Anachronismen. In Einzelfällen können allerdings auch Diagnosen gemacht werden, die über den primären Zweck eines Tests hinausgehen (z.B. klinische Diagnosen mittels HAWIE).

3. Gütekriterien diagnostischer Verfahren

Die im folgenden zu diskutierenden Gütekriterien wurden im Hinblick auf psychologische Tests entwickelt. Im Prinzip müßte aber von jeder diagnostischen Methode verlangt werden, daß sie diese Kriterien erfüllt, denn diese sind in der Summe nichts anderes als die Forderungen, die man an wissenschaftlich begründetes Arbeiten – auch wenn dieses in einem Anwendungskontext erfolgt – stellt.

(1) *Objektivität*: „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 Probanden zu gleichen Ergebnissen gelangten. Man spricht deshalb auch von ‚interpersoneller Übereinstimmung' der Untersucher" (Lienert, 1967, S.13). Bezogen auf die verschiedenen Phasen des diagnostischen Prozesses sollte Objektivität für die *Durchführung der Untersuchung, die Auswertung der Daten und die Interpretation der Ergebnisse gelten.*

(2) *Reliabilität* (Zuverlässigkeit, Genauigkeit): „Unter der Reliabilität eines Tests versteht man den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal mißt" (Lienert, 1967, S.14). Anders formuliert, betrifft die Reliabilität den Grad der formalen Meßgenauigkeit von Testwerten, und dies unabhängig davon, ob die Testwerte auch tatsächlich Rückschlüsse auf das diagnostisch Intendierte zulassen.

Bei methodenkritischen Untersuchungen fand man bald heraus, daß die Meßwerte in der Psychologie meist nicht die Genauigkeit besitzen, die man bei Messungen makrophysikalischer Größen erhält. Dies war Anlaß zur Entwicklung der sogenannten klassischen Testtheorie (Gulliksen, 1950), die im Grunde eine Reliabilitätstheorie ist. Auf die praktische Frage, wie man denn die Meßgenauigkeit eines diagnostischen Verfahrens abschätzen könnte, gibt es mehrere Antworten. Mit den Methoden der *Testwiederholung (Koeffizient der zeitlichen Stabilität eines Merkmales)*, der *Paralleltestmethode (Koeffizient der Äquivalenz)*, der *Testhalbierung* und der *Konsistenzanalyse (Homogenitätskoeffizienten)* liegen vier Operationalisierungen des Meßgenauigkeitsgedankens vor. Durch die Bestimmung von Reliabilitätskoeffizienten ist es im Rahmen der klassischen Testtheorie auch möglich, den Standardmeßfehler eines diagnostischen Verfahrens zu berechnen. Betrachtet man die verschiedenen Operationalisierungen zur Feststellung der Meßgenauigkeit eines diagnostischen Verfahrens, dann wird klar, daß es „die" Reliabilität eines Tests nicht gibt, sondern daß aufgrund dieser verschiedenen inhaltlichen Zugänge mehrere „Meßgenauigkeiten" zu unterscheiden sind (Michel, 1964, S.36).

(3) *Validität* (Gültigkeit): Für ein diagnostisches Verfahren genügt es nicht, seine Meßgenauigkeit unter Beweis gestellt zu haben, es muß auch „valide" sein, d. h. man muß die psychologische Bedeutung des mit einer diagnostischen Methode erhobenen Maßes kennen. Bekannt ist auch hier wieder Lienerts Definition (1967, S.16), wonach „die Validität eines Testes den Grad der Genauigkeit an(gibt), mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen soll oder zu messen vorgibt, auch tatsächlich mißt."

Auch hierbei sind wieder verschiedene Validitätskonzepte zu unterscheiden. Mit *inhaltlicher Validität („content validity")* verbindet man die Annahme, daß ein Test oder dessen Elemente so beschaffen sind, daß sie das in Frage stehende Persönlichkeitsmerkmal in optimaler Weise repräsentieren, daß also der Test selbst das beste Kriterium für das Persönlichkeitsmerkmal ist (z.B. sogenannte „Arbeitsproben"). Ob diese Art der Validitätsangabe ein für die psychologische Praxis in jedem Fall fruchtbarer Begriff sei, läßt sich anzweifeln. Der schwerwiegendste Einwand dagegen ist, daß diesem Konzept keine konzise empirische Überprüfungsmöglichkeit zugeordnet ist. Die Be-

rufung auf diese Art der Validität ist oft ein bloßes Dafürhalten, auch wenn es sich dabei um Urteile von Fachleuten handelt. Damit gerät diese Validität in enge Nachbarschaft zu der „face-validity", der „Augenscheinlichkeitsvalidität" (Drenth, 1969, S.187). Man sollte auch dann vorsichtig sein, wenn von „logischer Validität" gesprochen wird. Was mit „logisch" gemeint ist, ist zumeist eine vom Testautor als einsichtig empfundene Behauptung und nicht mehr. Auch in solchen Fällen kann man fast immer sicher sein, daß sich der Testautor um den schwierigen Weg der empirischen Bestätigung seiner Vermutungen drücken will.

Mit dem Stichwort der *empirischen Validität* (*„predictive and concurrent validity"*) ist gemeint, daß die Bedeutung eines diagnostischen Verfahrens über den empirischen Nachweis geführt wurde, welches Verhalten aus dem Verhalten in der diagnostischen Situation vorhergesagt werden kann. Je nachdem, ob das Verhalten, auf das geschlossen wird, in der Zukunft liegt oder gleichzeitig erhoben wird, spricht man von Gleichzeitigkeits- oder von Vorhersagevalidität; handelt es sich bei dem Kriterium um einen anderen Test, so spricht man von interner Validität, handelt es sich um ein anderes Kriterium, so spricht man von externer Validität. Die Höhe eines Korrelationskoeffizienten zwischen Testergebnis und Kriteriumsverhalten allein ist aber noch kein zureichendes Maß für die praktische Brauchbarkeit eines Verfahrens. Wesentlich sind noch (1) die Selektionsraten (d.h. wie viele von den getesteten Probanden sollen aufgenommen oder abgewiesen werden), (2) die Verteilung von Eignungs- und Nichteignungsquoten in der Population und (3) der Nutzen und der Schaden, die mit einer richtigen bzw. falschen Entscheidung verbunden sind.

Mit *Konstrukt-Validität* (*„construct validity"*) ist die Einordnung eines in einer Testsituation erhobenen Verhaltens in ein theoretisches Bezugssystem gemeint. Testdaten stellen nach dieser Interpretation nichts anderes dar als Operationalisierungsversuche von Persönlichkeitskonstrukten. Bei diesem Vorgang der Begriffsvalidierung handelt es sich um dasselbe wie um das Finden einer Theorie. Die Bedeutungsanalyse bezweckt letzten Endes das Auffinden und Bestätigen einer Theorie oder eines theoretischen Konstruktes, die (das) die Erklärung eines Testverhaltens ermöglicht (Westmeyer, 1972, S.64).

Die Frage der Validität eines diagnostischen Verfahrens muß damit in Zusammenhang gesehen werden, inwieweit ein Test die an ihn gestellten Forderungen erfüllt. Diese Ansprüche müssen nicht immer darin bestehen, ein Individuum auf einer Dimension mit bestimmter Metrik zu lokalisieren. Ein Test kann u.a. ganz pragmatisch dazu verwendet werden, eine Auslese vorzunehmen, ohne daß man auf Fähigkeiten oder Eigenschaften des Probanden rekurrieren muß. Man kann den Validitätsbegriff unter diesem pragmatischen Aspekt der Entscheidung für bestimmte Zwecke sehen und daher fragen, für welche Entscheidungen ein Test valide ist. „Validität" ist demnach wiederum nicht eine generelle Eigenschaft eines Tests, sondern ein Test kann für einen bestimmten Zweck valide sein und für einen anderen nicht.

(4) *Testnormierung*: Ein konkretes Testdatum ist nicht aus sich selbst heraus interpretierbar, sondern jedes Testergebnis muß in ein Bezugssystem eingeordnet werden. Bekanntlich stehen dafür drei Möglichkeiten zur Verfügung: die *intraindividuelle Norm* (*ipsative Norm*, z.B. der Vergleich der aktuellen Leistung eines Schülers mit seiner früheren), die *interindividuelle Norm* (*soziale Norm*, z.B. bei psychometrischen Persönlichkeitstests) und die *Idealnorm* (*objektive Norm, lehrzielorientierte Norm*, z.B. der Vergleich des realen Verhaltens eines Patienten mit einer idealen Verhaltensbeschreibung in einer kritischen Situation, z.B. im Rahmen der Verhaltenstherapie oder bei der kriteriumsorientierten Leistungsmessung).

Weitere Nebenkriterien, die für die Anwendung eines Testverfahrens wichtig sein können (Lienert, 1967, S.19), sind (5) *Ökonomie*, (6) *Nützlichkeit* und (7) *Vergleichbarkeit*.

4. Klassifikation von Tests

Für die Klassifikation von Tests werden die verschiedensten formalen und inhaltlichen Kriterien vorgeschlagen (Irle, 1966). Für die Datenbank PSYTKOM wurde primär von inhaltlichen und anwendungsbezogenen Gesichtspunkten Gebrauch gemacht. Bezogen auf die Hauptinhaltsgruppen stehen im deutschen Sprachraum die in Tabelle 24 aufgelisteten 2762 Testverfahren zur Verfügung (dabei kommen Doppelzählungen wegen der Möglichkeit eines Mehrfacheintrages vor). Die Zahl der Verfahren gibt einen groben Indikator für die diagnostische Wichtigkeit der einzelnen Bereiche ab.

Tabelle 26
Hauptinhaltsgruppen von Tests in der Datenbank PSYTKOM (Eberwein, 1991, S. 267)

Inhaltsgruppe	Anzahl dokumentierter Verfahren
1. Entwicklungstests	163
2. Intelligenztests	237
3. Spezielle Fähigkeits- und Eignungstests	59
– Allgemeine Sporttests	201
4. Leistungstests	48
5. Kreativitätstests	7
6. Schulleistungstests	219
7. Sensomotorische Fähigkeiten	103
8. Allgemeine Einstellungs- und Interessentests	397
9. Berufliche Einstellungs- und Interessentests	112
10. Familiäre und partnerbezogene Einstellungstests	105
11. Schulische Einstellungstests	123
12. Allgemeine Persönlichkeitsverfahren	356
13. Verfahren im Bereich der Klinischen Psychologie	345
14. Verfahren der verhaltenstheoretischen Diagnostik	88
15. Entfaltungs- und Gestaltungsverfahren	76
16. Verfahren zur Erfassung soziografischer Daten	21
17. Allgemeine Verhaltensskalen	13
18. Sonstige Verfahren	89

Nach der Praxis der Testveröffentlichungen dominiert als Konstruktionsprinzip für die Testerstellung die sogenannte klassische Testtheorie. Tests auf der Basis probabilistischer Testmodelle sind nur in Einzelfällen, solche auf der Grundlage einer kriteriumsbezogenen Testtheorie so gut wie gar nicht vorhanden. Aktuelle Entwicklungen im Bereich der Testtheorie (Kubinger, 1988), werden also in der Testpraxis so gut wie nicht berücksichtigt.

Die Dokumentation von PSYTKOM macht auch deutlich, daß in dem vorhandenen Angebot – selbst in den traditionsreichen Bereichen der Intelligenz-, Leistungs- und Persönlichkeitsdiagnostik – viele weitere Schwächen hinsichtlich der Erfüllung der Gütekriterien vorhanden sind. Auffällig ist dies vor allem bezüglich rein handwerklicher Kriterien, z.B. im Bereich der Normierung (mangelnde Aktualität und Repräsentativität der Vergleichsdaten). Nicht zuletzt ist dieser Mangel darauf zurückzuführen, daß die Testkonstruktion und -publikation in der Regel von Einzelpersonen getragen wird und keine institutionelle Unterstützung vorhanden ist, innerbetriebliche Entwicklungen hingegen der Allgemeinheit zumeist nicht zugänglich gemacht werden.

– Diagnostik
– Eignungsdiagnostik
– Intelligenzmessung
– Klinisch-psychologische Diagnostik
– Leistung
– Pädagogisch-psychologische Diagnostik
– Persönlichkeitsdiagnostik

Weiterführende Literatur

Fisseni, H.J. (1990). Lehrbuch der psychologischen Diagnostik. Göttingen: Hogrefe.

Groffmann, K.-J. & Michel, L. (Hrsg.). (1982). Grundlagen psychologischer Diagnostik. Enzyklopädie der Psychologie (Themenbereich B, Serie II, Bd. 1). Göttingen: Hogrefe.

Jäger, R.S. (Hrsg.). (1988). Psychologische Diagnostik. München: PVU.

Kubinger, K.D. (Hrsg.). (1988). Moderne Testtheorie – ein Abriß samt neuesten Beiträgen. München: PVU.

Strube, G. (Hrsg.). (1977). Binet und die Folgen (Band V der Psychologie des 20. Jahrhunderts). Zürich: Kindler.

Literatur

Binet, A. & Henri, V. (1898). La fatigue intellectuelle. Paris.

Brickenkamp, R. (Hrsg.). (1983). Erster Ergänzungsband zum Handbuch psychologischer und pädagogischer Tests. Göttingen: Hogrefe.

Brickenkamp, R. (Hrsg.). (1975). Handbuch psychologischer und pädagogischer Tests. Göttingen: Hogrefe.

Cattell, J. McKeen (1890). Mental tests and measurement. Mind, 15, 373-381.

Cranach, M. von & Frenz, H.-G. (1969). Systematische Beobachtung. In F. Graumann, L. Kruse & B. Kroner (Hrsg.), Handbuch der Psychologie (Bd. 7), Sozialpsychologie (S.269-331). Göttingen: Hogrefe.

Dorsch, F. (1982). Psychologisches Wörterbuch (10. neubearbeitete Aufl.). Bern: Huber.

Drenth, P. (1969). Der psychologische Test. München: Barth.

Ebbinghaus, H. (1897). Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. Zeitschrift für Psychologie, 13, 401-459

Eberwein, M. (1991). Intelligenztestverfahren. Trier: ZPID.

Ekman, G. (1955). Konstruktion und Standardisierung von Tests. Diagnostica, 1, 15-19.

Eysenck, H.J. (1962). Wege und Abwege der Psychologie. Hamburg: Rowohlt.

Fischer, G. (1974). Einführung in die Theorie psychologischer Tests. Bern: Huber.

Galton, F. (1870). Hereditary genius. New York: Macmillan.

Groffmann, K. (1964). Die Entwicklung der Intelligenzmessung. In R. Heiss, J. Groffmann & L. Michel (Hrsg.), Handbuch der Psychologie (Bd. 6), Psychologische Diagnostik (S.148-199). Göttingen: Hogrefe.

Groffmann, K.-J. & Michel, L. (Hrsg.). (1982). Grundlagen psychologischer Diagnostik. Enzyklopädie der Psychologie (Themenbereich B, Serie II, Bd. 1). Göttingen: Hogrefe.

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

Hartmann, H. (1970). Psychologische Diagnostik. Stuttgart: Kohlhammer.

Heiss, R. (1964). Psychologische Diagnostik. Einführung und Überblick. In R. Heiss, J. Groffmann & L. Michel (Hrsg.), Handbuch der Psychologie (Bd. 6), Psychologische Diagnostik (S.3-18). Göttingen: Hogrefe.

Hofstätter, P.R. (1971). Differentielle Psychologie. Stuttgart: Kröner.

Irle, M. (1956). Die Klassifikation von Tests. Diagnostica, 2, 61-66.

Klauer, K.J. (1987). Kriteriumsorientierte Tests. Lehrbuch der Theorie und Praxis lehrzielorientierten Messens. Göttingen: Hogrefe.

Kraepelin, E. (1895). Der psychologische Versuch in der Psychiatrie. Psychologische Arbeiten, 1, 1-91.

Lienert, G.A. (1967). Testaufbau und Testanalyse. Weinheim: Beltz.

Lukesch, H. (Hrsg.). (1992). PSYTKOM. Datenbank psychologischer und pädagogischer Testverfahren. Trier: ZPID.

Mees, U. (1977). Verhaltensbeobachtung in der natürlichen Umgebung. In U. Mees & H. Selg (Hrsg.), Verhaltensbeobachtung und Verhaltensmodifikation (S.14-32). Stuttgart: Klett.

Michel, L. (1964). Allgemeine Grundlagen psychometrischer Tests. In R. Heiss, K.J. Groffmann & L. Michel (Hrsg.), Psychologische Diagnostik. Handbuch der Psychologie (Bd. 6) (S.19-70). Göttingen: Hogrefe.

Rollett, B. (1976). Kriterienorientierte Prozeßdiagnostik im Behandlungskontext. In K. Pawlik (Hrsg.), Diagnose der Diagnostik (S. 131-148). Stuttgart: Klett.

Selg, H. & Bauer, W. (1971). Forschungsmethoden der Psychologie. Stuttgart: Kohlhammer.

Strube, G. (Hrsg.). (1977). Binet und die Folgen (Band V der Psychologie des 20. Jahrhunderts. Zürich: Kindler.

Sweetland, R.C. & Keyser, D.J. (1986). Tests. A comprehensive reference for assessments in psychology, education, and business (2nd ed.). Kansas City: Test Corporation of America.

Webb, E.J., Campbell, D.T., Schwartz, R.D. & Sechrest, L. (1966). Unobstrusive measures. Chicago: Rand McNally.

Westmeyer, H. (1972). Logik der Diagnostik. Grundlagen einer normativen Diagnostik. Stuttgart: Kohlhammer.

ZUMA. (1983). ZUMA-Handbuch sozialwissenschaftlicher Skalen. Mannheim: Zuma-Eigenverlag.