Enzyklopädie der Psychologie

Methoden der psychologischen Diagnostik

Psychologische Diagnostik

2

Hogrefe · Verlag für Psychologie
Göttingen · Bern · Toronto · Seattle

Enzyklopädie der Psychologie

ENZYKLOPÄDIE DER PSYCHOLOGIE

In Verbindung mit der
Deutschen Gesellschaft für Psychologie

herausgegeben von

Prof. Dr. Niels Birbaumer, Tübingen
Prof. Dr. Dieter Frey, München
Prof. Dr. Julius Kuhl, Osnabrück
Prof. Dr. Wolfgang Schneider, Würzburg
Prof. Dr. Ralf Schwarzer, Berlin

Themenbereich B

Methodologie und Methoden

Serie II

Psychologische Diagnostik

Band 2

Methoden der psychologischen Diagnostik

Hogrefe • Verlag für Psychologie
Göttingen • Bern • Toronto • Seattle

Methoden der psychologischen Diagnostik

herausgegeben von

Prof. em. Dr. Lutz F. Hornke, Aachen
Prof. em. Dr. Manfred Amelang, Heidelberg
Prof. Dr. Martin Kersting, Münster

Hogrefe • Verlag für Psychologie
Göttingen • Bern • Toronto • Seattle

Göttingen · Bern · Wien · Paris · Oxford · Prag · Toronto
Cambridge, MA · Amsterdam · Kopenhagen · Stockholm
Rohnsweg 25, 37085 Göttingen

http://www.hogrefe.de
Aktuelle Informationen · Weitere Titel zum Thema · Ergänzende Materialien

---

Satz: ARThür Grafik-Design & Kunst, Weimar
Format: PDF

ISBN 978-3-8409-1524-6

Autorenverzeichnis

Prof. em. Dr. Manfred Amelang

Universität Heidelberg
Psychologisches Institut
Hauptstraße 47
69117 Heidelberg
E-Mail: manfred.amelang@psychologie.uni-heidelberg.de

Prof. Dr. Martin E. Arendasy

Universität Graz
Institut für Psychologie
Universitätsplatz 2
A-8010 Graz
Österreich
E-Mail: martin.arendasy@uni-graz.at

Dr. Nicolas Becker

Universität des Saarlandes
Differentielle Psychologie und Psychologische Diagnostik
Campus A1.3
66123 Saarbrücken
E-Mail: nicolas.becker@mx.uni-saarland.de

Dr. Andreas Frey

Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN)
an der Universität Kiel
Olshausenstraße 62
24098 Kiel
E-Mail: frey@ipn.uni-kiel.de

Prof. Dr. Janet Harkness

Survey Research and Methodology Program
University of Nebraska-Lincoln
UNL Gallup Research Center
201 North 13th Street
Lincoln, NE 68588–0241
USA
E-Mail: jharkness2@unl.edu

Dr. Philipp Yorck Herzberg

Abteilung für Medizinische Psychologie und Medizinische Soziologie
Universitätsklinikum Leipzig AöR
Philipp-Rosenthal-Straße 55
04103 Leipzig
E-Mail: herzberg@medizin.uni-leipzig.de

Prof. em. Dr. Lutz F. Hornke

Institut für Psychologie
RWTH Aachen University
Jägerstraße 17/19
52056 Aachen
E-Mail: lutz.hornke@psych.rwth-aachen.de

Prof. Dr. Martin Kersting

Fachhochschule des Bundes für öffentliche Verwaltung
Fachbereich Finanzen
Gescher Weg 100
48161 Münster
E-Mail:
Martin@Kersting-internet.de

Prof. Dr. Peter Ph. Mohler

Kirchstr. 7
76829 Leinsweiler
E-Mail: pmohler798@aol.com

Prof. Dr. Helfried Moosbrugger

Institut für Psychologie
Johann Wolfgang Goethe-Universität
Frankfurt am Main
Mertonstraße 17
60054 Frankfurt am Main
E-Mail: moosbrugger@psych.uni-frankfurt.de

PD Dr. Beatrice Rammstedt

GESIS-Leibniz Institut
für Sozialwissenschaften
PF 122155
68072 Mannheim
E-Mail:
beatrice.rammstedt@gesis.org

Dr. Wolfgang Rauch

Institut für Psychologie
Johann Wolfgang Goethe-Universität
Frankfurt am Main
Mertonstraße 17
60054 Frankfurt am Main
E-Mail: wolfgang.rauch@psych.uni-frankfurt.de

Prof. Dr. Jürgen Rost

Leibniz-Institut für die Pädagogik
der Naturwissenschaften (IPN)
an der Universität Kiel
Olshausenstraße 62
24098 Kiel
E-Mail: an@j-rost.de

Prof. Dr. Frank M. Spinath

Universität des Saarlandes
Differentielle Psychologie und
Psychologische Diagnostik
Campus A1.3
66123 Saarbrücken
E-Mail: f.spinath@mx.uni-saarland.de

Mag. Markus Sommer

SCHUHFRIED GmbH
Hyrtlstraße 45
A-2340 Mödling
Österreich
E-Mail: Sommer@schuhfried.at

Jun.-Prof. Dr. Anja Strobel

TU Dresden
Professur Diagnostik und
Intervention
Helmholtzstraße 10
01069 Dresden
E-Mail: anja.strobel@tu-dresden.de

Dr. Oliver Walter

Leibniz-Institut für die Pädagogik
der Naturwissenschaften (IPN)
an der Universität Kiel
Olshausenstraße 62
24098 Kiel
E-Mail: walter@ipn.uni-kiel.de

Prof. Dr. Karl Westhoff

TU Dresden
Professur Diagnostik und Intervention
Helmholtzstraße 10
01069 Dresden
E-Mail: Karl.Westhoff@tu-dresden.de

Dr. Safir Yousfi

Bundesagentur für Arbeit
Psychologische Forschung und Entwicklung
Regensburger Straße 104
90478 Nürnberg
E-Mail:
Safir.Yousfi@arbeitsagentur.de

Vorwort

Eine Enzyklopädie ist zunächst sehr vereinfacht viel Arbeit, die auf viele Köpfe und fleißige Hände verteilt ist. Dabei handelt es sich um eine besondere Textsorte, die mit einer ganz besonderen fachlichen Absicht erstellt wurde.

Seit dem 5. Jh. v. Chr. versteht man unter „Enzyklopädie" die universelle Bildung oder doch zumindest die Zusammenstellung derselben (siehe dazu den Eintrag in der wohl größten Enzyklopädie der Gegenwart, der Wikipedia, die zudem noch internetbasiert und mit vielen Querverweisen organisiert und allseits zugänglich ist). Enzyklopädisten meinten damit vom Wortstamm her zunächst einen geschlossenen Kreis an Wissen, und später stand die Idee des kreisförmig immer wiederkehrenden und zu erweiternden bzw. erneuernden Wissens im Vordergrund.

Sicher wird niemand von den vorliegenden Enzyklopädiebänden eine vollständige Darstellung von Fragen, Methodiken und Antworten der Psychologischen Diagnostik erwarten, wohl aber Hinweise auf neue Trends und Ansätze. Die Herausgeber haben versucht, dies zu ermöglichen und mit weitgefassten Vorgaben an die Autoren sowie entsprechenden Begutachtungen und Hinweisen die Grundfragen und -antworten der Psychologischen Diagnostik zu Beginn des 21. Jahrhunderts vorzustellen. Wenngleich andere, historische Enzyklopädien viel umfangreicher waren, so kann sich diese hier doch sehen lassen: 47 Artikel von 74 Autorinnen und Autoren.

Dass eine solche Arbeit mehrere Jahre in Anspruch nahm, verwundert nicht. Autoren und Herausgeber haben mit großem Elan begonnen, aber manche mussten zwischenzeitlich ihre anfänglichen Zeitplanungen angesichts ihrer auch vielfältigen anderen Aufgaben revidieren. Doch auch die traurigen und erfreulichen privaten und beruflichen Lebensereignisse verzögerten die Arbeit: Wissenschaftler, auch und gerade Psychologen, sind Menschen; Herausgeber haben das zu akzeptieren und hoffen, all dem Beruflichen und Privaten durch Benevolenz begegnet zu sein. Die Leser werden das sicher respektieren.

Jeder Beitrag ist wie bei Zeitschriften auch entsprechend begutachtet worden. Dabei war stets eine Rückmeldung anonym, und die Herausgeber haben sich entschlossen, offen ihre Korrekturen, Kommentare und Diskussionen anzubrin-

gen. Hieraus ergab sich mit einigen Autoren ein sehr fruchtbarer Dialog; die Beiträge haben dadurch fraglos gewonnen, wenngleich die Autorinnen und Autoren jeweils den eigenen Beitrag selber verantworten, denn bei ihnen liegt letztlich die Expertise, wegen derer sie zur Mitarbeit eingeladen wurden.

Im Ergebnis ist durch die Aufgeschlossenheit und die Bemühungen aller eine Enzyklopädie der Psychologischen Diagnostik entstanden, die wie ihre Vorgänger fachlich über Fragen, Methoden und Antworten informiert. Auch wenn nicht alle anfangs geplanten Beiträge geschrieben werden konnten, so liegt doch nun ein Überblick vor, der allgemein Interessierte, Studierende und Fachwissenschaftler gleichermaßen, wenngleich aus unterschiedlicher Perspektive, ansprechen kann und soll. Zumindest hat die Psychologische Diagnostik wieder eine Plattform gefunden, von der aus Gegenwarts- und Zukunftsfragen gestellt und beantwortet werden können. Die hier vorliegende Enzyklopädie zeigt die Nähe aller Autorinnen und Autoren zueinander und zum Fach. Sie ist eine Art Zugehörigkeitsausweis; denn alle, die sich hier zu Wort melden, sind die Mitgestalter und -träger zukünftiger Entwicklungen in dieser Teildisziplin der Psychologie. Dass vieles zwar auch international mitteilbar ist, aber manches doch nur im deutschsprachigen Raum rezipiert wird, das liegt an der Natur des Gebietes: Diagnostik menschlichen Erlebens und Verhaltens geht zwar in seinen allgemeinen Grundfragen über den eigenen Sprachraum hinaus, aber die gesellschaftliche Anwendung vollzieht sich nun einmal dominant innerhalb einer (Sprach-)Kultur. Beides aber bedingt sich und ist allein genommen eher problematisch.

Nach allem gebührt aus Sicht der Herausgeber allen Autorinnen und Autoren der Dank, an dieser Enzyklopädieserie „Psychologische Diagnostik" mitgewirkt zu haben. Die vielen Leser werden ihren Dank dadurch abstatten, dass sie die Artikel zitieren, die Ideen aufgreifen, diskutieren und weiterentwickeln. Dazu gehört auch die Frage, ob die Textsorte Enzyklopädie zukünftig nicht doch auch einer anderen Präsentationsform oder einer Ergänzung durch diese bedarf.

Die Gesellschaft aber ist eingeladen, das Nachstehende zur Kenntnis zu nehmen und dort, wo es nicht unmittelbar verständlich ist, die Autorinnen und Autoren zu fragen, anzusprechen und sich mit ihnen auseinanderzusetzen. Gerade die vielen jungen Autorinnen und Autoren garantieren, dass fachlich „nachgehalten" werden kann. Nur so wird die eigentliche Absicht einer *Enkyklios Paideia* eingelöst: Psychologie und Psychologische Diagnostik im Besonderen ist der gesellschaftlichen Wirklichkeit und Aufklärung verpflichtet; hier entfaltet sie ihre vornehmsten Wirkungen.

Lutz F. Hornke, Aachen,
Manfred Amelang, Heidelberg und
Martin Kersting, Münster

Inhaltsverzeichnis

1. Kapitel: Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle
Von Wolfgang A. Rauch und Helfried Moosbrugger

2. Kapitel: Psychometrische Grundlagen von Large Scale Assessments
Von Oliver Walter und Jürgen Rost

3. Kapitel: Methoden der Item- und Skalenkonstruktion
Von Safir Yousfi

4. Kapitel: Automatisierte Itemgenerierung: Aktuelle Ansätze, Anwendungen und Forschungen
Von Martin Arendasy und Markus Sommer

5. Kapitel: Kriteriumsorientierte Diagnostik
Von Philipp Yorck Herzberg und Andreas Frey

6. Kapitel: Verhaltensbeobachtung
Von Frank M. Spinath und Nicolas Becker

7. Kapitel: Interview
Von Karl Westhoff und Anja Strobel

8. Kapitel: Psychodiagnostische Verfahren im Kulturvergleich
Von Beatrice Rammstedt, Janet Harkness und Peter Ph. Mohler

1. Kapitel

Klassische Testtheorie. Grundlagen und Erweiterungen für heterogene Tests und Mehrfacettenmodelle

Wolfgang A. Rauch und Helfried Moosbrugger

1 Konzeptuelle Grundlagen und Definitionen

1.1 Einführung

Die Klassische Testtheorie (KTT) umfasst eine Reihe von mathematisch-statistischen Modellvorstellungen zur quantifizierenden Beschreibung der Variabilität von Testwerten; Hauptziel ist die Schätzung der Reliabilität von Testwerten. Historisch gesehen stellt die KTT einen der wesentlichen Meilensteine in der Entwicklung des heutigen empirisch-psychologischen Forschungsparadigmas dar – sie fußt auf den Beiträgen von Charles Spearman (1904a, 1904b) zur psychologischen Messung und zur Intelligenzforschung. In den letzten Jahrzehnten hätte mancher Forscher die KTT auch gerne „nur" in den Rahmen einer historischen Abhandlung verwiesen, war doch die KTT Gegenstand z. T. vehementer Kritik und offener Ablehnung. In diesem Trend wurde die KTT auch implizit oder explizit als minderwertige Alternative zu Modellen der Item-Response-Theorie (IRT) dargestellt, vor allem deshalb, weil die KTT die Frage nicht behandelt, *wie* die Testwerte, mit denen sie sich beschäftigt, zustande kommen. Diese Kritik ist zumindest in gewisser Hinsicht nur schwer haltbar, da der wesentlichste Unterschied zwischen der „klassischen" KTT und Modellen der IRT in der Spezifikation von Verteilungsannahmen für die Testwerte bzw. für Antworten auf Testitems liegt. Daraus folgend erlauben Modelle der IRT eine stringentere Überprüfung der Modellannahmen. Diese zusätzlichen Aspekte rechtfertigen aber nur kaum, dass diesen Modellen ein gewissermaßen qualitativ hervorgehobener wissenschaftlicher Status zugewiesen wird, wie es manche frühen Verfechter

getan haben. Allerdings gibt es einen quantitativen Vorteil von IRT-Modellen: Die stringenten Annahmen erlauben, sofern sie empirisch erfüllt sind, auch eine genauere, modellkonforme individualdiagnostische Auswertung, d. h. die Schätzung von „latenten" Merkmalsausprägungen von Individuen.

Die Sinnhaftigkeit einer ausführlichen Abhandlung über die KTT ergibt sich also weniger hinsichtlich ihrer Anwendung für individualdiagnostische Zwecke; entsprechend werden keinerlei Ergebnisse dafür berichtet. Die hier eingenommene Perspektive ist die einer Komplementarität von KTT und IRT (vgl. auch Holland & Hoskens, 2003): Der wesentliche Nutzen der KTT und vor allem ihrer Erweiterungen im Bereich faktorenanalytischer Modelle und deren Fortführung in linearen Strukturgleichungsmodellen (SEM, *structural equation modeling*) sowie im Rahmen der Generalisierbarkeitstheorie (G-Theorie) liegt demnach in der Untersuchung von theoretischen Annahmen über die Struktur von Merkmalen in Populationen von Untersuchungsobjekten. Zumindest noch heute erlauben lineare Strukturgleichungsmodelle deutlich komplexere strukturelle Analysen als es mit IRT-Modellen bislang möglich ist. Solche Begrenzungen bei strukturellen Analysen im Rahmen von IRT-Modellen sind aber nur durch mangelnde „Computerpower" bedingt und werden in Zukunft wohl immer weniger bestehen. Bereits heute gibt es theoretische und z. T. auch praktische Arbeiten zur Verbindung von Strukturgleichungsmodellen auf der einen Seite und generalisierten Modellen für Antwortvariablen aus der IRT auf der anderen Seite (vor allem Skrondal & Rabe-Hesketh, 2004), so dass in Zukunft die Trennung zwischen „klassischen" Testtheorien und IRT-Modellen wohl nicht mehr sinnvoll sein wird. Vorweg geschickt sei, dass im vorliegenden Kapitel pragmatische Aspekte der Testkonstruktion nur am Rande behandelt werden. Für aktuelle Einführungen sei auf Bühner (2006) oder Moosbrugger und Kelava (2007) verwiesen. Bevor in Abschnitt 2 die KTT im engeren Sinne eingeführt wird, sollen noch kurz einige Grundlagen eingeführt und zentrale Grundbegriffe definiert werden.

1.2 Testtheorien und Voraussetzungen für deren Anwendung

Hauptanwendungsgebiet und historischer Ausgangspunkt von Testtheorien ist die psychologische Diagnostik, nämlich die

> „… regelgeleitete Sammlung und Verarbeitung von gezielt erhobenen Informationen, die für das Verständnis menschlichen Verhaltens und Erlebens bedeutsam sind. Die Fragestellungen der Psychologischen Diagnostik können sich dabei auf die Beschreibung und Klassifikation, Erklärung, Vorhersage und Evaluation von Zuständen und/oder Verläufen beziehen" (Eid & Petermann, 2006, S. 16).

Testtheorien können allerdings auch auf völlig andere Inhaltsgebiete übertragen werden als auf die „Fragestellungen der Psychologischen Diagnostik", denn „Beschreibung und Klassifikation, Erklärung, Vorhersage und Evaluation" charakterisieren viel allgemeiner die typischen Fragestellungen der meisten empirischen Wissenschaften. Entsprechend finden testtheoretische Ergebnisse keinesfalls nur in der psychologischen Diagnostik, sondern etwa auch in der wirtschafts- und sozialwissenschaftlichen Forschung oder der Biologie und Ökologie ihre Anwendung.

Die folgenden zwei Annahmen können als Grundvoraussetzungen für den Einsatz eines diagnostischen Verfahrens gesehen werden: (1) Personen – oder grundsätzlicher: Untersuchungsobjekte – unterscheiden sich voneinander, sonst wäre der Einsatz diagnostischer Verfahren sinnlos und (2) diese Unterschiede müssen (a) in irgendeiner Art von beobachtbaren Verhaltensäußerungen oder sonstigen empirisch erfassbaren Eigenschaften der Untersuchungsobjekte bestehen oder (b) mit den Verhaltensäußerungen oder beobachtbaren Eigenschaften zusammenhängen. Die Differenzierung von direkt oder indirekt beobachtbaren Unterschieden stellt wohl eines der zentralen Probleme psychologischer Messungen im engeren Sinne dar – gerade psychologische Tests zielen fast immer auf die Erfassung sogenannter *latenter*, d. h. nicht direkt beobachtbarer Konstrukte ab. Ein zentraler Problembereich von Testtheorien ist daher die Frage, ob die empirisch beobachteten Unterschiede zwischen den Untersuchungsobjekten auch tatsächlich in der theoretisch angenommenen Weise mit den Unterschieden auf den latenten Konstrukten zusammenhängen.

In der praktischen Anwendung werden latente Konstrukte unter Umständen „nur" als nützliche Datenreduktion angesehen werden. In der inhaltlich psychologischen Forschung aber wird latenten Konstrukten zumindest implizit, und meist auch explizit eine objektive Realität zugeschrieben (vgl. Borsboom, Mellenberg & van Heerden, 2003). Die unterschiedlichen Konzeptionen von latenten Konstrukten werden im Verlaufe dieses Beitrags immer wieder aufgegriffen, da sie von zentraler Bedeutung nicht nur für die Bewertung unterschiedlicher testtheoretischer Ergebnisse, sondern auch für inhaltlich-psychologische Anwendungen sind.

Im engeren Sinne lässt sich eine Testtheorie als eine Reihe von mathematisch-statistischen und praxisorientierten Regeln zum Umgang mit und zur Beschreibung von Testwerten und deren Zusammenhängen untereinander definieren. Gleichzeitig umfassen Testtheorien üblicherweise auch statistische und inhaltlich-interpretative Regeln zur Beurteilung der Qualität der Testwerte bzw. der bei deren Erhebung verwendeten Prozeduren. Mit *Testwert* ist hier eine aufgrund irgendeiner Verrechnungsvorschrift zustande gekommene Zahl gemeint, die im Allgemeinen eine Zusammenfassung des Ergebnisses eines Psychologischen Tests

oder dessen einzelner Bestandteile darstellt. Grundsätzlicher kann mit dem Begriff Testwert auch das zahlenmäßige Ergebnis irgendeiner empirischen Untersuchung mit dem Ziel der Unterscheidung zwischen den Untersuchungsobjekten gemeint sein. Testtheorien beschäftigen sich also bereits mit dem Ergebnis einer Messung und sind damit von *Messtheorien* abzugrenzen, die sich mit der Frage beschäftigen, inwieweit sich Beziehungen zwischen den Objekten empirischer Forschung durch die Beziehungen zwischen Zahlen repräsentieren lassen.

Da Testtheorien aber hauptsächlich für die Analyse von Psychologischen Tests im engeren Sinne verwendet werden, soll hier ebenfalls eine Definition des Begriffes *„Psychologischer Test"* in enger Anlehnung an Rost (2004) aufgeführt werden:

> Ein Psychologischer Test ist ein spezielles psychologisches Experiment mit dem Ziel, vergleichende Aussagen über die Personen abzuleiten. Vom Versuchsleiter bewusst hergestellte unterschiedliche Reizgegebenheiten (Testaufgaben oder Items) bilden die *unabhängige Variable* (UV); die Reaktion auf die Testaufgaben oder Items unter den so hergestellten verschiedenen Versuchsbedingungen wird als *abhängige Variable* (AV) beobachtet oder gemessen.

Die aufgeführte Analogie von Tests und Experimenten (vgl. Rost, 2004) ist von großem heuristischem und konzeptuellem Nutzen; die G-Theorie (vgl. Abschnitt 9.4) etwa liefert Methoden zur Untersuchung einer systematischen Variation von Test- und Itemeigenschaften, die nicht nur, aber vor allem auch einem experimentellen Ansatz dienen. Insgesamt ist die gemeinsame Untersuchung von interindividuellen Unterschieden und Effekten experimenteller Bedingungsvariation ein häufig genanntes Ziel der Allgemeinen Psychologie wie auch der persönlichkeitspsychologischen Forschung (vgl. Moosbrugger & Rauch, 2009). Auch aus der Perspektive der IRT wird immer mehr auf eine vergleichbare Zusammenführung der „zwei psychologischen Disziplinen" abgezielt (Cronbach, 1957; De Boeck & Wilson, 2004).

1.3 Gütekriterien als Anforderungen an die Qualität psychologischer Diagnostik

Im vorangegangenen Abschnitt wurde als Gegenstand (und implizit auch als Zweck) von Testtheorien unter anderem die Beurteilung der Qualität von Testwerten und der Prozeduren zu ihrer Erhebung genannt. Die Qualitätsanforderungen dabei sind eine Teilmenge der grundsätzlichen Anforderungen, *wie* eine wissenschaftliche Untersuchung durchzuführen ist. Solche normativen Grundsätze entstehen üblicherweise als mehr oder weniger informeller Konsens einer

wissenschaftlichen Gemeinschaft (vgl. Messick, 1989). Gerade für die psychologische Diagnostik wurden Qualitätsgesichtspunkte in den letzten Jahren allerdings immer stärker formalisiert (vgl. Moosbrugger & Höfling, 2006). Insbesondere in der deutschsprachigen Psychologie werden solche Qualitätsanforderungen meist unter dem Oberbegriff *Gütekriterien* behandelt. Für Qualitätsanforderungen an berufsbezogene Eignungsbeurteilungen existiert sogar eine DIN-Norm (die DIN 33430; DIN, 2002).

Testtheorien haben in erster Linie eine normative Aufgabe zur Sicherung der Gütekriterien der *Reliabilität* und *Validität*. Allerdings wird die Validität in der KTT und der IRT nicht direkt thematisiert. Untersuchungen der Validität erfolgen stattdessen mit einem Methodenarsenal, das über rein testtheoretische Überlegungen hinausgeht, und dabei auf Ergebnisse aus den Testtheorien zurückgreift. Das dritte Hauptgütekriterium, die *Objektivität,* im Sinne einer möglichst weitgehenden Unabhängigkeit der Untersuchungsergebnisse von der Person desjenigen, der die Untersuchung durchführt, wird im Rahmen von Testtheorien meist implizit vorausgesetzt und nicht explizit behandelt.

Im Folgenden soll die Definition dieser Gütekriterien zunächst als normative Anforderung erfolgen, d. h. noch ohne Bezug zu einer spezifischen Testtheorie. Zu beachten ist dabei, dass der konzeptuelle Unterschied zwischen Gütekriterien als Qualitätsanforderungen und quantitativen Angaben in Form von Reliabilitäts- und Validitätskoeffizienten zur Beurteilung der Erfüllung dieser Anforderungen in der Literatur häufig nicht deutlich gemacht wird. Insbesondere der Begriff der Reliabilität ist mit der KTT verknüpft, und in der modernen, vor allem englischsprachigen psychometrischen Literatur wird der Begriff Reliabilität sehr häufig im Hinblick auf seine Formalisierung im Rahmen der KTT definiert (z. B. Lucke, 2005b; Zinbarg, Revelle, Yovel & Li, 2005).

1.3.1 Reliabilität als normative Anforderung

Zur Einführung des Reliabilitätsbegriffes wird zunächst auf zwei in der Literatur verbreitete Definitionen zurückgegriffen:

> „Unter der Reliabilität eines Testes versteht man den *Grad der Genauigkeit*, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal misst, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht“ (Lienert, 1969, S. 14; Hervorhebung im Original).
>
> „Ein Test ist dann perfekt reliabel, wenn er das Merkmal, das er misst, exakt, das heißt ohne Messfehler, misst. Dabei wird das Ausmaß der Reliabilität vom Anteil des Messfehlers an der Messung bestimmt“ (Schermelleh-Engel, Kelava & Moosbrugger, 2006, S. 421).

Gemeinsam ist beiden Definitionen, dass Reliabilität als etwas Abstufbares aufgefasst wird („Ausmaß" bei Schermelleh-Engel et al., 2006; „Grad" bei Lienert, 1969). Ein Unterschied liegt darin, dass in der Definition von Schermelleh-Engel et al. (2006) explizit auf das Konzept „Messfehler" (und dessen unabänderliches Vorhandensein) rekurriert wird, während der Begriff der Genauigkeit in der Definition von Lienert (1969) prinzipiell allgemeiner ist. *Un*genauigkeit kann nämlich nicht nur durch bei der Messung aufgetretene Fehler entstehen, sondern auch durch zu geringe Differenzierung: Beispielsweise kann ein Lineal zwar bei der Messung im Millimeterbereich gute Hilfen leisten, aber für deutlich kleinere Strecken meist keine sichtbare Unterscheidbarkeit herstellen. Allerdings kann der Begriff der „Genauigkeit" auch irreführend sein, da dieser Begriff im Alltagsgebrauch nicht unbedingt auch das grundsätzliche Problem einschließt, das sich hinter dem Begriff „Messfehler" verbirgt: *Messfehler* im engeren Sinne entstehen nämlich durch „eine Vielzahl von Faktoren, die in der Messprozedur nicht kontrolliert wurden" (Lord & Novick, 1968, S. 38; Übersetzung durch die Autoren)[1]. Immerhin macht Lienerts (1969) Reliabilitätsdefinition explizit, dass Reliabilität und Validität getrennt voneinander zu beurteilen sind („gleichgültig, ob [der Test] dieses Merkmal auch zu messen beansprucht"), was vor dem Hintergrund einer naiven, alltagssprachlichen Verbindung der Begriffe „Messfehler" und „Zuverlässigkeit" wichtig erscheint.

Schließlich wird ein wesentliches Ergebnis der moderneren psychometrischen Literatur in beiden Definitionen vernachlässigt: Ein Test als solcher ist nicht reliabel oder unreliabel. Die Bewertung der Reliabilität kann nur erfolgen, wenn in bestimmten Situationen erhobene Testwerte vorliegen, d. h. wenn im Sinne der obigen Definitionen situationsbedingte Reaktionen auf die Testaufgaben aufgezeichnet wurden und dabei Reaktionsunterschiede zwischen den Untersuchungsobjekten beobachtet werden können. Es ist unmittelbar ersichtlich, dass der Genauigkeitsbegriff die Unterscheidung zwischen den Reaktionen von Objekten erfordert; entsprechend muss der Aspekt der Abhängigkeit der Reliabilität von den Untersuchungsobjekten und Untersuchungssituationen berücksichtigt werden. Zusammenfassend schlagen wir die folgende Definition vor:

> Reliabilität ist eine Qualitätsanforderung an Testwerte: Je weniger die Testwerte durch unkontrollierbare situative Zufallseinflüsse verfälscht werden und je feinere Unterscheidungen zwischen den Untersuchungsobjekten gemacht werden können, desto höher ist die Reliabilität.

---

1 Interessanterweise bezeichnet Spearman (1904b) in seinem Beitrag zur Schätzung der von „Messfehlern" bereinigten Korrelation zweier Testreihen, der einer der wichtigsten Ausgangspunkte der psychometrischen Literatur ist, die Messfehler im heutigen Sinne als „systematic deviations" und unterscheidet diese von „accidental deviations", die aus dem Stichprobenfehler resultieren. Erst später setzt er den Begriff „accidental deviations" ein, der dem heutigen Messfehlerkonzept als im experimentellen Sinne unkontrollierbare Zufallsvariation entspricht.

1.3.2 Validität als normative Anforderung

Schon Spearman (1904b) beschäftigte sich in dem bereits als Ausgangspunkt der KTT genannten Artikel mit der Frage der „Kriteriumsvalidität", nämlich mit der Schätzung der Korrelation zweier „Testreihen" im Zusammenhang mit der Unreliabilität der Testreihen. Dennoch gibt es Autoren, die grundsätzlich verneinen, dass die Validitätsanforderungen im Rahmen des Regelwerks der KTT überhaupt behandelt werden kann (Fischer, 1974; in gewisser Weise auch Borsboom & Mellenbergh, 2002). Unzweifelhaft ist die Operationalisierung der Reliabilitätsanforderung wohl der Kernbestandteil der KTT. Die Behandlung der Validitätsanforderung ist im Rahmen der KTT hingegen nur unter Zuhilfenahme weiterer Konzepte oder unter strengen Zusatzannahmen möglich.

Diese Herauslösung der Validitätsfrage aus dem Zusammenhang der KTT ist wohl am besten durch eine kurze historische Behandlung des Begriffs der Validität zu verstehen. Bei Spearman (1904b) und anderen wurde die aus heutiger Sicht als Teilaspekt zu verstehende Frage der „Kriteriumsvalidität" behandelt, nämlich als Frage, inwieweit sich aus einem Testergebnis sinnvolle Vorhersagen für einen anderen Inhaltsbereich treffen lassen. Bald zeigte sich aber, dass die Beurteilung der Kriteriumsvalidität alleine höchst unbefriedigend ist. Eine Beschränkung auf die Vorhersage für andere Inhaltsbereiche vernachlässigt nämlich das Problem der theoretischen und inhaltlichen Interpretation von Testwerten. Darüber hinaus kann aber auch das Problem einer *regressio ad infinitum* auftreten, wenn das Kriterium anstelle einer theoretischen und inhaltlichen Analyse selbst wieder durch zu validierende Testwerte repräsentiert wird, etwa beim häufig genannten Beispiel der Validierung eines Intelligenztests durch Korrelation mit anderen Intelligenztests.

Aus diesen Gründen beschäftigte sich der wissenschaftliche Diskurs dann auch mit der Frage der *Inhaltsvalidität*, bei der es darum geht, (a) inwieweit ein Test inhaltlich das Gebiet abbildet, für das der Test genutzt werden soll, und/oder (b) inwieweit das im Test verlangte Verhalten das für das jeweilige Gebiet relevante Verhalten abbildet, und/oder (c) inwieweit die für die Bearbeitung des Tests notwendigen Prozesse die für das jeweilige Gebiet relevanten Prozesse repräsentieren (vgl. Messick, 1989)[2]. Allerdings wurde auch das Konzept der Inhaltsvalidität Gegenstand harscher Kritik; beim Beispiel von Intelligenztests führt eine falsch verstandene Inhaltsvalidierung wiederum zur *regressio ad infinitum*,

2 Die Unterscheidung der drei Aspekte der Inhaltsvalidität geht in neueren Definitionen der Inhaltsvalidität meist verloren, und entsprechend wird die Inhaltsvalidität von vielen Autoren auch eher als nebensächlich betrachtet. Gerade aber die Analyse der Antwortprozesse, die für die Bearbeitung eines Tests notwendig sind, wird zuweilen als unbedingt notwendig für die Beurteilung der Validität eines Tests angesehen (Borsboom, Mellenbergh & van Heerden, 2004).

wenn unter operationalistischer und quasi-behavioristischer Perspektive „intelligentes Verhalten“ als das Verhalten definiert wird, das für die Lösung der Intelligenztestaufgaben erforderlich ist.

In den 1940er und 1950er Jahren wurde das Problem der Validität dann vor allem durch die Arbeit einer „task force“ der American Psychological Association (APA) zur Entwicklung von Teststandards (vgl. Moosbrugger & Höfling, 2006) aufgegriffen. Als Meilenstein dieser Entwicklung ist der Artikel von Cronbach und Meehl (1955) anzusehen, der den Begriff der *Konstruktvalidität* einführte. In diesem Artikel wird vor allem auch der Begriff des Konstruktes näher beleuchtet; in loser Übersetzung definieren wir:

> Ein Konstrukt ist eine postulierte Eigenschaft von Untersuchungsobjekten, hinsichtlich derer sich die Untersuchungsobjekte unterscheiden[3] (vgl. Cronbach & Meehl, 1955, S. 283).

Die etwas umständlich erscheinende Formulierung „eine postulierte Eigenschaft“, die von den Originalautoren übernommen wurde, kann hier auch als Vermeidung möglicher erkenntnistheoretischer Fallstricke aufgefasst werden – unabhängig von der Annahme, ob ein Konstrukt tatsächlich in einer „objektiven“ Realität existiert, oder ob es sich stattdessen etwa um eine „nützliche“ oder sonst wie geartete Konstruktion ohne Anspruch auf Entsprechung in der „realen“ physischen Welt handelt, beide Ansätze lassen sich unter dem Begriff der „Postulierung“ unterbringen. Von Bedeutung ist nur, dass sich Untersuchungsobjekte hinsichtlich des Konstruktes unterscheiden lassen.

Innerhalb des von Cronbach und Meehl (1955) gesteckten Rahmens der Konstruktvalidität genügt nicht mehr eine einzelne Korrelation, um das Ausmaß der Konstruktvalidität als Validitätskoeffizient auszudrücken. Da sich die Validitätseinschätzung auf qualitativ unterschiedliche Quellen bezieht, wird der einzelne Validitätskoeffizient durch ein Programm der Konstruktvalidierung ersetzt. Dieses Programm erscheint fast wie das grundsätzliche Forschungsprogramm der empirischen Psychologie an sich (vgl. Messick, 1989): Inwieweit lassen sich Belege und Begründungen dafür finden, dass die Testwerte auch so interpretiert werden können, wie es bei der Testkonstruktion und -anwendung intendiert war? Solche Belege und Begründungen können aus der Testbearbei-

3 Cronbach und Meehl (1955) definieren den Konstruktbegriff tatsächlich direkt unter Rückgriff auf einen Test: „A construct is some postulated attribute of people, assumed to be reflected in test performance“ (S. 283); im Rahmen unserer Darstellung gehen wir allerdings davon aus, dass Konstrukte unabhängig von einem (spezifischen) Test definierbar sind. Insofern ist die Frage, inwieweit sich ein Konstrukt tatsächlich in der Testleistung abbildet, keine Frage der Definition eines Konstruktes, sondern eine Kernfrage der Testvalidierung.

tung selbst wie auch aus den Zusammenhängen mit anderen Variablen stammen, wobei diese anderen Variablen entweder als erklärende Variablen oder als abhängige (= Kriteriums-)Variablen mit den Testwerten zusammenhängen. Das heißt, die Konstruktvalidierung kann sowohl experimentelle als auch korrelative Untersuchungen umfassen und erfordert grundsätzlich die vorherige Formulierung von Hypothesen über die zu erfassenden Konstrukte und über deren Beziehungen. Die im Rahmen einer Konstruktvalidierung zu prüfenden Hypothesen können sich beispielsweise auf den Zusammenhang der Testwerte mit Testwerten in einer Kriteriumsvariablen bzw. Maßen anderer Konstrukte beziehen, aber genauso auch auf die intrapsychischen Prozesse, die für die Bearbeitung der Testaufgaben nötig sind („Inhaltsvalidität"). Dementsprechend werden in vielen modernen Darstellungen Kriteriumsvalidität und Inhaltsvalidität als untergeordnete Aspekte der Konstruktvalidität betrachtet. Die traditionelle Dreigliederung in Kriteriums-, Inhalts- und Konstruktvalidität, wie sie etwa noch bei Lienert und Raatz (1998) üblich war, kann daher mittlerweile als überholt angesehen werden (Hubley & Zumbo, 1996), nicht nur aus der Perspektive von Messick (1989), sondern auch aus der Perspektive der aktuellen alternativen Validitätskonzeption von Borsboom, Mellenbergh und van Heerden (2004).

Der vorliegende Beitrag orientiert sich wesentlich an der Validitätskonzeption von Messick (1989), da diese im Vergleich mit anderen umfassender erscheint und sich weitestgehend von einer spezifischen erkenntnistheoretischen Position freizuhalten bemüht. In loser Anlehnung an Messick (1989) definieren wir Validität als Qualitätsanforderung daher wie folgt:

> Validität ist eine Anforderung an Testwerte. Das Erheben von Testwerten dient im Allgemeinen einem gewissen Zweck; je stärker die empirischen Belege darauf hinweisen und die theoretischen Herleitungen sicherstellen, dass Interpretationen der Testwerte und Schlussfolgerungen aus diesen Interpretationen diesem Zweck gerecht werden, d. h. je besser die Interpretationen der Testwerte und Schlussfolgerungen anhand wissenschaftlicher Kriterien begründet werden können, desto höher die Validität.

Im Vergleich mit den Formulierungen von Borsboom et al. (2004; „Ein Test ist valide für eine Eigenschaft, wenn a) diese Eigenschaft existiert und b) Variation in der Eigenschaft Variation in den Testwerten kausal erzeugt", S. 1061; Übersetzung durch die Autoren) oder mit der Definition von Lienert[4] (1969; „Die Validität eines Testes gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er mes-

4 In Lienerts (1969) Definition ist darüber hinaus der Begriff der Genauigkeit, der ja genauso auch in der Reliabilitätsdefinition verwendet wird, äußerst unglücklich.

sen soll oder zu messen vorgibt, tatsächlich misst", S. 16) erscheint die hier vorgenommene Definition möglicherweise zunächst kompliziert und technologisch bzw. pragmatisch-funktional, da der „Zweck" der Erhebung von Testwerten einbezogen wird. Aber selbst derjenige Wissenschaftler, der „nur" Erkenntnis sucht, verfolgt ein Ziel und erhebt Testwerte zu diesem Zweck. Vor allem aber wird in der hier vorgenommenen Definition wie schon bei der Definition des Konstruktbegriffs das Problem der Existenz eines zu messenden Konstruktes ausgeklammert. Daher kann im Rahmen dieser Definition sowohl unter einer konstruktivistischen oder funktionalistischen Perspektive als auch unter einer radikal-realistischen Perspektive eine Einschätzung der Validität von Testwerten vorgenommen werden. Dennoch sei erwähnt, dass sich manche Forscher wohl am zugegebenermaßen unklaren Begriff der „Begründbarkeit" stören mögen; die dahinter stehende Annahme ist, dass die Entscheidung darüber, was als wissenschaftlich „wahr" oder zumindest „gesicherte Erkenntnis" akzeptiert wird, immer auch mehr oder weniger von einer sozialen Übereinkunft der wissenschaftlichen Gemeinschaft abhängt. Die Definitionen von Lienert (1969) oder Borsboom et al. (2004) dagegen behandeln das Problem, wie die Erfüllung der jeweiligen Validitätsanforderungen gezeigt werden kann, tatsächlich überhaupt nicht. Schließlich macht die Validitätsdefinition nochmals deutlich, dass die von Hubley und Zumbo (1996) ironisch als „Dreifaltigkeit" bezeichnete Dreiteilung von Kriteriums-, Inhalts- und Konstruktvalidität obsolet ist, da für die Begründbarkeit von Schlussfolgerungen und Handlungen aus Testwerten letztlich wissenschaftliche Belege unterschiedlicher Art gemeinsam angeführt werden können und müssen (vgl. Hartig, Frey & Jude, 2007).

Für die weitere Darstellung der KTT und ihrer Erweiterungen ist zu beachten, dass die hier gewählte Validitätskonzeption sehr viel umfassender ist als diejenige, die in psychometrischen Veröffentlichungen gerade etwas älteren Datums zu finden ist. Noch stärker als beim Begriff der Reliabilität stellen wir nicht nur an uns selbst, sondern auch an zukünftige Veröffentlichungen die Forderung, explizit und deutlich zwischen der Validität als Qualitätsanforderung und den empirischen Statistiken numerischen Größen, die als Belege für die Validität angeführt werden, zu unterscheiden.

1.3.3 Testwert, Subtests, Testlets und Testitems

Die Begriffe „Test" und „Testwert" wurden bereits in Abschnitt 1.2 eingeführt. Insbesondere Psychologische Tests bestehen in der Regel aus mehreren unterscheidbaren Teilen: Die kleinste Einheit, in die ein psychologischer Test zerlegt werden kann, wird im Allgemeinen als *Item* bezeichnet. Sehr häufig werden aber in einem hierarchischen Schema auch mehrere Items zu *Subtests* zusammengefasst, und zwar fast immer auf der Grundlage eines inhaltlichen Zusammen-

hangs. Beispielsweise enthielt der in PISA 2003 eingesetzte Mathematiktest drei „Subskalen[5]“ (vgl. Prenzel et al., 2004). Derselbe Test enthält aber noch eine weitere Gliederungsebene: Die „Subskalen“ sind aus mehreren (komplexen) „Aufgaben“ zusammengesetzt, deren kleinste Einheiten nun die Items darstellen, wobei sich die Items auf unterschiedliche Aspekte derselben Aufgabe beziehen. Um Konsistenz mit der angloamerikanischen Literatur zu erzielen und um Verwechslungen vorzubeugen[6], soll für die erste Gliederungsebene oberhalb der Itemebene im Folgenden der Begriff *Testlet* verwendet werden. Im (häufigeren) Fall, dass ein Test nur eine Gliederungsebene oberhalb der Itemebene hat, werden die Begriffe „Subtest“ und „Testlet“ synonym verwendet.

Der Begriff *Testlet*, der insbesondere auch im Rahmen von IRT-Modellen populär geworden ist, wurde ursprünglich verwendet, um eine schwierigkeitshomogene Itemgruppe im Rahmen eines adaptiven Tests zu bezeichnen, eine „Gruppe von Items, die sich auf ein einziges Inhaltsgebiet beziehen und als Einheit entwickelt werden“ (Wainer & Thissen, 1996, S. 190; Übersetzung durch die Autoren). In späteren Veröffentlichungen (Wainer & Thissen, 1996; Lee, Brennan & Frisbie, 2000) wurde der Umfang des Begriffes allerdings erweitert: Ein Testlet ist demnach eine Gruppe von Items, die sich auf irgendeine Art (statistisch und/oder inhaltlich) gegen Items eines anderen Testlets abgrenzen lassen. Testlets können außerdem *a priori* definiert werden, wie es etwa in der ursprünglichen Definition der Fall ist. Aber häufig zeigen empirische Ergebnisse erst *post hoc* die Notwendigkeit einer gesonderten Behandlung einzelner Testlets.

2 True Scores und Fehlerwerte

2.1 Definition der KTT

Im Speziellen definieren wir den Begriff „Klassische Testtheorie“ wie folgt:

> Die Klassische Testtheorie (KTT) stellt mathematische Regeln für die Analyse der Variabilität von Testwerten zur Verfügung. Sie ist ein Bündel von Annahmen, das historisch auf unterschiedliche Weise formalisiert wurde. Die wesentlichsten Annahmen sind die folgenden: Für jeden Testwert existiert ein „wahrer“ Wert (True Score; Existenzannahme), der nicht mit dem „beobachteten“ Testwert identisch sein muss, aber mit dem beobachteten Wert linear zusammenhängt (Additivitätsannahme). Aus der Differenz von

5 Der Begriff „Skala“ oder „Subskala“ ist allerdings mehrdeutig und bezieht sich zuweilen auch auf das Ergebnis von messtheoretischen Operationen oder auf das Ergebnis der Anwendung eines IRT-Modells.

6 Die Begriffe „Item“ und „(Test-)Aufgabe“ werden – anders als im PISA-Test – meist synonym verwendet.

wahren und beobachteten Werten ergeben sich „Messfehler", die untereinander und mit den wahren Werten unkorreliert sind.

Natürlich hat eine solche Definition den Schönheitsfehler, dass die meisten der verwendeten Begriffe wiederum definitionsbedürftig sind. Entsprechend wird ein großer Teil des folgenden Abschnitts für die zentralen Begriffe „True Score" und „Messfehler" verwendet. Tatsächlich wollen wir der Definition von True Scores deutlich mehr Platz einräumen als es in den meisten Darstellungen der KTT üblich ist; dieses Problem ist von fundamentaler Bedeutung für Überlegungen zur Reliabilitätsschätzung und für Fragen der Testvalidität und Generalisierbarkeit.

Darüber hinaus werden unter dem Oberbegriff KTT aber nicht nur die in der Definition genannten Konzepte von True Score und Fehlerwerten mit den daraus folgenden Ableitungen behandelt; Stumpf (1996) etwa definiert die KTT zunächst als „ein Arsenal pragmatisch orientierter Prinzipien oder Regeln zu Konstruktion, Erprobung und Evaluation psychometrischer Tests und zur Interpretation von Testergebnissen" (S. 411). Im Rahmen eines solchermaßen erweiterten „klassischen" Arsenals von Prinzipien zur Testkonstruktion und -evaluation werden etwa Begriffe wie die der Itemanalyse (vgl. Kelava & Moosbrugger, 2007) zumindest implizit ebenfalls dem Sachgebiet der KTT zugeordnet, auch wenn Konzepte wie „Itemschwierigkeit", „Itemvarianz" oder „Trennschärfe" in Modellen der IRT meist eine sehr viel genauere Behandlung erfahren (vgl. etwa Rauch, Schweizer & Moosbrugger, 2008, für ein didaktisch orientiertes Anwendungsbeispiel).

2.2 Einführung in das Problem der Messwiederholung

Der zentrale Gesichtspunkt der KTT ist die additive Zusammensetzung von wahrem Wert und Messfehler, aufbauend auf der Beobachtung, dass Testwerte bei Wiederholung der Messung variieren, selbst unter der theoretischen Annahme, dass diese stabil sein sollten. Das Grundproblem besteht in den Fragen, aus welchem Grund die (unerwünschte) Variation, nämlich die Messfehler entstehen, was die „wahren" (nämlich nicht variierenden) Werte bedeuten und wie in diesem Zusammenhang die Reliabilität der Testwerte geschätzt werden könnte.

2.2.1 Das Uhrenbeispiel

Brennan (2001) führt in das Problem der Variabilität von Testwerten und der damit verbundenen *Un*zuverlässigkeit wie folgt ein: „Eine Person mit einer Uhr weiß, wie spät es ist; eine Person mit zwei Uhren ist nie ganz sicher" (S. 7; Übersetzung durch die Autoren). Dieses einfach erscheinende Beispiel lässt sich heranziehen, um zentrale Konzepte und Problembereiche der KTT (aber auch der psychologischen Diagnostik im Allgemeinen) zu verdeutlichen.