ps psychologie

Markus Bühner

Einführung in die Test- und Fragebogenkonstruktion

3., aktualisierte Auflage

PEARSON Studium

Einführung in die Test- und Fragebogenkonstruktion

Markus Bühner

Einführung in die Test- und Fragebogenkonstruktion

3., aktualisierte und erweiterte Auflage

PEARSON Studium

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

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über *http://dnb.ddb.de* abrufbar.

10 9 8 7 6 5 4 3 2 1

13 12 11

ISBN: 978-3-8689-4033-6

ein Imprint der Pearson Education Deutschland GmbH,
Martin-Kollar-Straße 10-12, D-81829 München/Germany

www.pearson-studium.de
Lektorat: Andra Riemhofer, ariemhofer@pearson.de; Alice Kachnij, akachnij@pearson.de
Korrektorat: Margret Neuhoff, München
Einbandgestaltung: adesso 21, Thomas Arlt, München
Herstellung: Claudia Bäurle, cbaeurle@pearson.de
Satz: mediaService, Siegen (www.media-service.tv)
Druck und Verarbeitung: Becker, Kevelaer

Printed in Germany

Inhaltsverzeichnis

Vorwort zur 3. Auflage

Der Erfolg des Buches hat auch in der zweiten Auflage unverändert angehalten. Da nun seit deren Erscheinen schon wieder vier Jahre vergangen sind, folgt jetzt die dritte Auflage dieses Buches in einer Neukonzeption unter Mitarbeit von Moritz Heene und Matthias Ziegler. Die strikte Anwendungsorientierung ist jedoch von der Neukonzeption unberührt.

In die Neukonzeption sind neben der Lehrerfahrung der vergangenen Jahre auch die Beratungspraxis sowie Erfahrungen aus der Begutachtung von Zeitschriftenartikeln eingeflossen. Da viele Studierende mehr Hintergrundinformationen nachfragten, wurde die Beschreibung der Grundlagen durch Exkurse stark erweitert und alle Kapitel neu und übersichtlicher strukturiert. Es finden sich jetzt auch mehr Checklisten im Buch, die die Umsetzung von Analysen und die Bewertung von Tests erleichtern sollen. Darüber hinaus wurden auch die Anwendungsteile der Kapitel stark erweitert. Kapitel 5 widmet sich nun beispielsweise ausschließlich der Umsetzung der Testkonstruktion mit den Anwendungsprogrammen SPSS und G*Power. In diesem Kapitel sind jetzt auch die Testnormierung sowie die Cut-off-Wertbildung genau beschrieben. Es liegen nun auch neue Datensätze vor, mit deren Hilfe man verschiedenste Analysemethoden veranschaulichen kann. Es ist auch notwendig geworden, neue Anwendungsprogramme mit einzubeziehen. Manche Analysen, die sich als sehr sinnvoll erwiesen haben, lassen sich nicht mit den bekannten Softwarepaketen SPSS, AMOS und WINMIRA umsetzen. Daher sind im Buch nun auch Analyseprozeduren mit den Programmpaketen *R*, *G*Power* und *MPlus* beschrieben. Besonders viel Raum im Vergleich zur vorangegangenen Auflage wird in dieser Auflage dem Prozess der Itemkonstruktion bzw. Testkonstruktion eingeräumt. Das neue didaktische Konzept ist in Kapitel 1 ausführlich dargelegt, und ich hoffe, dass die dritte Auflage von den Lesern ebenso gut angenommen wird wie die beiden vorangegangenen Auflagen.

An dieser Stelle möchte ich mich nun bei den Menschen bedanken, die zum Gelingen des Buches sehr viel beigetragen haben. An erster Stelle natürlich Moritz Heene und Matthias Ziegler. Matthias hat mich insbesondere bei Kapitel 3 unterstützt sowie beim Schreiben des Kapitels über Latent-State-Trait-Modelle, und Moritz beim Schreiben von Kapitel 7 sowie bei der Herstellung von Abbildungen im Buch. Moritz und Matthias haben das komplette Buch gelesen und sehr viele konstruktive Verbesserungsvorschläge gemacht. Herrn Prof. Dr. Helmut Küchenhoff (StabLab) danke ich herzlich für seine Verbesserungsvorschläge, die vielen hilfreichen Diskussionen und die ausgezeichnete Zusammenarbeit, sowie Herrn Prof. Dr. Martin Arendasy für seine konstruktiven Hinweise zu Kapitel 8 und zur AIG. Auch Julia Gsottschneider, Alexandra Welz und Alexandra Zech möchte ich besonders erwähnen. Julia hat alle Kapitel gründlich Korrektur gelesen, und ihre Anmerkungen und Korrekturen waren einfach klasse. Alexandra Zech hat mich bei der Beschreibung des Programmpakets *eRm* in *R* unterstützt und Alexandra Weltz beim Erstellen des Literaturverzeichnisses und der Korrektur. Herrn Schneider vom Pearson Verlag danke ich herzlich für die lange und ausgezeichnete Zusammenarbeit und für die Möglichkeit, überhaupt mein Lehrkonzept in einem Buch zu veröffentlichen. Mein Dank gilt auch den Kollegen und Studierenden, die mich auf Inkonsistenzen, Fehler und Verständnisschwierigkeiten hingewiesen haben. Darüber hinaus danke ich einem tollen Team, das mich beim

Durchsehen der Druckfahnen unterstützt hat: Stella Bollmann, Marie Peterseil, Patricia Weber, Hannah Klaas, Daniela Konrad, Kathrin Ortner, Sabine Krenn und Sylvia Opriessnig. Auch Harry Freudenthaler möchte ich herzlich danken. Er hat sich mit mir die emotionalen Intelligenztestitems aus Kapitel 9 einfallen lassen und mich beim Schreiben der Alphamax-Story tatkräftig unterstützt.

Auch in der dritten Auflage meines Buches gilt ein ganz besonderer Dank meinen Eltern Lothar und Barbara Bühner, meiner Partnerin Sabine Schaal und meiner Tochter Luisa, die sehr oft auf einen chronisch überarbeiteten Menschen gestoßen sind, mit dem manchmal sehr wenig anzufangen war.

Es sei mir auch erlaubt, in dieser Auflage meinem Freund Prof. Dr. Friedrich Försterling zu gedenken, der leider am 6. August 2007 viel zu früh verstorben ist und für mich ein ganz besonderer Mensch und Wissenschaftler war. Ich bedanke mich auch sehr herzlich bei Frau Kachnij und Frau Riemhofer vom Pearson-Verlag für ihre Unterstützung bei den Korrekturen im Buch.

Am Ende dieses Vorworts möchte ich darauf hinweisen, dass ich stark an konstruktiver Kritik zu diesem Buch interessiert bin. Ihre Anmerkungen können Sie ganz einfach per E-Mail an die folgende Adresse senden: *psychologie@pearson-studium.de*.

Markus Bühner

Einführung

1

ÜBERBLICK

1

Dieses einführende Kapitel dient zum einen zur Einstimmung auf die folgenden Kapitel und enthält daher zunächst in *Kapitel 1.1* eine kurze Beschreibung der Inhalte der folgenden *Kapitel 2 bis 8*. Im anschließenden *Kapitel 1.2* soll zum anderen die Motivation geschaffen werden, sich mit dem Thema intensiver zu befassen. Das folgende *Kapitel 1.3* enthält eine Klassifikation von psychologischen Tests und eine kurze Darstellung eines Tests aus jeder Kategorie. Im letzten *Kapitel 1.4* wird beschrieben, welche Diagnosemöglichkeiten mithilfe von Tests bestehen.

1.1 Ziel des Buches

Ziel des vorliegenden Buches ist es, den Lesern ein Grundverständnis der Testkonstruktion zu vermitteln. Dabei wird auf Verständlichkeit und Anwendungsbezug besonderer Wert gelegt. Es handelt sich bei den Inhalten um eine Mischung aus Grundlagen und Anwendungshinweisen für Statistiksoftware. Der Leser soll nach der Lektüre des Buches in der Lage sein, einen psychologischen Test zu bewerten und auch einen Fragebogen oder Test selbst zu konstruieren. Dazu werden die für die Testkonstruktion notwendigen Grundlagen dargestellt. Im Rahmen einer neu eingeführten **Einstiegshilfe** wird auf **Voraussetzungen in Kapiteln anderer Lehrbücher** verwiesen, die ein besseres Verständnis der beschriebenen Inhalte ermöglichen. Das Buch eignet sich nicht, um völlig frei von jedem Vorwissen einen Test schnell zu konstruieren. Dazu ist der Prozess der Testkonstruktion insgesamt zu kompliziert. Um jedoch Lesern mit entsprechendem Vorwissen (Grundlagen der Statistik) die Testkonstruktion verständlich zu vermitteln, erfolgt eine besonders detaillierte Darstellung, wie man von den theoretischen Grundlagen mithilfe von **Statistiksoftware** und Interpretationen der Ergebnisausgaben der Anwendungsprogramme einen Test praktisch von der Konstruktion bis zur Normierung erstellen kann. Die Kapitel enthalten dabei didaktische Elemente wie **Exkurse**, die eine Vertiefung der Inhalte ermöglichen, sowie **Praxistipps**, welche die Umsetzung der theoretisch dargestellten Sachverhalte mithilfe von Statistikprogrammen wie SPSS, G*Power, AMOS, MPlus, R oder WINMIRA aufzeigen. Die Zusammenfassungen am Ende der Kapitel dienen als Gedankenstütze. **Checklisten** werden vor allem in den anwendungsbezogenen Kapiteln präsentiert, sie dienen als Strukturierungs- und Entscheidungshilfe im Bewertungs- oder Konstruktionsprozess. Im Folgenden wird zu jedem Kapitel in diesem Buch ein kurzer inhaltlicher Abriss gegeben.

Kapitel 2: Testtheoretische Grundlagen

Im zweiten Kapitel werden zwei unterschiedliche Arten dargestellt, ein Konstrukt zu definieren: zum einen mit **reflektiven**, zum anderen mit **formativen Indikatoren**. Konstrukte sind nicht beobachtbare Eigenschaften oder Fähigkeiten, die mithilfe von Indikatoren erschlossen werden. Indikatoren sind Items, die beispielsweise Einstellungen oder Verhalten erfassen. Formative Indikatoren stellen Indizes dar. Die Indikatoren des Indexes können dabei unkorreliert sein. Reflektive Indikatoren von Konstrukten sind hingegen korreliert, und deren Interkorrelationen werden durch so genannte **latente Variablen** erklärt. Die nachfolgend dargestellten beiden Testtheorien beziehen sich dabei nur auf Konstrukte mit reflektiven Indikatoren. Zunächst wird die **Klassische Testtheorie** dargestellt. Sehr viele Tests basieren immer noch auf dieser Theorie. Sie stellt den theoretischen Rahmen zur Schätzung der Messgenauigkeit eines Messwerts zur Verfügung. Diese Schätzung ist jedoch an bestimmte Annahmen gebunden, die in diesem Kapitel ebenfalls vorgestellt werden. Die klassische Testtheorie lässt

sich sehr knapp schildern, dabei wird jedoch nicht deutlich, an welche Annahmen gerade die Reliabilitätsschätzung gebunden ist. Die zweite Testtheorie, die vorgestellt wird, ist die **Probabilistische Testtheorie**, wobei in diesem Zusammenhang nur kurz auf das Rasch-Modell eingegangen wird. Da dieses Modell für die Testkonstruktion besonders wichtig ist, wird es in Kapitel 8 noch einmal detaillierter dargestellt. Der zweite Teil von Kapitel 2 beschäftigt sich mit **Gütekriterien** von Tests bzw. Messwerten. Dabei werden **Haupt- und Nebengütekriterien** unterschieden. Am Ende des Kapitels wird eine **Kurzcheckliste zur Beurteilung von Tests** vorgestellt, die für Einsteiger eine Hilfe bei der Testbeurteilung darstellen soll. Darüber hinaus werden Hinweise gegeben, wie man sich über die Qualität bestimmter Testverfahren informieren kann.

Kapitel 3: Testkonstruktion

Im dritten Kapitel wird der **Konstruktionsprozess** eines Tests skizziert. Im Groben können drei Schritte der Testkonstruktion unterschieden werden: die **Erstellung des Testentwurfs**, die empirische Überprüfung des Testentwurfs und die Normierung und/oder Cut-Off-Ermittlung der endgültigen Testversion. Dabei wird nur der erste Schritt in Kapitel 3 näher erläutert. Zunächst erfolgt die Festlegung, ob **subjektive** oder **objektive Indikatoren** zur Erfassung einer Eigenschaft oder Fähigkeit genutzt werden sollen. Im Anschluss daran muss festgelegt werden, welche **Zielgruppe** untersucht und welche **Konstruktionsstrategie** angewendet werden soll, beispielsweise, ob Gruppen getrennt oder die Ausprägungen bestimmter Eigenschaften oder Fähigkeiten von Personen festgestellt werden sollen. Sind diese Schritte getan, müssen **Indikatoren für das Konstrukt** gesammelt und eine **Eingrenzung des Konstrukts** vorgenommen werden. Dies dient als Basis für die Erstellung einer **Definition des Messgegenstandes**. Anhand der Definition erfolgt dann die **Auswahl eines Itemformats**. Dabei werden die Vor- und Nachteile der verschiedenen Itemformate besprochen sowie allgemeine Probleme der Itembeantwortung, beispielsweise Verfälschbarkeit von Fragen. Am Ende des Kapitels werden dann Hinweise zur **sprachlichen Gestaltung von Items** gegeben.

Kapitel 4: Reliabilität

Das vierte Kapitel widmet sich sehr ausführlich der **Reliabilitätsschätzung**. Es wird zunächst erklärt, wie die **Reliabilität** im Rahmen der Klassischen Testtheorie **definiert** ist. Im nächsten Schritt werden die **Voraussetzungen** dargelegt, welche zur präzisen Schätzung der Reliabilität notwendig sind, und schließlich unterschiedliche **Methoden der Reliabilitätsschätzung** beschrieben. Schließlich erfolgt die Darstellung von **Formeln zur Reliabilitätsschätzung**. Dabei ist diese Darstellung auf die Formeln beschränkt, die häufiger in Testhandbüchern zu finden sind. Im Anschluss daran wird auf die **Trennschärfeberechnung** eingegangen. Die Höhe der Trennschärfen von Items beeinflusst direkt die Reliabilität der Messwerte. Andere Einflussgrößen sowie **Möglichkeiten der Reliabilitätsverbesserung** sind ebenfalls dargestellt. Im vorletzten Abschnitt des vierten Kapitels wird auf die **psychometrische Einzelfalldiagnostik** eingegangen. Es handelt sich hierbei um eine sehr wichtige Anwendung der Reliabilität, da der Messfehler eines Tests bei der Interpretation von Messwerten berücksichtigt werden muss. Dies gilt in gleichem Maße für die Interpretation von Messwertunterschieden einer Person in zwei Tests, einer Person im selben Test nach einer Intervention, im Rahmen einer Verlaufsdiagnostik oder von zwei unterschiedlichen Personen in einem Test. Im letzten Abschnitt des Kapitels wird beschrieben, wie man **Korrelationen von Tests so schätzen** kann, dass sie **messfehlerfrei** sind.

1

Kapitel 5: Itemanalyse

Kapitel fünf ist neu und umfasst die gesamte **Itemanalyse nach der Klassischen Testtheorie mit SPSS** sowie den gesamten Prozess der Normierung bzw. Ermittlung von Cut-off-Werten. Diese Schritte schließen direkt an die Testkonstruktion, wie sie in Kapitel 3 geschildert wurde, an. Zunächst werden die **Itemcodierung** und die **Schwierigkeits-** sowie die **Verteilungsanalyse** der Items vorgestellt. **Trennschärfe** und **Reliabilitätsanalyse** sind gemeinsam in diesem Kapitel dargestellt und nicht mehr wie in der alten Auflage getrennt. Auch eine **Bewertung der inhaltlichen Passung der Items** ist nun in die dargestellte Analyse implementiert. Im Anschluss werden die **Zusammenhänge zwischen Itemschwierigkeit, Standardabweichung, Trennschärfe und Reliabilität** am dargestellten Beispiel erläutert. Im zweiten Teil des Kapitels wird das Thema **Normierung**, **norm- und kriterienorientiertes Testen** sowie die **Erstellung von Cut-off-Werten** für die Gruppentrennung beschrieben sowie deren praktische Ermittlung mithilfe von SPSS dargestellt.

Kapitel 6: Exploratorische Faktorenanalyse

Das sechste Kapitel stellt die **exploratorische Faktorenanalyse** vor. Im Rahmen der Testkonstruktion ist sie praktisch nicht zu ersetzen. So ist beispielsweise das Auffinden von Items mit korrelierten Messfehlern mithilfe der explorativen Faktorenanalyse sehr effektiv, da hier alle Itemladungen auf allen Faktoren angezeigt werden. Liefert eine konfirmatorische Faktorenanalyse ein nur schwer zu interpretierendes Ergebnis, wird in der Regel auch auf eine exploratorische Faktorenanalyse zurückgegriffen, um neue Hypothesen über die Faktorenstruktur zu generieren. Dargestellt wird zunächst das **Modell mehrerer gemeinsamer Faktoren** mit einem kleinen Exkurs zur Testtheorie und dann der **Grundgedanke der Faktorenanalyse**. Im Fall der Testkonstruktion ist eine Faktorenstruktur dann günstig, wenn die Itemkorrelation nur durch einen Faktor erklärt werden kann. Der erste Schritt, der überlegt werden muss, ist die **Wahl der faktorenanalytischen Methode**. Daher werden zwei gängige faktorenanalytische Methoden vorgestellt: Hauptachsenfaktorenanalyse und Maximum-Likelihood-Faktorenanalyse. Als ausführlicher Exkurs wird auch die Hauptkomponentenanalyse beschrieben, die eigentlich gar keine faktorenanalytische Methode darstellt. Sie wird jedoch häufig angewandt und deshalb ausführlich behandelt. Die sich daran anschließende Frage ist, **wie viele Faktoren extrahiert werden sollen**. Dazu werden mehrere Methoden vorgestellt, z.B. das Eigenwertkriterium größer Eins, der Scree-Test, die Parallelanalyse und der Minimum-Average-Partial-Test. Schließlich entscheidet man sich für eine **Rotationsmethode**, bei der man davon ausgeht, dass die Faktoren entweder korreliert sind (oblique, z.B. Promax-Rotation) oder unkorreliert (orthogonal, z.B. Varimax-Rotation). Ziel der Rotationsmethode ist, eine gut interpretierbare Faktorenstruktur zu erhalten. Daher wendet sich das anschließende Kapitel diesem Thema zu. Wichtig im Rahmen der Faktorenanalyse ist auch, die Ausprägung der Personen auf den Faktoren zu bestimmen. Man bezeichnet diese Werte auch als **Faktorwerte**. Die verschiedenen Methoden, Faktorwerte zu schätzen, werden ebenfalls in einem Kapitel beschrieben. Bevor dann die konkrete Durchführung vorgestellt wird, werden **Voraussetzungen für die Durchführung** der Faktorenanalyse beschrieben. Schließlich werden **Kriterien zur Bewertung der Durchführbarkeit einer Faktorenanalyse** beschrieben, wie beispielsweise Maße der Eignung der Korrelationsmatrix für die Faktorenanalyse oder die Reliabilität von Faktoren. Am Ende des theoretischen Blocks wird eine Checkliste zur Durchführung

einer Faktorenanalyse vorgestellt, die dem Anwender bei den zu tätigenden Entscheidungen helfen soll. In den beiden letzten Teilen des Kapitels erfolgen die **Darstellung der Durchführung einer Faktorenanalyse mit dem Programm SPSS** und die Interpretation der Ergebnisausgaben für die Faktorenstruktur der Extraversions- und Neurotizismusskalen aus dem NEO-FFI. Darüber hinaus wird auch kurz auf die Durchführung von Faktorenanalysen mit dichotomen oder ordinalen Items mithilfe von **MPlus** eingegangen.

Kapitel 7: Konfirmatorische Faktorenanalyse

Die **konfirmatorische Faktorenanalyse** wird zur **Überprüfung eines a priori bestehenden, gut fundierten theoretischen Modells** eingesetzt. Sie wird im Rahmen der Testkonstruktion insbesondere dafür benötigt, um die Voraussetzungen zu prüfen, die für die Reliabilitätsbestimmung eines Tests notwendig sind. Im Rahmen einer Itemanalyse können diese Voraussetzungen nicht geprüft werden. Die konfirmatorische Faktorenanalyse kann aber auch zur Überprüfung der Validität von Tests durchgeführt werden. Zunächst wird in diesem Kapitel wie im vorangegangenen Kapitel das **Grundmodell** beschrieben und dargestellt, wie man Modelle grafisch veranschaulicht, sowie die **Notation** (Bezeichnung) der Parameter (z.B. Ladungen, Fehlervarianzen, Korrelationen) einer konfirmatorischen Faktorenanalyse vorgestellt. Anschließend wird gezeigt, wie man ein Modell in **Definitionsgleichungen** zerlegt, so dass man im Anschluss mithilfe von **Strukturgleichungen** die Parameter für das zu prüfende Modell schätzen kann. Bevor jedoch die Schätzung der so genannten Modellparameter möglich ist, muss das Modell identifiziert sein. **Identifikation** meint, dass mehr bekannte als unbekannte Werte in den Strukturgleichungen enthalten sind und dass latenten Variablen und Fehlervariablen eine Einheit zugewiesen werden kann. Schließlich wird gezeigt, wie die **Parameterschätzung** erfolgt. Nachdem nun alle Korrelationen, Ladungen und Fehlervarianzen des Modells geschätzt wurden, bleibt die Frage offen, ob das so spezifizierte Modell durch die Daten verworfen bzw. abgelehnt werden muss oder nicht. Diese Frage wird durch **Modelltests** beantwortet. Man kann hier Modelle mithilfe von statistischen Tests, beispielsweise dem χ^2-Test (Chi-Quadrat-Test), oder auch mithilfe von so genannten Fitindizes überprüfen. Es wird in diesem Kapitel auch versucht, den Nutzen konfirmatorischer Faktorenanalysen für die Testkonstruktion zu verdeutlichen. Dazu gehört ein Hinweis auf die Äquivalenz von Messungen und des **Multitrait-Multimethod-Ansatzes**. Bevor letztendlich die **Durchführung und Interpretation** von Strukturgleichungsmodellen mithilfe des **Programmpakets AMOS** beschrieben wird, werden Möglichkeiten der **Modellmodifikation** sowie **Schätzprobleme** besprochen. Neu hinzugekommen ist im Anwendungsteil die Darstellung von **Latent-State-Trait-Modellen.**

Kapitel 8: Probabilistische Testtheorie

In diesem Kapitel wird eine kleine Auswahl probabilistischer Testmodelle vorgestellt. Einen besonderen Stellenwert nimmt dabei das Rasch-Modell ein. Dieses hat Eigenschaften, die dem Ideal einer echten Messung näher kommen als jedes andere Testmodell. Zunächst erfolgt eine kleine Einführung zum Thema **Messen**, welche die Sinnhaftigkeit der Anwendung Probabilistischer Testmodelle deutlich macht. Vorgestellt werden dann die Rasch-Modelle für **dichotome** und **ordinale Daten** sowie das **Mixed-Rasch-Modell**. Das Mixed-Rasch-Modell ermöglicht nicht nur die Quantifizierung der Testleistung, sondern auch die Zuordnung von Personen zu Klassen. In der Praxis

sind neben dem Rasch-Modell noch das so genannte **Birnbaum-Modell** sowie das **Drei-Parameter-Modell**, in dem auch die Ratewahrscheinlichkeit bei einer Aufgabe berücksichtigt wird, weitverbreitet. Die Schätzung von Item- und Personenparameter wird im Rahmen des Rasch-Modells erklärt. Dabei steht der Itemparameter für die Schwierigkeit eines Items und der Personenparameter quantifiziert die Leistung oder Eigenschaftsausprägung einer Person in einem Test. Ein Vorteil des Rasch-Modells ist auch die Testbarkeit des Modells und der Annahmen, die mit dem Modell verbunden sind. Daher werden auch **Modelltests** im Rahmen des Rasch-Modells in Kapitel 8 vorgestellt. Schließlich wird noch auf das **adaptive Testen** eingegangen, welches im Rahmen probabilistischer Testmodelle möglich ist. Am Ende des Kapitels wird die Durchführung einer Rasch-Analyse und einer Mixed-Rasch-Analyse mit **WINMIRA** dargestellt, dabei wird auch kurz auf das **Programmpaket R** eingegangen.

1.2 Testanwendungsbereiche

Warum muss ich wissen, wie ein Test entwickelt wird?

Testverfahren werden in der Forschung und in den Anwendungsbereichen der Psychologie (Neuropsychologie, Klinische Psychologie, Arbeitspsychologie, Werbepsychologie usw.), der Medizin sowie der Sozial- und Wirtschaftswissenschaften eingesetzt. Für viele Fragestellungen existieren bereits Testverfahren. Aber für sehr viele andere sind keine Testverfahren oder Fragebögen verfügbar, vor allem dann, wenn sehr spezifische Fragestellungen zu beantworten sind. Die folgende Auflistung macht deutlich, dass in zahlreichen Anwendungsbereichen die Testentwicklung eine bedeutende Rolle spielt.

Kliniken

- Psychische Störungen
- Psychosomatische Störungen
- Somatische Störungen
- Auf somatische Ursachen zurückgehende psychische Störungen (z.B. Hirnschäden)
- Belastung und Schmerz
- Krankheitsbewältigung
- Therapieverlaufs- und Erfolgskontrolle

Beratungsstellen

- Allgemeine psychosoziale Beratung
- Familienberatung
- Eheberatung
- Erziehungsberatung
- Suchtberatung

Staatliche Verwaltung

- Berufsberatung
- Auslese
- Jugendhilfe
- Verkehrseignung (TÜV)

Forensischer Bereich

- Strafvollzug (Haftentlassung)
- Im Rahmen von Gerichtsverfahren (zivil- und strafrechtliche Verantwortlichkeit, Glaubwürdigkeit von Zeugen, Sorgerecht)

Betriebe/Personalverwaltung

- Eignung/Auslese
- Arbeitsplatzgestaltung/Ergonomie
- Arbeitsanalyse
- Personalentwicklung

Schulen, pädagogischer Bereich

- Entwicklung/Entwicklungsstörungen
- Lernprozesse
- Schulreife
- Sonderschulbedürftigkeit/geistige Behinderung
- Spezifische Lernschwierigkeiten
- Leistungsstörungen
- Hochbegabung
- Schulleistung
- Hochschuleignung
- Intelligenzdiagnostik allgemein

Militär

- Eignungsprüfungen

Marktforschung/Werbepsychologie

- Produktbeurteilung
- Werbung
- Einstellungsmessung

Forschung

- Einstellungen/Interessen
- Eigenschaften
- Momentane Zustände
- Verlaufsprozesse

Die Erfahrung zeigt, dass es Studenten anfangs oft unverständlich erscheint, warum sie sich mit statistischen Methoden zur Testkonstruktion beschäftigen sollen. Daher ist es sinnvoll, am Anfang plausibel zu machen, warum sich der Erwerb von Kenntnissen in diesem Bereich lohnt.

Nach einer Untersuchung von Jiménez und Raab (1999) sind Aufgabenbereiche, die Psychologen von Laien zugeordnet werden „Tests zur Vorhersage von Berufseignung" und „Feststellung der Schulreife von Kindern". Grundlegendes Wissen auf diesem Gebiet ermöglicht nicht nur die Konstruktion von Test- und Fragebogenverfahren, sondern ist auch Grundvoraussetzung für die Beurteilung bereits bestehender Tests. So richten sich Auswahl und Interpretation von Test- und Fragebogenergebnissen nach so genannten Testgütekriterien. Testgütekriterien geben Auskunft über die Genauigkeit und den Vorhersageerfolg der Testergebnisse. Auswahl und Interpretation von Test- und Fragebogenergebnissen zählen zu den Routineaufgaben in der späteren beruflichen Praxis. Auch in der neuen DIN-Norm 33430 für „Berufsbezogene Eignungsdiagnostik", erschienen im Beuth Verlag 2002, werden Anforderungen an Personen, die in diesem Berufsfeld tätig sind (z.B. Psychologen, Betriebswirte), spezifiziert (S. 10): (1) Grundkenntnisse über Verfahren der Eignungsbeurteilung; (2) **statistisch-methodische Grundlagen** (siehe *Kapitel 5* und *6*); (3) **Testtheorien (Klassische und Item-Response-Theorien)**, **Messtheorien** (siehe *Kapitel 2* und *8*); (4) Evaluationsmethodik einschließlich Kosten-Nutzen-Aspekten; (4) **Konstruktionsgrundlagen** (siehe *Kapitel 3*); (5) Einsatzmöglichkeiten; (6) Durchführungsbedingungen; (7) **Gütekriterien** (siehe *Kapitel 2*); (8) Gutachtenerstellung.

Die durch Fettdruck hervorgehobenen Bereiche der DIN-Norm werden in diesem Buch behandelt.

- Wissen im Bereich „Testkonstruktion" ermöglicht die Konstruktion von Testverfahren mit spezifischer Fragestellung.
- Wissen im Bereich „Testkonstruktion" erleichtert die Auswahl und Interpretation von bereits gebräuchlichen Tests.
- Wissen im Bereich „Testkonstruktion" wird als Schlüsselkompetenz angesehen (Amelang, 1999).

1.3 Arten von Tests

Welche Arten von Tests sollte ich kennen?

Es gibt eine große Vielfalt von Testverfahren. Eine Hilfe bei der Kategorisierung dieser Verfahren liefern Brähler, Holling, Leutner & Petermann (2002). Sie unterteilen Test- und Fragebogenverfahren grob in *drei* große Bereiche (S. XI ff.):

Leistungstests

- Entwicklungstests
- Intelligenztests
- Allgemeine Leistungstests

- Schultests
- Spezielle Funktionsprüfungs- und Eignungstests

Psychometrische Persönlichkeitstests

- Persönlichkeitsstrukturtests
- Einstellungstests
- Interessentests
- Klinische Tests

Persönlichkeitsentfaltungs-Verfahren

- Formdeuteverfahren
- Verbal-thematische Verfahren
- Zeichnerische und
- Gestaltungsverfahren

Leistungstests erfassen, wie der Name schon sagt, eine Leistung. Die Messung einer Leistung setzt voraus, dass die Leistung nach bestimmten Kriterien als richtig oder falsch klassifiziert werden kann. Schmidt-Atzert und Amelang (2006, S. 112) unterteilen Leistungstests in Schnelligkeitstests (Speedtests) und Niveautests (Powertests, siehe Kasten). Diese Unterscheidung ist allgemein üblich und für das Verständnis der Methoden der Testanalyse notwendig (siehe Kasten).

Definition: Speed- und Powertests

Schnelligkeitstests (Speedtests) enthalten leichte oder mittelschwere Aufgaben, die bei der Bearbeitung ohne Zeitbegrenzung von jedem gelöst werden können. Es wird jedoch eine Zeitbegrenzung vorgegeben, so dass kein Proband innerhalb dieser Zeit alle Aufgaben beantworten kann. Das Testergebnis (Score) ergibt sich entweder aus der Anzahl der bearbeiteten Aufgaben oder aus der Anzahl richtig gelöster Aufgaben (Tempowert). Es kann auch sein, dass eine bestimmte Anzahl von Aufgaben vorgegeben wird und die Bearbeitungszeit jeder einzelnen Aufgabe gemessen wird oder alle Aufgaben so schnell und genau wie möglich bearbeitet werden sollen. In diesem Fall wird die Bearbeitungszeit, beispielsweise in Millisekunden, als Testwert erhoben. Es kommt also vorwiegend auf Schnelligkeit (englisch: Speed) an. Das heißt nicht (!), dass den Probanden in diesen Tests keine Fehler unterlaufen. Diese werden meist in einem weiteren Kennwert berücksichtigt oder mit dem Tempowert verrechnet.

Niveautests (Powertests) enthalten Aufgaben, die im Schwierigkeitsgrad kontinuierlich ansteigen. Es wird keine oder eine großzügige Zeitbegrenzung vorgegeben, aber auch bei unbegrenzter Zeitvorgabe können alle Aufgaben (Items) kaum von einem Probanden richtig gelöst werden. Schnelligkeit spielt eine eher untergeordnete Rolle, es geht nur um die Ermittlung des intellektuellen Niveaus bzw. der „Denkkraft“ (englisch: Power).

1

Häufig wird zwischen zeitbegrenzten Niveautests und reinen Niveautests nicht weiter unterschieden. Dabei wird als Argument verwendet, dass Leistungen in zeitbegrenzten und nicht zeitbegrenzten Powertests stark zusammenhängen. Das heißt eigentlich erfassen diese Tests die gleiche Fähigkeit. Somit ist die Zeitbegrenzung bei Powertests zu vernachlässigen. Jedoch handelt es sich, wie Carroll (1993) richtig bemerkt, bei diesem Zusammenhang um eine partielle Eigenkorrelation. Das bedeutet, dass z.B. ein zeitbegrenzter Test schon allein deshalb mit seinem nicht zeitbegrenzten Test hoch korrelieren muss, da er die gleichen Aufgaben verwendet. Wilhelm und Schulze (2002) konnten zeigen, dass zeitbegrenzte Intelligenztests deutlich höher mit zeitbegrenzten Speedtests zusammenhängen als zeitunbegrenzte Intelligenztests. Das heißt, dass diese zeitbegrenzten Intelligenztests noch etwas anderes messen als Intelligenz (ihre Messeigenschaften sind andere), beispielsweise Verarbeitungsgeschwindigkeit. Dies muss in der Praxis kein Nachteil sein, sollte aber bei der Testkonstruktion und Ergebnisinterpretation berücksichtigt werden. Insgesamt spiegelt die vorliegende Einteilung nur sehr grobe Kategorien wider, die aber vorerst zum weiteren Verständnis der folgenden Kapitel ausreichen.

Im Folgenden soll für die gerade dargestellten Kategorien Leistungstests, psychometrische Persönlichkeitstests und Persönlichkeitsentfaltungs-Verfahren je ein Test vorgestellt werden. Der Fokus liegt auf der Beschreibung des theoretischen Hintergrunds. Die Bilder der beteiligten Testkonstrukteure sowie von wichtigen Persönlichkeiten im Rahmen der vorangehenden Theorieentwicklung sind in *Abbildung 1.1* dargestellt. Die Darstellungen sollen auf die folgenden Kapitel einstimmen und bereits in der Einführung einen Überblick über psychologische Tests geben. Die folgenden Verfahren werden kurz dargestellt:

- Leistungstests ▶ Intelligenztests ▶ Matrizentest von J. C. Raven (1938)
- Psychometrische Persönlichkeitstests ▶ Persönlichkeitsstrukturtests ▶ NEO-Persönlichkeitsinventar (NEO-PI) und NEO-Fünf-Faktoren-Inventar (NEO-FFI; Kurzform des NEO-PI) von P. T. Costa und R. R. McCrae (1992)
- Persönlichkeits-Entfaltungstests ▶ Verbal-thematische Verfahren ▶ Thematischer Apperzeptionstest (TAT) von H. A. Murray und C. D. Morgan (1935)

Abbildung 1.1: Porträts der Testkonstrukteure und wichtiger Persönlichkeiten in Bezug auf die den Tests vorangegangene Theorieentwicklung.

1.3.1 Raven Progressive Matrices

Die Raven Progressive Matrices (RPM) stellen nach Angaben der Autoren einen kulturunabhängigen Intelligenztest dar. Die Konstruktion basiert auf der **Zweifaktorentheorie der Intelligenz** von Spearman, die auch als Generalfaktorentheorie bezeichnet wird. Der Theorie lag die Beobachtung von Spearman zugrunde, dass Schulleistungen, die per Augenschein wenig oder gar nichts miteinander zu tun hatten, beispielsweise Mathematik und klassische Sprachen, positiv miteinander korrelierten (vgl. das sehr lesenswerte Buch von Rost, 2009, S. 26 ff.). Spearman (1904) schloss daraus, dass alle Aufgaben, zu deren Bewältigung eine gewisse kognitive Leistungsfähigkeit erforderlich ist, positiv miteinander korrelieren. Das Gemeinsame aller kognitiven Leistungen wurde von Spearman als g-Faktor bezeichnet. Das heißt im Durchschnitt geht eine höhere Leistung bei einer kognitiven Aufgabe auch mit einer höheren Leistung bei einer anderen kognitiven Aufgabe einher. Diese Hypothese wurde auch als Hypothese der positiven Mannigfaltigkeit (Positive Manifold) der intellektuellen Leistungsfähigkeit bezeichnet, auf der die Theorie von Spearman basiert. Damit können Unterschiede in der Erbringung einer kognitiven Leistung in zwei Anteile aufgespalten werden: zum einen Anteile, die auf das Gemeinsame verschiedener kognitiver Leistungen zurückgehen, und zum anderen Anteile, die für jeden Test spezifisch sind.

Die RPM liegen in verschiedenen Versionen vor. Allen Versionen liegt das allgemeine Prinzip des Matrizentests zugrunde: Ein Muster, dem ein Teil zur Komplettierung fehlt, welcher von einer Person aus einer Reihe von Vorlagen richtig ausgewählt werden muss (vgl. Item in Abbildung 1.2). Immer nur eine der Vorlagen passt im Rahmen der jeweiligen Logik in das Muster. Die Schwierigkeit der Items nimmt dabei zu. In der Praxis sicher am häufigsten angewandt werden die „Standard Progressive Matrices“ (SPM). Es handelt sich dabei um die ursprüngliche Fassung des Matrizentests, die erstmals 1938 von J. C. Raven veröffentlicht wurde. Dieser Test umfasst fünf Untertests (A bis E), die jeweils 12 Items enthalten. Insgesamt bestehen die SPM also aus 60 Items. Bei den Untertests A und B muss der Proband die richtige Lösung aus sechs Vorlagen auswählen, bei C bis E sind es sogar acht, die ihm zur Auswahl stehen. Innerhalb eines Untertests bleibt die Logik der Muster stets bestehen, jedoch werden die Items sukzessive schwieriger. Alle Muster sind in schwarzer Farbe auf weißem Hintergrund abgebildet. Der Test ist für Probanden ab 6 Jahren einsetzbar. Die Durchführungsdauer beträgt ca. 45 Minuten.

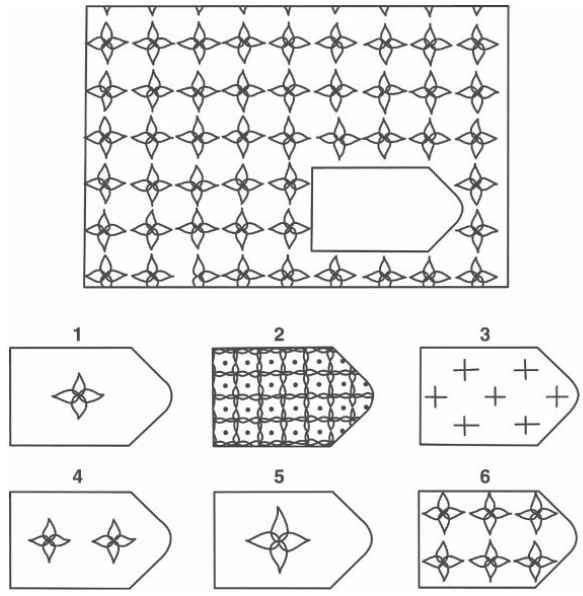

Abbildung 1.2: Item aus den Standard Progressive Matrices von J. C. Raven (deutsche Bearbeitung Bulheller & Häcker, 2009).

Hinsichtlich der qualitativen Aufgabenanalyse heißt es im Testhandbuch sinngemäß, dass sich die Items grob in wahrnehmungsbezogene und logisch-konzeptuelle Aufgaben unterteilen, wobei die wahrnehmungsbezogenen das Erkennen von Gestaltprinzipien und die logisch-konzeptuellen neben Wahrnehmungsleistungen abstraktes sowie induktives Denken erfordern. Induktion bedeutet, dass von Einzelbeobachtungen auf ein allgemeines Prinzip geschlossen wird. Dr. John Raven meint in diesem Zusammenhang:

„Mit den Raven Progressive Matrices verhält es sich wie im richtigen Leben: Man muss zuerst das Problem erkennen. Manchmal ist das Problem nicht von vornherein klar ersichtlich, vielleicht hat man anfangs nur die Ahnung eines Problems. Mit den Items in den Matrizentests ist es ähnlich: Man sieht ein Problem; etwas scheint falsch zu sein; ein Teil fehlt; man experimentiert mit den verschiedenen Lösungen, die zur Auswahl stehen, und versucht, das richtige Item zu finden. Wenn der erste Teil nicht passt, probiert man einfach den nächsten und so weiter. Aber manchmal kann es auch passieren, dass man selbst das Problem als solches gar nicht erkennt …“

1.3.2 NEO-FFI und NEO-PI-R

Ein bedeutender Vertreter der psychometrischen **Persönlichkeitstests** ist das NEO-Fünf-Faktoren-Inventar (NEO-FFI) in der zweiten Auflage von Borkenau und Ostendorf (2008). Das NEO-FFI basiert auf dem NEO-Persönlichkeitsinventar nach Costa und McCrae (1992) in der revidierten Fassung (deutsche Fassung von Ostendorf & Angleitner, 2004). Beide Verfahren stellen **psychometrische Persönlichkeitsstrukturtests** dar und basieren auf dem Fünf-Faktoren-Modell (FFM), das aus dem psycholexikalischen Ansatz hervorgegangen ist. Im Rahmen dieses Ansatzes wird angenommen, dass sich

Persönlichkeitsmerkmale in der Sprache ausdrücken: dies wird auch als Sedimentationshypothese bezeichnet. Ist ein Persönlichkeitsmerkmal besonders wichtig, treten Worte oder Begriffe, die dieses Merkmal beschreiben, in der Sprache besonders häufig auf. Es wird nun davon ausgegangen, dass durch die Analyse der Sprache die wichtigsten Persönlichkeitsfaktoren gefunden werden können. Allport und Odbert (1936) erstellten eine Liste von fast 18.000 persönlichkeitsrelevanten Begriffen (überwiegend Adjektive) aus einem Wörterbuch. Die Begriffe ordneten die Autoren mehreren Kategorien zu. Die Kategorie „Personal Traits" erwies sich dabei als besonders wichtig, da dieser Begriffe zugeordnet wurden, die eine stabile und konsistente Anpassung einer Person an ihre Umgebung beschreiben, dabei jedoch nicht wertend sind. In diese Kategorie fielen etwa 4500 Wörter. Cattell (1943) reduzierte diese Wortliste unter anderem faktorenanalytisch weiter auf 35 Variablencluster (Amelang & Schmidt-Atzert, 2006). Diese Variablencluster dienten dann als Grundlage weiterer Analysen und führten letztendlich zu fünf nahezu unkorrelierten Faktoren der Persönlichkeit: Neurotizismus (emotionale Stabilität), Extraversion, Offenheit, Verträglichkeit sowie Gewissenhaftigkeit. Was diese Skalen messen, soll für die Skalen Extraversion und Neurotizismus kurz beschrieben werden, da ein Datensatz mit diesen beiden Skalen im Buch analysiert wird:

„Personen mit hohen Werten in Neurotizismus neigen dazu, nervös, ängstlich, traurig, unsicher und verlegen zu sein und sich Sorgen um ihre Gesundheit zu machen. Sie neigen zu unrealistischen Ideen und sind weniger in der Lage, ihre Bedürfnisse zu kontrollieren und auf Stresssituationen angemessen zu reagieren."

„Probanden mit hohen Werten in Extraversion sind gesellig, aktiv, gesprächig, personenorientiert, herzlich, optimistisch und heiter. Sie mögen Anregungen und Aufregungen."

Goldberg (1993) nannte die aus Adjektivstudien resultierenden Faktoren Big 5, und Costa und McCrae verwendeten den Begriff Fünf-Faktoren-Modell. Costa und McCrae nutzten zur Erfassung des „Fünf-Faktoren-Modells" kurze Statements.

Die Items des NEO-FFI stellen eine Teilmenge der Items aus dem NEO-PI-R dar. Das NEO-PI-R erfasst jeden Faktor der „Big 5" mit sechs Subskalen (siehe *Abbildung 1.3*), die sich ihrerseits in je acht Items aufgliedern. Insgesamt besteht das NEO-PI-R also aus 240 Items und das NEO-FFI aus 60 Items (je zwölf pro Faktor). Das Antwortformat besteht aus einer fünfstufigen Likertskala mit den Antwortmöglichkeiten *starke Ablehnung*, *Ablehnung*, *neutral*, *Zustimmung* und *starke Zustimmung*.

Neurotizismus (N)	Extraversion (E)	Offenheit für Erfahrungen (O)	Verträglichkeit (A)	Gewissenhaftigkeit (C)
N1: Ängstlichkeit	E1: Herzlichkeit	O1: für Fantasie	A1: Vertrauen	C1: Kompetenz
N2: Reizbarkeit	E2: Geselligkeit	O2: für Ästhetik	A2: Freimütigkeit	C2: Ordnungsliebe
N3: Depression	E3: Durchsetzungsfähigkeit	O3: für Gefühle	A3: Itruismus	C3: Pflichtbewusstsein
N4: soziale Befangenheit	E4: Aktivität	O4: für Handlungen	A4: Entgegenkommen	C4: Leistungsstreben
N5: Impulsivität	E5: Erlebnissuche	O5: für Ideen	A5: Bescheidenheit	C5: Selbstdisziplin
N6: Verletzlichkeit	E6: positive Emotionen	O6: für Werte- und Normensysteme	A6: Gutherzigkeit	C6: Besonnenheit

Abbildung 1.3: Faktoren und Facetten des NEO-PI-R.

Die Auswertung erfolgt durch den Testleiter oder bei der Computerversion durch den PC. Die Antworten der Versuchsperson werden auf einem Durchschreibbogen abgebildet, bei inversen Items automatisch umgepolt und als Zahlenwerte von 0 bis 4 dargestellt. Eine Umpolung ist nötig, da manche Fragen invertiert sind. Eine invertierte

Frage für die Dimension Extraversion wäre beispielsweise: Ich bin kein geselliger Mensch. Eine Person, die dieser Frage stark zustimmt, erhielte ohne Umpolung die höchste Punktzahl von vier Punkten bei diesem Item, ist jedoch wenig extravertiert. Durch die Umpolung erhält die Person, wenn sie starke Zustimmung angibt, null Punkte. Eine nicht invertierte Frage würde lauten: Ich bin ein geselliger Mensch. Stimmt eine Person dieser Frage stark zu, wird es sich wohl um eine extravertierte Person handeln, und sie erhält zu Recht vier Punkte auf die Frage. Die Zuweisung der Zahlen zu den Antwortmöglichkeiten wird auch als Codierung bezeichnet. Der Testleiter addiert nun die Werte der getesteten Person für jede Skala, z.B. Extraversion (Item 2 + Item 7 + Item 12 + … + Item 57), und erhält somit einen Summenwert.

1.3.3 Thematischer Apperzeptionstest (TAT)

Der Thematische Apperzeptionstest (TAT) stellte ein **projektives Testverfahren** dar. Er wurde zur **Messung unbewusster Motive** konzipiert. Erstmals entwickelt wurde dieses Verfahren 1935 von Henry A. Murray und Christina D. Morgan. Seither wurden von den verschiedensten Autoren neuere, abgewandelte Versionen des TAT herausgegeben. Das Prinzip des TAT besteht darin, dass Personen mehrdeutige Abbildungen vorgelegt werden, zu denen sie eine Geschichte erzählen sollen. Die Instruktion lautet dabei: „Erzählen Sie eine Geschichte zu diesem Bild, die möglichst dramatisch ist. Berichten Sie, wie es zu dieser Szene kam, was jetzt vor sich geht und wie die Geschichte ausgeht."

Der TAT von Murray besteht aus insgesamt 30 schwarz-weißen Bildtafeln, die hauptsächlich Menschen in alltäglichen Situationen darstellen, jedoch verschiedenartig interpretiert werden können (vgl. Beispieltafel in *Abbildung 1.4*). Es gibt außerdem eine 31. Bildtafel, welche vollständig weiß ist. Die Tafeln sind auf der Rückseite nummeriert und teilweise für ein bestimmtes Geschlecht bzw. eine bestimmte Altersstufe vorgesehen. Für die erste Untersuchung wird eine Serie von zehn Bildtafeln zusammengestellt, die der Person sukzessive vorgelegt werden. Pro Bild ist eine Erzählzeit von fünf Minuten vorgesehen, so dass die Durchführung des TAT insgesamt circa 50 Minuten in Anspruch nimmt. In einer zweiten Sitzung werden der Versuchsperson wiederum zehn Bildtafeln vorgelegt. Diese zweite Sitzung sollte durch ein mindestens 24-stündiges Intervall von der ersten getrennt sein. Außerdem sollte eine ausführliche biografische Anamnese mit der Person durchgeführt werden, um die erzählten Geschichten mit der jeweiligen Lebensgeschichte in Verbindung bringen und den Ursprung der zum Vorschein gebrachten Motive herausfinden zu können.

Die Auswertung nach Murray besteht in einer so genannten Need-Press-Analyse. Bei den „Needs" handelt es sich um die Bedürfnisse und Wünsche der Hauptfigur, die von den Probanden in ihren Geschichten angesprochen werden. Bei den „Presses" hingegen handelt es sich um die Erwartungen und Zwänge, die von der Außenwelt ausgehen und das Verhalten der Hauptfigur beeinflussen. Zuerst wird entschieden, mit welcher Person auf der Bildtafel sich die Versuchsperson identifiziert. Dann werden Satz für Satz „Needs" und „Presses" analysiert. Im Idealfall kann somit ein Einheitsthema („Unity Theme") ausfindig gemacht werden, welches mit der Biografie des Probanden korrespondieren sollte. Da dieses Auswertungsverfahren jedoch recht aufwendig ist, hat es sich nicht vollständig durchsetzen können; stattdessen haben sich andere Auswertungsverfahren entwickelt, die hier jedoch nicht vorgestellt werden.

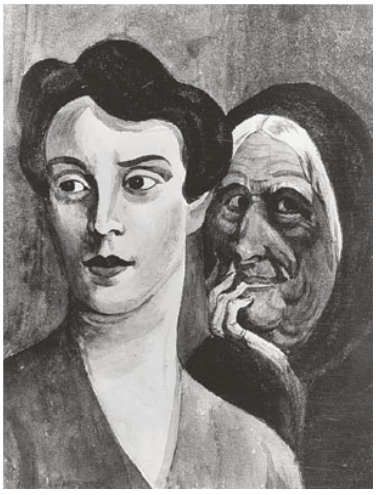

Abbildung 1.4: Bildtafel aus dem TAT.

1.4 Diagnosemöglichkeiten mit Tests

Für welche allgemeinen Fragestellungen kann ich Tests einsetzen?

Man unterscheidet zwei Arten von **Diagnosemethoden**, die Querschnittdiagnose und die Längsschnittdiagnose. Während die Querschnittdiagnose einen aktuell gegebenen Zustand darstellt, zielt die Längsschnittdiagnose auf Veränderungen über einen größeren Zeitraum mit zwei oder mehreren Messzeitpunkten ab. Im Folgenden sollen einige Beispiele für Fragestellungen der Diagnosemethoden erläutert werden.

Querschnittdiagnose

- *Position einer Person innerhalb einer Gruppe* vergleichbarer Individuen hinsichtlich eines oder mehrerer Merkmale (relativer Grad der Ausprägung). Wichtig ist diese Einschätzung vor allem im Rahmen einer Begutachtung. Eine Fragestellung könnte beispielsweise lauten, ob ein Kind hochbegabt ist oder nicht. Diese Fragestellung könnte dann mithilfe eines Intelligenztests, der im oberen Intelligenzbereich Unterscheidungen zwischen Kindern zulässt, beantwortet werden.
- *Unterschiede der Merkmalsausprägung zwischen Personen* oder Gruppen. Eine Fragestellung im Rahmen der Personalselektion könnte z.B. lauten, welcher von mehreren Bewerbern besser für einen Arbeitsplatz geeignet ist. Diese Fragestellung könnte mit objektiven Leistungstests, spezifischen Fragebögen zur Berufseignung oder auch anderen diagnostischen Methoden, wie einem strukturierten Interview, beantwortet werden.
- *Feststellung individueller Merkmalskombinationen (Profil).* Oftmals wird bei Kindern oder Erwachsenen gefragt, ob sie ihr potenzielles Leistungsniveau ausschöpfen. Dazu werden häufig sprachliche und nicht sprachliche Intelligenztests eingesetzt, beispielsweise der HAWIE-III für Erwachsene (Tewes, 1991) oder der HAWIK-III für Kinder (Tewes, Rossmann & Schallberger, 2002). Hier wird zwischen Verbal- und

Handlungsintelligenz unterschieden. Die Leistung in einem Intelligenztest wird in einen Intelligenzquotienten umgerechnet, der als IQ bezeichnet wird. Ist der Handlungs-IQ bedeutsam oder signifikant (über eine bestimmte Zufallswahrscheinlichkeit, in der Regel 5 Prozent, hinaus) höher als der Verbal-IQ, wird dies häufig als Indiz dafür gewertet, dass die geistigen Möglichkeiten des Erwachsenen oder des Kindes nicht vollständig ausgeschöpft werden. Des Weiteren kann man sowohl für den Leistungsbereich (z.B. Intelligenz) als auch den Persönlichkeitsbereich so genannte Profilvergleiche durchführen. Dabei können mehrere Fragestellungen unterschieden werden, z.B.: (1) Unterscheiden sich zwei oder mehrere Merkmalsausprägungen eines Probanden voneinander? (2) Weichen eine oder mehrere Merkmalsausprägungen eines Probanden von einem mittleren Profil einer bestimmten Bezugsgruppe ab?

- *Entscheidung über (Nicht-)Erfüllung einer Bedingung/Erreichen eines Kriteriums.* Ein Patient erhält beispielsweise die Diagnose „Depression“, wenn er genau festgelegte Kriterien für eine Depression erfüllt. Die Diagnose einer Depression wird über Klassifikationssysteme bestimmt, wie z.B. DSM-IV (Saß, Wittchen, Zaudig und I. Houben, 2003) oder ICD-10 (Dilling, Schmidt, Mombour & Schulte-Markwort, 2000). Solche Kriterien sind beispielsweise für eine Depression nach DSM-IV: schwere Beeinträchtigung, die mehr als zwei Wochen andauert, von mindestens fünf Symptomen gekennzeichnet und nicht durch Faktoren wie Medikamente/Drogenmissbrauch und körperliche Erkrankung bedingt ist (Comer, 1995).

Längsschnittdiagnose

- *Merkmalsveränderungen über die Zeit* (Verlaufsprofil) *für Individuen oder Gruppen.* Z.B.: Verändert sich die Konzentrationsfähigkeit einer Person oder mehrerer Personen mit zunehmendem Alter?
- *Prognosen.* Welche psychologischen Tests oder psychologischen Methoden erlauben eine Vorhersage von zukünftigem Verhalten? Dies ist insbesondere im Rahmen der Eignungsdiagnostik relevant. Hier werden psychologische Tests verwendet, um eine Prognose auf den späteren Berufserfolg vorzunehmen. Dabei wird Berufserfolg beispielsweise über Vorgesetztenbeurteilungen, Umsatz oder Gehalt erfasst.

Testtheoretische Grundlagen

2

ÜBERBLICK

In diesem Kapitel wird zunächst die Frage geklärt, was man unter einer Testtheorie versteht, anschließend der Gegenstand einer Testtheorie beschrieben und schließlich die Definition eines psychometrischen Tests gegeben. Danach wird auf die Klassische Testtheorie eingegangen. Die Klassische Testtheorie war die erste Testtheorie, die für psychologische Tests entwickelt wurde. In der Regel vergibt man für die richtige Lösung einer Aufgabe in einem Test einen Punkt und für die falsche Lösung null Punkte. Bei Fragebögen, die eine mehrstufige Ratingskala für die Antworten beinhalten, werden die einzelnen Ratingkategorien ebenfalls mit Zahlen versehen: *starke Ablehnung* = 0, *Ablehnung* = 1, *neutral* = 2, *Zustimmung* = 3, *starke Zustimmung* = 4. Aufgaben oder Fragen eines Tests werden auch als **Items** bezeichnet. Die Punkte der Items werden dann im einfachsten Fall über alle Items zusammengezählt und man erhält einen Testwert. Es wird davon ausgegangen, dass es sich bei diesem Testwert um einen Messwert handelt.

Die Klassische Testtheorie trägt dem Umstand Rechnung, dass die Messung fehlerbehaftet ist und Messungen zu verschiedenen Zeitpunkten unterschiedliche Messwerte für Personen ergeben können. Neben der Klassischen gibt es auch die Probabilistische Testtheorie, die eine Vielzahl von Testmodellen beinhaltet, beispielsweise auch das Rasch-Modell. Betrachten wir beispielsweise das folgende Item: *Ich bin ein geselliger Mensch*. Für dieses Item könnten unterschiedliche Antwortmöglichkeiten eingesetzt werden, beispielsweise *Ja* und *Nein* oder Ratingskalen mit den Antwortstufen *starke Ablehnung*, *Ablehnung*, *neutral*, *Zustimmung* und schließlich *starke Zustimmung*. Für solche Items ist das Rasch-Modell das Modell der Wahl und wird im Anschluss an die Klassische Testtheorie kurz und in *Kapitel 8* ausführlich besprochen.

Am Ende dieses Kapitels wird noch detailliert auf Gütekriterien von Messungen eingegangen. Diese sind von besonderer Bedeutung, da sie für alle Arten psychologischer Messungen gelten. Am Kapitelabschluss befindet sich schließlich eine Kurzcheckliste zur Beurteilung von psychometrischen Tests. Diese Checkliste soll es einem Anwender ermöglichen, einen schnellen Überblick über wesentliche Gütekriterien von Tests zu erhalten.

Einstiegshilfe Günstige Voraussetzungen, um die Inhalte des Kapitels besser nachzuvollziehen, sind Kenntnisse über Zufallsvariablen (vgl. Bühner & Ziegler, 2009, Kapitel 3.1.1 sowie 4.1.3 und 4.1.4) sowie ein Basisverständnis von Skalenniveaus, Varianz, Kovarianz und Korrelation (vgl. Bühner & Ziegler (2009), Kapitel 2.1, Kapitel 2.2.3, Kapitel 3.1 und Kapitel 7.1.1 bis 7.1.4 sowie 7.2).

2.1 Gegenstand einer Testtheorie

Über welche theoretischen Grundlagen sollte ich verfügen?

Begriffsklärung Nach Rost (2004, S. 17 ff.) gibt es in der sozialwissenschaftlichen Methodenlehre für den Begriff Testtheorie zwei unterschiedliche Definitionen. Die erste bezeichnet eine Theorie über statistische Schlüsse: Man schließt aufgrund von Stichprobendaten auf bestimmte Verhältnisse in einer Population (siehe Bühner & Ziegler, 2009, Kapitel 4). Ein solcher statistischer Schluss ist das Ergebnis eines statistischen Tests. Dieser Begriff von Testtheorie wird im Folgenden nicht verwendet.

Arten und Gegenstand psychologischer Tests Die zweite Definition bezieht sich auf psychologische Tests. Bevor diese Definition eingeführt werden kann, sollen zunächst verschiedene Begriffe und Annahmen bezüglich psychologischer Tests dargestellt werden, die für die weiteren Ausführungen notwendig sind. Unter psychologischen Tests im engeren Sinne versteht man **Leistungstests**, **psychometrische Persönlichkeitstests** und **Persönlichkeits-Entfaltungs-Verfahren**. Im weiteren Sinne zählen zu psychologischen Tests auch **standardisierte Interviews** und **standardisierte Beobachtungen**. Psychologische Tests erfassen sowohl **Fähigkeiten**, **Eigenschaften** und **Fertigkeiten** als auch **Zustände** von Personen. Während Fähigkeiten und Eigenschaften zeitlich überdauernde und über Situationen hinweg stabile Merkmale einer Person darstellen, kennzeichnen Zustände momentanes und nicht überdauerndes Befinden einer Person. Unter Fertigkeiten versteht man Aufgabenwissen und Aufgabenkönnen. Aufgabenwissen kann eine Person erwerben, z.B. indem sie sich den Stoff für eine Testtheorieklausur aneignet. Aufgabenkönnen kann eine Person erlangen, wenn sie sich beispielsweise aneignet, wie man einen psychologischen Test sachgerecht konstruiert. Es sei bereits hier gesagt, dass Fähigkeiten, Eigenschaften, Fertigkeiten und Zustände nur dann Gegenstand psychologischer Tests sein können, wenn sie eindimensional sind. Da Eigenschaften, Fähigkeiten, Fertigkeiten und Zustände in den meisten Fällen nicht direkt beobachtbar sind, werden sie aus dem beobachtbaren Verhalten einer Person erschlossen. Daher werden sie auch als **Konstrukte** bezeichnet. Hinter den jeweiligen Konstrukten stehen dabei psychologische Theorien, die empirisch untersucht und gesichert sein müssen.

Psychologische Messungen sind eindimensional Psychologische Tests, die Fähigkeiten, Eigenschaften, Fertigkeiten sowie Zustände erfassen, sollen **eindimensional** sein, da sie nach Westmeyer (2006) als Diagnoseinstrument bzw. Diagnosehilfe zur (1) Klassifikation von Personen eingesetzt werden, an die sich (2) eine konkrete therapeutische Intervention anschließt. (3) Diagnosen dienen nach Westmeyer auch als Erklärungsmodell für bestimmte „auffällige“ Verhaltensweisen. Würde die Testleistung einer Person in einem Test unterdurchschnittlich ausfallen, so ist zu fragen, warum dies der Fall ist. Ist der Test eindimensional, gibt es für das Abschneiden der Person nur eine einzige Erklärung, nämlich die niedrige Ausprägung auf der einen latenten Variablen bzw. dem gemessenen Konstrukt. Damit ist auch eine eindeutige Diagnose verbunden, und es lassen sich so gezielte Interventionsmaßnahmen einleiten. Würden Items mehrere Konstrukte messen, wäre nicht klar, warum eine Person in diesem Test „auffällig“ abgeschnitten hat. Es sind mehrere Erklärungen als Ursache denkbar. Es wäre doch sehr unbefriedigend, einer Person mitzuteilen, dass sie in einem psychologischen Test unterdurchschnittlich abgeschnitten hat, man aber als Psychologe nicht wisse, warum dies so ist.

Gegenstand Testtheorien Theorien sollen nach Rost (2004, S. 29 ff.) vor allem erklären und nicht nur beschreiben. Eine Testtheorie ist deshalb eine erklärende Theorie, da mithilfe eines **Konstrukts** die Zusammenhänge (auch als „Korrelation“ oder „Kontingenz“ bezeichnet) zwischen den **Antworten auf Aufgaben oder Fragen eines Tests kausal** erklärt werden. Konstrukte werden im Rahmen der Testtheorie auch als **latente Variablen** bezeichnet. „Latent“ bedeutet dabei verborgen. Man bezeichnet **Items** auch als **beobachtbare** oder **manifeste Variablen**. Die Items eines Tests werden dabei als Indikatoren der latenten Variablen angesehen. Eine latente Variable könnte z.B. das Konstrukt Intelligenz sein, das über verschiedene einzelne Items gemessen wird. Die

Antworten von Personen auf diese Items stellen dabei das beobachtbare Verhalten der Personen dar. Durch die Wahl einer Antwortkategorie eines Items wird das Antwortverhalten zunächst nur **zählbar** gemacht. Inwieweit mit Items tatsächlich etwas gemessen wird, muss erst überprüft werden. Es besteht also ein Unterschied zwischen Zählen und Messen, denn Messen ist die Zuordnung von Zahlen zu Merkmalen von Personen, so dass die Zahlen die Relationen dieses Merkmals zwischen den Personen abbilden. Näheres hierzu wird in *Kapitel 8* erläutert.

Reflektive Indikatoren Wie bereits eingeführt, besitzen latente Variablen so genannte **Indikatoren**. Indikatoren der latenten Variablen sind dabei **Items** bzw. Itemantworten. Es wird angenommen, dass nur eine **latente Variable** für das Zustandekommen der Unterschiede in den Itemantworten von Personen „verantwortlich" ist und daher deren beobachtbare Zusammenhänge bzw. Korrelationen „produziert". Das heißt Unterschiede in der latenten Variablen erklären Unterschiede im Antwortverhalten *aller* Items, die als Indikatoren der latenten Variablen fungieren. Diese Items spiegeln also die latente Variable wider, daher stammt der Begriff reflektiver Indikatoren. Eine Betrachtung der Ausprägung auf einzelnen Items ist daher nicht mehr nötig. Darüber hinaus gibt es keine weitere(n) Variable(n), die das Antwortverhalten beeinflussen. Wenn diese Erklärung stimmt, müssen sich die Zusammenhänge bzw. Korrelationen zwischen den Items auf null reduzieren, wenn man den Einfluss der latenten Variablen „ausschaltet". Für Ausschalten können auch die Begriffe „konstant halten" oder „auspartialisieren" verwendet werden. In *Abbildung 2.1* wird der beschriebene Gedankengang nochmals veranschaulicht.

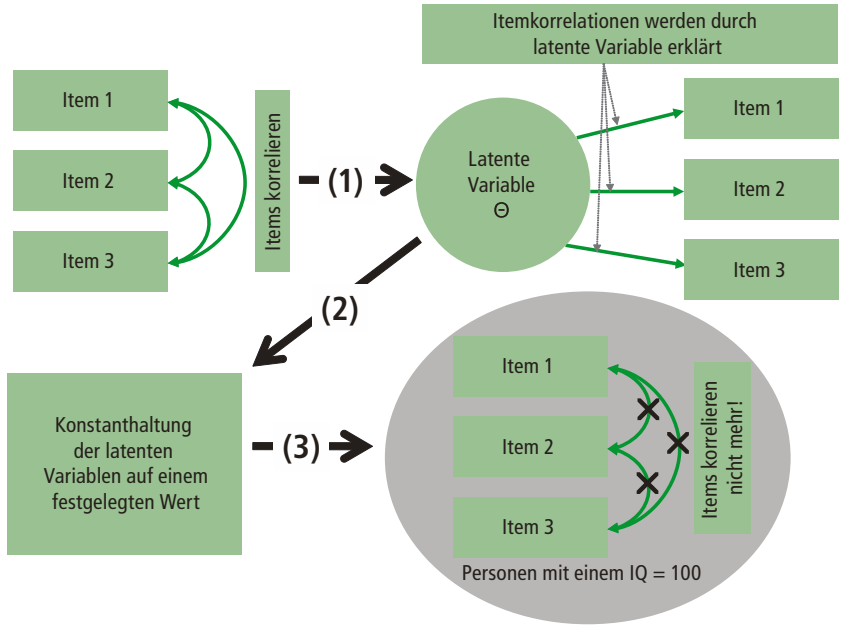

Abbildung 2.1: Lokale Unabhängigkeit reflektiver Indikatoren.

Wie in *Abbildung 2.1* dargestellt, wird davon ausgegangen, dass die Unterschiede der latenten Variablen für die Korrelationen der manifesten Variablen bzw. Items ursächlich verantwortlich sind. Die Wirkrichtung läuft von der latenten Variablen auf die Items:

Die Pfeile zeigen von der latenten Variablen Theta (θ) weg auf die Items. Unterschiede zwischen den Personen in der latenten Variablen sagen folglich Unterschiede in der Itembeantwortung vorher. Da angenommen wird, dass alle Items dieselbe latente Variable messen, müssen die Items miteinander korrelieren. Betrachtet man Personen mit einem festen Wert auf der latenten Variablen, beispielsweise nur Personen mit einem IQ von 100 bei einem Intelligenztest, variiert die latente Variable für diese Personengruppe nicht mehr. Damit können in dieser speziellen Gruppe Unterschiede in der latenten Variablen für Zusammenhänge zwischen den Items nicht mehr verantwortlich sein. Schließlich hat die latente Variable hier keine Varianz mehr, kann also auch nicht mehr für unterschiedliches Antwortverhalten sorgen. Würden dennoch Korrelationen zwischen den Items in dieser speziellen Gruppe bestehen, können diese Korrelationen nicht mehr auf die Variation der latenten Variablen, die konstant gehalten wurde, zurückgeführt werden. Vielmehr ist anzunehmen, dass auch andere Variablen für Unterschiede im Antwortverhalten verantwortlich sind. Man bezeichnet solche Residualkorrelationen auch als Korrelationen zwischen systematischen Messfehlern der betreffenden Items. Residualkorrelation meint, dass die gesamte Korrelation zwischen den Items nicht mehr alleine auf die latente Variable zurückzuführen ist, sondern der durch die latente Variable nicht erklärte Rest an Variabilität korreliert. Ist jedoch *alleine* die Variation der latenten Variablen für die Korrelation zwischen den manifesten Variablen kausal verantwortlich, bestehen in der Gruppe von Personen mit demselben IQ keine Korrelationen mehr zwischen den manifesten Variablen. Man bezeichnet diese Eigenschaft von Items als **lokale Unabhängigkeit** (siehe auch *Kapitel 8*).

Manche Autoren (Fischer, 1974, S. 33) bezeichnen lokale Unabhängigkeit auch als linear experimentelle Unabhängigkeit. In der Klassischen Testtheorie wird angenommen, dass die Messfehler generell unkorreliert sind. Diese Annahme wird zur Definition der Genauigkeit bzw. Reliabilität einer Messung benötigt. Im Rahmen der Probabilistischen Testtheorie wird hingegen tatsächlich geprüft, ob Items lokal stochastisch unabhängig sind, was von manchen Autoren, beispielsweise Fischer (1974), als experimentelle Unabhängigkeit bezeichnet wird. Dies ist eine mathematisch strengere Definition als lokale Unabhängigkeit und wird ebenfalls in *Kapitel 8* ausführlich besprochen.

Exkurs 2.1 **Eindimensionalität**

Eindimensionalität bedeutet eigentlich, dass die Beantwortung der Items von nur einer einzigen Fähigkeit oder Eigenschaft abhängt. Es sind jedoch in der Regel viele Prozesse an der Itembeantwortung beteiligt. So korrelieren Intelligenzmessungen auch mit Gedächtnis und Konzentration. Falls diese Fähigkeiten bei jeder Person im gleichen Ausmaß die Testleistung beeinflussen, sind diese empirisch nicht voneinander zu trennen. Das heißt konkret, wenn das Ergebnis bei der Bearbeitung jedes Items in gleichem Ausmaß von solchen Fähigkeiten abhängt, dann kann die Annahme der Eindimensionalität durch keine psychometrische Analyse verworfen werden (Bejar, 1983). Formal hat Eindimensionalität zur Folge, dass lokale Unabhängigkeit vorliegt. Liegt lokale Unabhängigkeit vor, hat dies wiederum zur Folge, dass Items unkorrelierte Messfehler aufweisen. Die Interpretation in der umgekehrten Richtung ist nicht notwendigerweise zutreffend.

Nehmen wir an, es liegen Daten von Personen vor, die einen zeitbegrenzten Intelligenztest bearbeitet haben. Nehmen wir weiter an, jede Aufgabe dieses Tests wird an einem Computer mit einer festgelegten knappen Zeitvorgabe präsentiert. Hier wirkt sich folglich die Zeitbegrenzung auf alle Personen und alle Items gleichermaßen aus, so dass der Test sowohl schlussfolgerndes Denken als auch Bearbeitungsgeschwindigkeit erfasst. Damit sind schlussfolgerndes Denken und Bearbeitungsgeschwindigkeit nicht voneinander zu trennen. Das heißt selbst wenn Personen mit gleicher Fähigkeitsausprägung im schlussfolgernden Denken untersucht werden, liegen keine korrelierten Messfehler vor. Dennoch liegt weder lokale Unabhängigkeit vor noch Eindimensionalität. Testmodelle überprüfen daher vielmehr eine statistisch definierte Eindimensionalität, aber nicht zwingend eine psychologische. Darüber hinaus ist es vor allem schwierig, möglicherweise vorhandene Mehrdimensionalität nachzuweisen, wenn zwischen zwei oder mehreren Merkmalen hohe Korrelationen bestehen wie etwa zwischen Arbeitsgedächtnis und Intelligenz.

Aus der Unterscheidung zwischen statistischer und psychologischer Eindimensionalität sei daher hier nochmals ein Plädoyer dafür gehalten, eine sorgfältige Itemkonstruktion durchzuführen, die auf einer starken psychologischen Theorie beruht. Daneben ist es wichtig, im Rahmen der Testevaluation weitere Tests mitlaufen zu lassen, die etwas Ähnliches, aber nicht dasselbe wie der zu evaluierende Test messen. Würde man in unserem Beispiel einen Test mitlaufen lassen, der nur die Schnelligkeit der Aufgabenbeantwortung misst, und diesen mithilfe einer Faktorenanalyse gemeinsam mit dem zeitbegrenzten Intelligenztest untersuchen, könnte man wahrscheinlich Mehrdimensionalität feststellen. Auch probabilistische Testmodelle erlauben dies, wenn a priori eine Hypothese darüber besteht, welche anderen Merkmale mit dem zu erfassenden Merkmal zusammenhängen, beispielsweise mehrdimensionale Rasch-Modelle.

Formative Indikatoren Konstrukte können nicht nur aus reflektiven Indikatoren bestehen, sondern auch aus formativen Indikatoren (vgl. Edwards & Bagozzi, 2000). Indikatoren können Items im Sinne von Fragen sein, aber auch Größen wie Länge oder Breite. Indikator wird hier als Begriff für ein Item oder irgendeine andere Größe verwendet. Wie der Begriff es nahelegt, „formen“ diese Items das Konstrukt: Hier verursachen die Indikatoren die Ausprägung auf der latenten Variablen. Veränderungen in der Ausprägung eines einzelnen Indikators gehen mit einer Veränderung der Ausprägung der latenten Variablen einher. Eine Veränderung der Ausprägung des Konstrukts geht hingegen nicht zwangsweise mit einer Veränderung der Ausprägung aller Indikatoren einher. Ob und inwieweit sich die Ausprägung der anderen Indikatoren verändert, wenn sich ein Indikator in der Ausprägung verändert, ist nur durch die Höhe der Korrelationen zwischen den beobachteten Indikatoren bestimmt. Die Wahl der Indikatoren hängt alleine von der Definition und den Bestandteilen des Konstrukts ab, die eben nicht alle dasselbe messen müssen und nicht untereinander korreliert sein müssen (siehe Qualitätssicherung Pflege: Beispiel). Man bezeichnet solche Konstrukte als **Indizes**. **Index** bedeutet übersetzt unter anderem „Anzeiger“. Ein Index ist also demnach eine Größe, die den Inhalt einer definierten Itemmenge anzeigt. Die Güte eines Index hängt davon ab, ob er vollständig im Sinne des Konstrukts ist und kein wesentlicher

Bestandteil vergessen wurde. Im Gegensatz zu Indikatoren eines reflektiven Konstrukts müssen formative Indikatoren von Konstrukten nicht eindimensional sein. Betrachten wir zur Veranschaulichung die folgenden beiden Beispiele.

Beispiel 2.1 **Beispiele formativer Indikatoren**

Betrachten wir als Beispiel Items einer Testtheorieklausur. Im Rahmen einer solchen Klausur müssen vom Prüfling Fragen zu verschiedenen Wissensteilbereichen beantwortet werden. Nun ist nicht anzunehmen, dass die Itemzusammenhänge der Klausur durch eine einzige Variable, wie z.B. Testtheoriewissen, erklärt werden können. Wahrscheinlich erklären unterschiedliche Variablen oder Indikatoren, wie beispielsweise mathematisches Verständnis, Gedächtniskapazität oder die verwendete Lernzeit, die Itemzusammenhänge. Welche Konstrukte bzw. Indikatoren an der erfolgreichen Bewältigung der Klausur in welcher Gewichtung beteiligt sind, ist zunächst auch nicht von Interesse: Wichtig ist, dass die Klausurfragen den Stoff abdecken. Daher kann eine vorher festgelegte Frage auch nicht im Nachhinein aus dem Test entfernt werden, da sie Teil des Wissens und der vorher definierten Itemmenge ist. Eine Frage könnte lediglich unklar gestellt sein und daher der sprachlichen Überarbeitung bedürfen. Im Rahmen einer Testtheorieklausur können aus Zeitgründen sicherlich auch nicht alle denkbaren Fragen zur Testtheorie gestellt werden. In einem solchen Fall ist es jedoch wichtig, die inhaltlich wichtigen Fragen auszuwählen. Diese müssen nicht repräsentativ für alle denkbaren Items sein. Beispielsweise könnte die Vorlesung auf die statistischen Aspekte der Testtheorie fokussieren und nicht auf die Itemkonstruktion an sich. Daher werden die meisten Fragen statistischer Art sein. Die Note, die sich nun als Ergebnis aus den ganzen Aufgaben ergibt, repräsentiert also nicht eine latente Variable, sondern wurde aus den einzelnen Indikatoren geformt.

Fragebogen zur Qualitätssicherung. Auch die Entwicklung eines Fragebogens zur Qualitätssicherung in der Pflege älterer Menschen stellt ein Konstrukt mit formativen Indikatoren oder einen Index dar. Auch hier sind vermutlich unterschiedliche, möglicherweise von Experten festgelegte Erfordernisse zu berücksichtigen, wie beispielsweise emotionale Zuwendung, Körperpflege und pünktliches Verabreichen von Medikamenten. Daher können die Zusammenhänge zwischen allen Items nur schwer durch eine einzige latente Variable erklärt werden. Umgekehrt jedoch können die Items das Konstrukt „Güte der Pflege“ sehr gut beschreiben. Das heißt bei formativen Indikatoren erklären die Items das Konstrukt. Im Gegensatz zu z.B. der Testtheorieklausur muss in diesem Fall wirklich „Qualität der Pflege“ erschöpfend definiert sein und auch abgefragt werden. Es reicht an dieser Stelle nicht aus, eine repräsentative Itemmenge bei jeder Prüfung einer Pflegeeinrichtung abzufragen. Es wäre sicherlich nicht hinzunehmen, dass ein relevanter Bereich bei der Einschätzung der Qualität einer solchen Einrichtung fehlt. Die Vollständigkeit, mit der der Index „Pflege“ erfasst, macht seine Güte aus.

Formative Indikatoren müssen folglich nicht miteinander korrelieren. Formative Konstrukte müssen daher nicht zwingend eindimensional sein. In vielen Fällen ist vor allem die Korrelation des Index mit einem Kriterium wichtig. Wie in *Abbildung 2.2* zu erkennen ist (siehe Pfeilspitze), sagen formative Indikatoren das Konstrukt vorher bzw. erklären es: Die Wirkrichtung läuft von den Indikatoren auf die latente Variable. Formative Indikatoren können dabei untereinander völlig unkorreliert sein. Allerdings kann man bei einer Veränderung des Konstrukts nicht darauf schließen, dass sich *alle* Indikatoren verändern, wie dies bei reflektiven Indikatoren der Fall ist. Im Extremfall könnte sogar nur ein Indikator dafür verantwortlich sein, dass sich die Ausprägung des Konstrukts verändert. Formative Modelle sind schwer zu konstruieren: Das Weglassen eines Bestandteils des Index führt zu einer deutlichen inhaltlichen Veränderung des Konstrukts. Da reflektive Indikatoren aus einem gedachten Itemuniversum lediglich beispielhaft für das Konstrukt sind, sind sie in einem gewissen Rahmen (beispielsweise anhand von Trennschärfe und Schwierigkeit) austauschbar.

Für formative Indikatoren ist die Klassische Testtheorie nicht anwendbar. Auch Schätzungen der Genauigkeit einer Messung, wie beispielsweise Cronbach-α (siehe *Abschnitt 4.4*), sind für formative Konstrukte nicht angemessen.

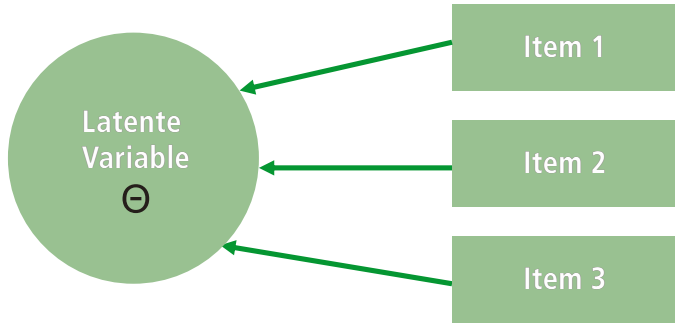

Abbildung 2.2: Formative Indikatoren.

Zusammenfassend lassen sich formative und reflektive Indikatoren eines Konstrukts inklusive der konzeptionellen Ebene, wie in *Abbildung 2.3* abgebildet, darstellen. Es ist zu beachten, dass der grün hinterlegte Bereich in der Abbildung den Geltungsbereich von Testtheorien kennzeichnet.

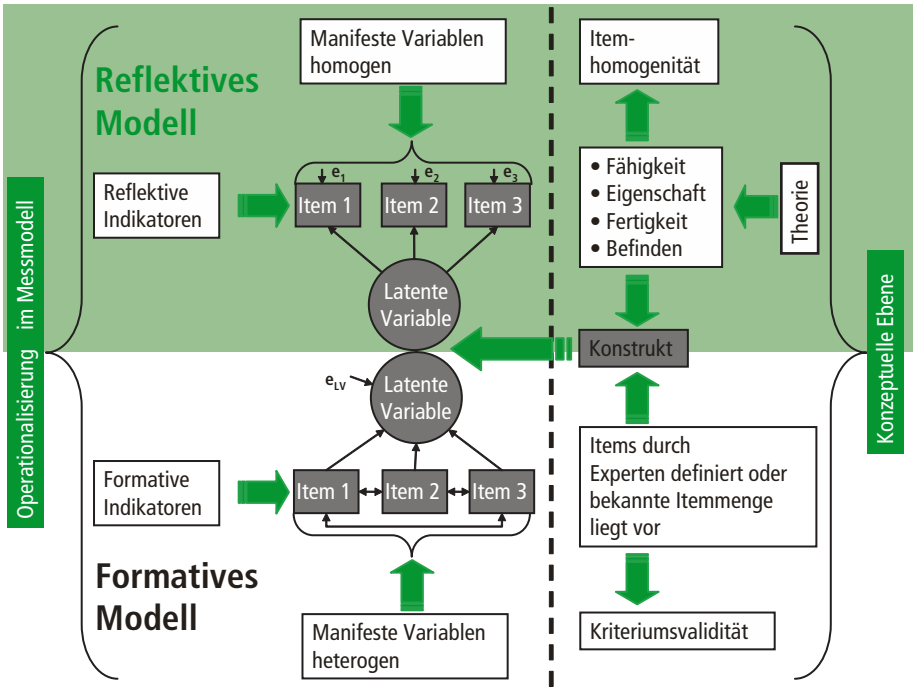

Abbildung 2.3: Reflektive und formative Indikatoren eines Konstrukts.

Psychologische Tests Die am Beginn des Abschnitts angesprochene zweite Bedeutung des Begriffs Testtheorie bezeichnet eine **Theorie über „psychologische Tests“**. Testtheorien in diesem Sinne sind die Klassische und die Probabilistische Testtheorie. Beide Testtheorien werden meistens so angewandt, dass jeweils nur eine psychologisch definierte Fähigkeit, Fertigkeit, Eigenschaft oder ein Zustand durch einen Test gemessen wird, und besitzen daher reflektive Indikatoren. Testtheorien dienen generell dazu, die Qualität einer Messung zu definieren. Qualität ist in diesem Zusammenhang weit gefasst: (1) Qualität kann sich auf die Genauigkeit einer Messung beziehen wie in der Klassischen Testtheorie, (2) jedoch auch auf deren Gültigkeit selbst. Gültigkeit meint hier, ob überhaupt etwas gemessen wird. Dies wird auch als das **„Gütekriterium“ der Skalierbarkeit** bezeichnet und unter dem *Kapitel 2.4* Gütekriterien genauer beschrieben.

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

Merkmale psychometrischer Tests Für die Testkonstruktion sollten eine Theorie über das Persönlichkeitsmerkmal sowie genaue Überlegungen zur Erfassung des angestrebten Persönlichkeitsmerkmals vorliegen. Das interessierende Persönlichkeitsmerkmal sollte genau definiert sein. Der Prozess muss dabei sorgfältig im Handbuch des Tests dokumentiert werden. Anhand der Definition wird dann nach Indikatoren für das Persönlichkeitsmerkmal gesucht. In der Regel ist das Itemuniversum unendlich groß, und es muss sichergestellt werden, dass die ausgewählten Items für das zu messende Persönlichkeitsmerkmal repräsentativ sind. Das bedeutet, das Persönlichkeitsmerkmal muss in all seinen Facetten abgebildet werden. Wie dies möglich ist, wird in *Kapitel 3* ausführlich besprochen. Die Items werden dann je nach angewandter Testtheorie verschiedenen psychometrischen Analysen unterzogen. Darüber hinaus werden bestimmte Gütekriterien für die Messung, die mit einem Test vorgenommen wird, empirisch und über inhaltliche Überlegungen ermittelt. Schließlich ist es notwendig zu zeigen, dass das gemessene Konstrukt auch in der Lage ist, tatsächliches Verhalten, das mit dem Konstrukt im Zusammenhang steht, zu erfassen. Beispielsweise sollte ein Intelligenztest in der Lage sein, Schulleistungen vorherzusagen. Erst am Ende all dieser Schritte kann beurteilt werden, ob ein Test in der Lage ist, tatsächlich den Ausprägungsgrad eines Persönlichkeitsmerkmals zu messen bzw. die Zuordnung einer Person zu einer Gruppe erlaubt.

Definition: Psychometrischer Test
Ein psychometrischer Test ist ein wissenschaftliches **Routineverfahren** zur Untersuchung eines oder mehrerer empirisch abgrenzbarer **Persönlichkeitsmerkmale** (vgl. Lienert & Raatz, 1998, S. 1). Das Ziel eines psychometrischen Tests besteht darin, die **absolute** oder **relative Ausprägung** einer **Eigenschaft**, einer **Fähigkeit** oder eines **Zustands** bei einer oder mehreren Personen zu messen oder aber eine **qualitative Aussage** zu treffen, welcher **Personenklasse** Personen zugeordnet werden können (vgl. Rost, 2004). Psychometrische Tests sind nach der Klassischen oder Probabilistischen Testtheorie entwickelt, sind **theoretisch fundiert** und genügen genau definierten **Gütekriterien (Haupt- und Nebengütekriterien).**

Im Folgenden werden nun vor allem die Gütekriterien und die Klassische Testtheorie näher erläutert. Es erfolgt der Vollständigkeit halber auch ein kurzer Verweis auf die Probabilistische Testtheorie. Da insbesondere das Rasch-Modell als wichtigstes dieser Testmodelle von herausragender Bedeutung für die Testkonstruktion ist, werden dieses Modell und seine Varianten in *Kapitel 8* ausführlicher besprochen.

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

An dieser Stelle soll eine kurze Zusammenfassung gegeben werden, bevor eine Einführung in die Klassische Testtheorie folgt:

Unter psychologischen Tests versteht man Leistungstests, Fragebögen, standardisierte Interviews und standardisierte Beobachtungen. Psychologische Tests erfassen Fähigkeiten, Eigenschaften, Fertigkeiten und auch Zustände von Personen, die wiederum als Konstrukte bezeichnet werden. Handelt es sich um ein zu messendes Konstrukt, spricht man auch von einer latenten Variable. Indikatoren für eine latente Variable sind Items. Itemantworten stellen dabei das beobachtete Verhalten der Testpersonen dar. Durch die Beantwortung der Items wird das Verhalten von Personen zunächst nur zählbar gemacht.

Konstrukte können formative und reflektive Indikatoren besitzen. Die latente Variable erklärt die Zusammenhänge reflektiver Indikatoren. Ein sehr wichtiges Ziel der Testkonstruktion ist es, die Eindimensionalitätsannahme nach sorgfältiger Testkonstruktion nicht verwerfen zu müssen. Nur eindimensionale Messungen sind Gegenstand der Klassischen Testtheorie und auch von einigen Probabilistischen Testmodellen. Das heißt konkret, mit einer Änderung der Ausprägung auf dem Konstrukt geht auch eine Änderung der Ausprägung auf allen Items einher. Bei formativen Indikatoren geht hingegen eine Änderung des Konstrukts nicht zwangsweise mit einer Änderung auf allen Indikatoren einher. Im Extremfall ist dafür nur ein Indikator verantwortlich. Formative Indikatoren erklären das Konstrukt. Entscheidend für die Auswahl formativer Indikatoren ist daher nicht die Itemhomogenität wie bei reflektiven Indikatoren, sondern die Inhaltsvalidität.

Psychometrische Tests genügen definierten Gütekriterien, sind nach der Klassischen oder der Probabilistischen Testtheorie konstruiert und weisen einen theoretischen Rahmen auf. Ihr Ziel ist die quantitative Erfassung einer Merkmalsausprägung oder die Zuordnung einer Person zu einer Klasse von Personen.

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

2.2 Klassische Testtheorie

Was muss ich über die Klassische Testtheorie wissen?

Die Klassische Testtheorie ist gegenwärtig die Grundlage der meisten psychologischen Testverfahren. Nach Rost (1999, S. 140) basieren 95 Prozent aller Tests auf der Klassischen Testtheorie. Klassisch heißt sie deshalb, weil sie die erste Theorie war, die zur Konstruktion von psychologischen Tests herangezogen wurde. Es gibt unterschiedliche Arten, die Klassische Testtheorie darzustellen. Die Darstellung in diesem Kapitel orientiert sich an Lord und Novick (1968) sowie Steyer und Eid (2001). Ein großer Vorteil der Klassischen Testtheorie liegt in ihrer einfachen Anwendbarkeit (Henard, 2000). Bevor nun die Grundannahmen der Klassischen Testtheorie erläutert werden, sollen anhand eines Beispiels aus dem Sport einige Überlegungen angestellt werden, die zu einem besseren Verständnis für die Nützlichkeit einer Testtheorie führen sollen.

Variation beboachteter Werte Nehmen wir als Beispiel einen Hochsprungwettbewerb. Das Ziel eines solchen Wettbewerbs ist es, die Leistungsfähigkeit eines Hochspringers zu messen. Hochspringer haben für jede Höhe, die ihnen vorgegeben wird, drei Sprungversuche. Mit zunehmender Höhe wird es schwieriger, die Latte zu überspringen. Das heißt den Hochspringern werden unterschiedlich schwere Höhen (Items) vorgegeben, um ihre Leistungsfähigkeit zu messen. Es gewinnt der Hochspringer, der

das schwierigste Item löst und damit am höchsten springt. Allerdings wird ein Hochspringer nicht in jedem Wettkampf die gleiche Höhe erzielen. Bei einem Wettkampf wird er vielleicht 2.10 Meter überspringen und bei einem anderen 2.15 Meter. **Seine erzielten Leistungen werden also in einem bestimmten Bereich schwanken.**

Zufällige Messfehler Wenn wir annehmen, dass die Wettkampfbedingungen von Messung zu Messung konstant sind, sollten die Wettkampfleistungen eines Hochspringers einer bestimmten individuellen Verteilung folgen. Das heißt Leistungen, die am ehesten seiner Fähigkeit entsprechen („wahre" Leistungsfähigkeit), werden bei wiederholten Sprüngen häufiger vorkommen, extrem schlechte oder gute Leistungen seltener. Unter der Konstanz von Wettkampfbedingungen versteht man, dass sich die Wettkampfbedingungen während der Wettkämpfe nicht durch Regen, stärkeren Gegenwind oder Ähnliches verändern. Diese Konstanz mag innerhalb eines Wettkampfs manchmal noch realisierbar sein, in zeitlich aufeinanderfolgenden Wettkämpfen ist jedoch die Bedingungskonstanz selten gegeben. Die erzielten Höhen werden sich aber selbst unter perfekter Konstanz der Wettkampfbedingungen unterscheiden, da auch Faktoren, die in der Person liegen, variieren können, wie z.B. Müdigkeit oder Motivation. Folglich entstehen Messfehler, zum einen aufgrund der äußeren Bedingungen und zum anderen aufgrund von Faktoren, die innerhalb der getesteten Person liegen.

Prinzip wiederholter Messungen Das heißt selbst unter optimalen Bedingungen ist die Wahrscheinlichkeit sehr gering, dass Personen immer die gleiche Leistung erzielen. Es ist also sinnvoll, mehrere Messungen vorzunehmen, um die Leistung eines Hochspringers zu ermitteln. In der Regel hat jeder Hochspringer für jede Höhe drei Versuche. Die Leistung wird aber nicht nur über einen Wettkampf zusammengefasst, sondern auch über verschiedene Wettkämpfe. Kann man die unterschiedlichen Leistungen zwischen den Wettkämpfen durch zufällige Fehler erklären, würden sich diese über die wiederholten Messungen hinweg immer mehr ausmitteln. Wiederholte Messungen führen also zu einer genaueren Einschätzung der Wettkampfleistung oder der Hochsprungfähigkeit einer Person.

Einzelmessungen als Fehlerquelle Angenommen, man würde die Hochsprungleistung nicht durch verschiedene Durchgänge mit unterschiedlichen Höhen messen, sondern nach einem einzigen Sprung ermitteln, indem man die gesprungene Höhe mit einer Kamera festhält. Zieht man nur diesen Einzelwert zur Beurteilung der Leistungsfähigkeit eines Hochspringers heran, so könnte dies zu Fehlschlüssen führen. In einem solchen Fall ist es möglich, dass ein Hochspringer mit einer „tatsächlich" niedrigeren Leistungsfähigkeit einen Hochspringer mit einer „tatsächlich" höheren Leistungsfähigkeit in einem Wettkampf besiegt (siehe *Abbildung 2.4*). Dieses Ergebnis kann man damit erklären, dass dem Hochspringer mit geringerer Leistungsfähigkeit ein extrem guter Sprung (löst Item mit hoher Schwierigkeit) gelungen ist und dem Hochspringer mit der höheren Leistungsfähigkeit nur ein extrem schlechter Sprung (scheitert an Item mit geringer Schwierigkeit). Die Wahrscheinlichkeit für ein solches Wettkampfergebnis ist gering, da man annehmen muss, dass ein Hochspringer mit einer hohen Fähigkeit auch mit einer höheren Wahrscheinlichkeit eine bessere Höhe erzielt als ein Hochspringer mit einer niedrigen Fähigkeit. Dieses Beispiel zeigt, dass es sinnvoll ist, mehrere Sprünge (Items) mit unterschiedlichen Höhen (Schwierigkeiten) heranzuziehen, um die „wahre" Leistungsfähigkeit eines Springers zu beurteilen.

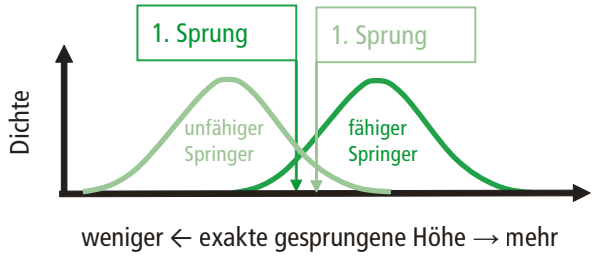

Abbildung 2.4: Darstellung Einzelmessungen als Fehlerquelle.

Schlussfolgerungen Folgende Schlüsse lassen sich aus dem beschriebenen Beispiel ziehen:

- Auch wenn die Bedingungen für jeden Sprung (konstante Wettkampfbedingungen) gleich sind, werden die Leistungen eines Hochspringers variieren. Grund dafür sind nicht kontrollierbare Einflüsse der Situation und der Person. Daher kann die Kenntnis eines einzelnen Wertes zu falschen Schlussfolgerungen führen.
- Es ist notwendig, die Leistung einer Person über mehrere Messgelegenheiten zu erheben. Diese Messgelegenheiten können mehrere Sprungversuche (Items) oder aber auch mehrere Wettkämpfe (Testwiederholungen) sein. Dieses Aggregationsprinzip erhöht die Genauigkeit der Messung.
- Ob sich Personen in ihrer Hochsprungfähigkeit unterscheiden, erkennt man, indem man ihnen Items (Sprungversuche) unterschiedlicher Schwierigkeit (unterschiedliche Höhen) vorgibt: beispielsweise 2.10 Meter und 2.15 Meter.

2.2.1 Grundannahmen der Klassischen Testtheorie

Was sind die Grundannahmen der Klassischen Testtheorie (KTT)?

Im Folgenden werden die Grundannahmen der Klassischen Testtheorie einfach und ohne formale Herleitung oder Einbettung geschildert. Sehr umfassend ist die Klassische Testtheorie bei Steyer und Eid (2001) dargestellt. Der interessierte Leser sei zur Vertiefung der Klassischen Testtheorie ausdrücklich auf dieses Lehrbuch verwiesen. Bei der hier gewählten Darstellung wird dabei eine Aufteilung in (1) Kern bzw. Definitionen, (2) Folgerungen und (3) Zusatzannahmen vorgenommen. Die Klassische Testtheorie stellt das Gerüst dar, das die Basis für die Definition der Messgenauigkeit einer Messung und all ihrer Anwendungen liefert. Wie die Messgenauigkeit dann genau definiert ist, wird in *Kapitel 4* dargestellt. Es sei bereits hier darauf hingewiesen, dass die Klassische Testtheorie davon ausgeht, dass bereits Messwerte vorliegen. Das heißt, dass die Testitems eines Tests tatsächlich eine latente Variable messen und beispielsweise die Summation der Items eine gültige Verrechnungsvorschrift der Items darstellt. Diese Annahmen werden im Rahmen der Klassischen Testtheorie nicht überprüft.

Definitionen Die Klassische Testtheorie trägt dem Umstand Rechnung, dass Messwerte einzelner Personen über verschiedene Messungen hinweg variieren. Steyer und Eid (2001, S. 102) nennen dafür zwei unterschiedliche Gründe. Übertragen wir diese auf unser obiges Beispiel, kann zum einen die Ausprägung einer Person „hochzuspringen" durch ein besonderes Aufbautraining (**Übungs- und Transfereffekte**) verbessert werden. Zum anderen ist es möglich, dass die Messungen durch **unsystematische äußere Einflüsse,** wie Wind und Regen, oder **unsystematische innere Einflüsse,** wie Müdigkeit oder mangelnde Motivation, zufällig schwanken. Diese Einflussfaktoren treten meist in Kombination auf. Abgesehen vom systematischen Trainingseffekt führen diese Faktoren zu einer verzerrten Messung und werden daher auch als Messfehler bezeichnet.

Messfehler als „zufällig" definiert Der Fehlerbegriff in der Klassischen Testtheorie berücksichtigt jedoch nur **unsystematische Messfehler**, die häufig normalverteilt sind. Darüber hinaus erfolgen in der Klassischen Testtheorie **keine Annahmen** darüber, wie Items beantwortet werden oder wie Testleistungen zustande kommen (Fischer, 1974, S. 124), sondern nur, aus welchen **Komponenten** Messwerte bestehen. Die Klassische Testtheorie ist eine reine **Messfehlertheorie**.

Ziehungsprozess als Basis Novick (1966) geht davon aus, dass die getestete Person zufällig aus einer Population entnommen wurde und das Testergebnis einer Person über wiederholte Messungen aufgrund von Messfehlern zufallsabhängig variiert. Im Rahmen wiederholter Messungen erzielt eine Person folglich unterschiedliche Messwerte. Diese Messwerte entstammen der intraindividuellen Werteverteilung der getesteten Person. Diese intraindividuelle Verteilung stellt eine hypothetische Verteilung dar, die sich ergeben würde, wenn man diese Person unendlich oft unter identischen Bedingungen testen würde. Man nimmt weiter an, dass die beobachteten Messwerte einer festen Person eine Varianz besitzen und diese endlich ist (Krauth, 1995, S. 238). Aus der Verteilung dieser Messwerte einer Person wird bei einer Testung zufällig ein Wert beobachtet. Mit welcher Wahrscheinlichkeit die Werte einer Person beobachtet werden, wird durch die intraindividuelle Verteilung der beobachteten Werte einer Person festgelegt. Dieser Prozess ist in *Abbildung 2.5* noch einmal veranschaulicht. Nimmt man für die beobachteten Werte einer Person eine Normalverteilung an, haben die Werte um den Erwartungswert der Verteilung die höchste Wahrscheinlichkeit, beobachtet oder gezogen zu werden, und extreme Werte eine geringere Wahrscheinlichkeit. Der Erwartungswert stellt dabei den Mittelwert einer Person über unendlich viele Messungen dar.

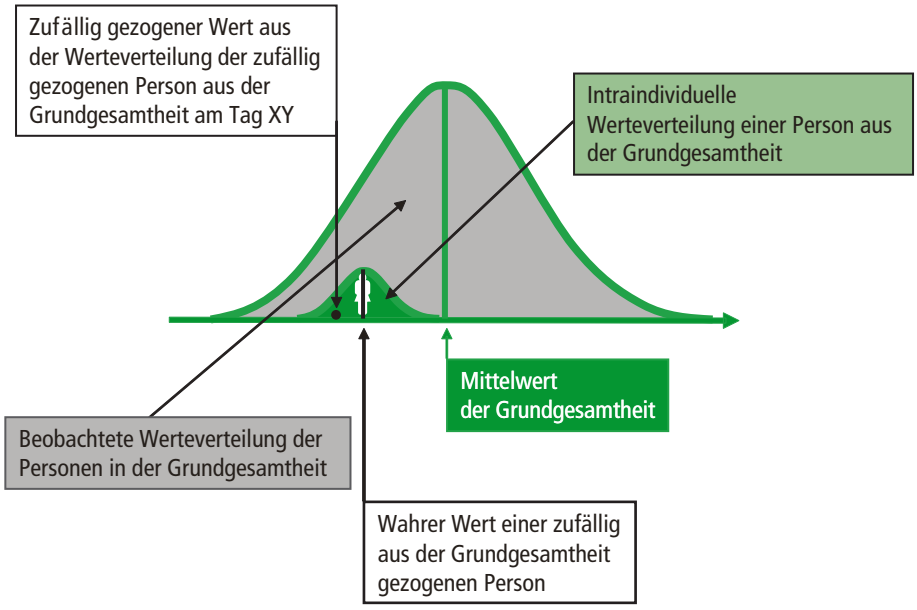

Abbildung 2.5: Ziehungsprozess im Rahmen der Klassischen Testtheorie.

Beispiel 2.2 **Sportwettkampf**

Die beiden folgenden Definitionen legen fest, dass es einen wahren Wert gibt, wie er ermittelt werden kann und wie die Ausprägung des Messfehlers im Rahmen einer Messung bestimmt wird. Diese Definitionen werden auch manchmal als **Axiome** bezeichnet. Axiome sind Aussagen, die nicht begründet werden müssen, sondern von sich aus offensichtlich richtig sind.

Konstruktion des wahren Werts Der **wahre Wert (Tau, τ)** einer **Person (v)** ist als **Mittelwert** über unendlich oft **wiederholte unabhängige Messungen (t)** der **beobachteten Werte (x)** einer **Person (v)** definiert. Es handelt sich folglich um den **Erwartungswert E (X)** der intraindividuellen Verteilung der beobachteten Werte x einer Person. Da es sich um unabhängige Messungen handeln soll, ist an dieser Stelle schon eine weitere Voraussetzung bzw. Definition nötig, nämlich die **(lokale stochastische) Unabhängigkeit** der Messungen für jede Person. Auf diese wird im Zusammenhang mit der

Zusatzannahme und in der Zusammenfassung noch einmal genauer eingegangen. Weiterhin wird an dieser Stelle angenommen, dass der beobachtete Wert auch wirklich etwas misst, man bezeichnet dies als **„per-fiat"-Messung**. Per fiat bedeutet aus dem Lateinischen übersetzt „so sei es". Zunächst einmal wird jedoch ein wahrer Wert für jede Person definiert bzw. konstruiert. Diese Definition wird auch als Existenzaxiom bezeichnet (Moosbrugger & Hartig, 2003).

$$\tau_v = \mathbf{E}(\mathbf{X}_v) \qquad \rightarrow \textbf{Definition 1}$$

Exkurs 2.2 Zufallsvariablen

Zufallsexperimente sind Experimente mit zufälligem Ausgang. Eine Zufallsvariable ist eine Funktion, die den Ergebnissen eines Zufallsexperiments eine Zahl zuweist. Das heißt nicht, dass der Wert einer getesteten Person durch Zufall zustande gekommen ist. Es heißt lediglich, dass man bei Ziehung einer Person aus einer Population den Wert der Person zunächst nicht kennt. Bei diesen und den folgenden Formeln werden nun große und kleine Buchstaben verwendet. Ein großer Buchstabe steht für eine Zufallsvariable. Ein kleiner Buchstabe steht für den konkreten Wert, den die Zufallsvariable annehmen kann.

Beispielsweise ist der beobachtete Wert x von Markus zum Zeitpunkt t in einem Depressionsfragebogen $x_{vt} = 18$. Die Zufallsvariable X ist dann der Depressionswert X selbst, der folglich je nach zufällig gezogener Person v einen unterschiedlichen Wert annehmen kann und für Matthias beispielsweise drei Punkte betragen könnte. Der beobachtete Wert stellt aber auch für t wiederholte Messungen an einer einzelnen Person v eine Zufallsvariable X dar, denn der beobachtete Wert einer festen Person, beispielsweise Markus, kann von Messung zu Messung variieren: beispielsweise wird für Markus heute ein Depressionswert von $x_{vt} = 18$ Punkten und bei der Messung morgen ein Wert von $x_{vt} = 15$ Punkten ermittelt.

Eine weitere Zufallsvariable ist der Messfehler Epsilon, da dieser sowohl für unterschiedliche Personen als auch innerhalb einer festen Person bei wiederholter Messung t variiert: Beispielsweise beträgt der Messfehler bei der heutigen Messung für Markus $\varepsilon_{vt} = 5$ und für Matthias $\varepsilon_{vt} = 6$ Punkte. Bei einer Messung am nächsten Tag kann der Messfehler für Markus hingegen einen Wert von $\varepsilon_{vt} = 2$ Punkte annehmen.

Der wahre Wert einer Person τ_v stellt, betrachtet man **eine einzelne Person**, keine Zufallsvariable dar, da es sich für jede Person um einen konstanten Wert handelt: Er variiert für eine Person in wiederholten Messungen nicht. Daher fehlt auch der Index t für den Zeitpunkt. Betrachtet man verschiedene Personen zu einem Zeitpunkt t, stellt der wahre Wert hingegen wieder eine Zufallsvariable dar. Bleiben wir beim Beispiel Depression. Depression ist hier wieder die Zufallsvariable und verschiedene Personen können unterschiedliche wahre Werte in einem Depressionsfragebogen erzielen, beispielsweise Markus 18 und Matthias 4 Punkte.

Beispiel 2.3 **Definition 1**

Für unser Sportbeispiel heißt dies zunächst, dass die wahre Hochsprungfähigkeit einer Person die mittlere Leistung über unendlich viele Hochsprungwettbewerbe darstellt.

Diese Definition hat in der Anwendung ein Problem: Die meisten psychologischen Tests verwenden Summenwerte als Indikator für die Testleistung. Es handelt sich dabei um Summen der Itemantworten über die Items eines Tests. Für solche so genannten Summenwerte wird der Mittelwert über verschiedene wiederholte Messungen demnach kein ganzzahliger Wert sein. Dasselbe gilt für Normwerte. Die Ausprägung einer Person in einem psychologischen Konstrukt wird in der Regel durch Normwerte definiert. Diese Normwerte sind ebenfalls ganzzahlig. Damit kann der wahre Wert in einem Zufallsexperiment, wie oben beschrieben, nicht realisiert werden (vgl. Lumsden, 1976). Das heißt, dass es relativ unwahrscheinlich ist, den wahren Wert wirklich zu beobachten. Ziehen wir zur Veranschaulichung einen IQ-Wert heran. Eine Person wird immer einen ganzzahligen IQ-Wert in einem Test erzielen (z. B. 110), auch wenn der wahre Wert durch die Mittelung nicht ganzzahlig ist (z. B. 109.5). Somit ist nach Krauth (1995) bei der Veranschaulichung des wahren Werts Vorsicht geboten. Es ist jedoch möglich, den wahren Wert zu schätzen – dieser geschätzte Wert ist wiederum selten ganzzahlig (siehe *Abschnitt 4.8*). Steyer und Eid (2001) lehnen den Begriff „wahrer Wert“ gänzlich ab und verstehen den Erwartungswert über unendlich viele Testwiederholungen als Verhaltenstendenz einer Person in einer konkreten Situation. Würden wir Intelligenz als kontinuierlichen Messwert annehmen, dann kann die Konstruktion eines wahren Werts für Intelligenz wie in *Abbildung 2.6* veranschaulicht werden.

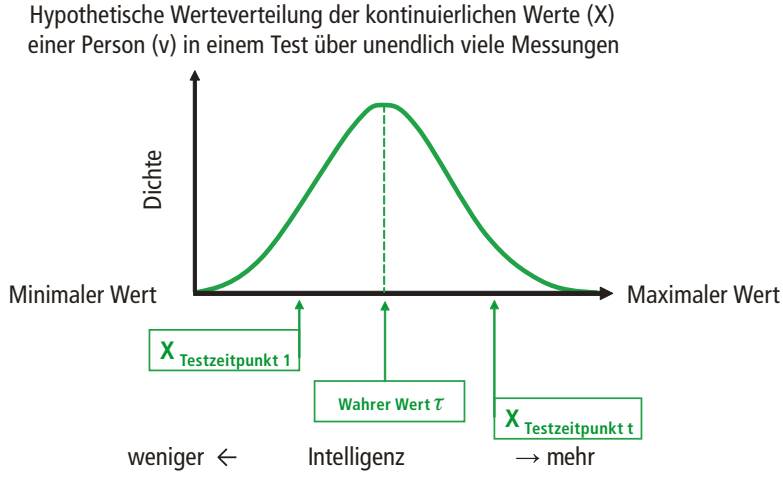

Abbildung 2.6: Konstruktion des wahren Werts.

Konstruktion des Messfehlers Nachdem nun der wahre Wert definiert ist, wird eine Fehlerdefinition vorgenommen: Der **Messfehler (Epsilon, ε)** einer **Person (v)** zu einem **Zeitpunkt (t)** setzt sich aus der Differenz zwischen **beobachtetem Messwert (x)** zum Zeitpunkt t und konstantem, über Zeitpunkte hinweg nicht variierenden, **wahren Wert (Tau, τ)** zusammen. Der Messfehler repräsentiert dabei alle unkontrollierten und unsystematischen Störeinflüsse bei der Messung:

$$\varepsilon_{vt} = \mathbf{x}_{vt} - \tau_v \qquad \rightarrow \textbf{Definition 2}$$

Beispiel 2.4 **Definition 2**

Das heißt der Fehler, der sich bei der Messung im Rahmen eines Hochsprungwettbewerbs für eine Person ergibt, wird ermittelt, indem vom beobachteten Ergebnis bei diesem einen Wettbewerb die wahre Hochsprungfähigkeit, als Erwartungswert der intraindividuellen Verteilung der Leistung eines Springers über unendlich viele Wettbewerbe, abgezogen wird.

Aus diesen Festsetzungen bzw. Definitionen von wahrem Wert und Messfehler ergeben sich weitere Folgerungen.

Folgerungen aus dem Kern der Klassischen Testtheorie

Erste Folgerung Aus den genannten Definitionen folgt unmittelbar, dass die Varianz der beobachteten Messwerte einer **Person** *v* über *t* **wiederholte Messungen** die Messfehlervarianz darstellt. Würden nämlich die Messungen messfehlerfrei sein, würde der beobachtete Wert einer Person in diesen Messungen nicht variieren, er wäre konstant. Da der wahre Wert eine Konstante darstellt, beeinflusst er auch nicht die Varianz.

$$\sigma^2(\mathrm{x}_{vt}) = \sigma^2(\mathrm{E}_{vt}) \qquad \rightarrow \textbf{Folgerung 1}$$

Beispiel 2.5 **Folgerung 1**

Das heißt, zieht man vom wahren Wert eines Hochspringers die beobachteten Werte in jedem Wettbewerb ab, quadriert die Differenzen, summiert sie auf und teilt das Ergebnis durch die Anzahl der Sprünge, erhält man die Varianz der beobachteten Werte. Diese Varianz ist ein Indikator für die Variabilität der Hochsprungleistung des Springers. Die gemittelten quadrierten Abweichungen von der wahren Hochsprungleistung des Springers stellen per Definition ($\varepsilon_{vt} = \mathbf{x}_{vt} - \tau_v$) den Messfehler dar.

Zweite Folgerung Die zweite Folgerung ergibt sich aus der Umformung der Gleichung der zweiten Definition $\varepsilon_{vt} = \mathbf{x}_{vt} - \tau_v$. Durch die Umformung erhält man die Bestandteile des beobachteten Werts. Demnach setzt sich der **beobachtete Messwert (x)** einer **Person (v)** zum **Zeitpunkt (t)** aus dem konstanten **wahren Wert (τ)** dieser Person und einem **Messfehler (ε)** zusammen.

$$\mathbf{x}_{vt} = \tau_v + \varepsilon_{vt} \qquad \rightarrow \textbf{Folgerung 2}$$

Beispiel 2.6 **Folgerung 2**

Das heißt bezogen auf unser Hochsprungbeispiel: Die Hochsprungleistung einer Person im Rahmen eines Wettbewerbs setzt sich aus der wahren Hochsprungleistung (ermittelt als Erwartungswert der Leistungen über unendlich viele Messwiederholungen bzw. Wettbewerbe) und dem Messfehler bei der aktuellen Messung zusammen.

Diese Folgerung wird auch als Verknüpfungsaxiom bezeichnet (Moosbrugger & Hartig, 2003; Bühner, 2006). Nach dem hier dargestellten Vorgehen, das sich an Steyer und Eid (2001) anlehnt, ist dies eine Folgerung aus den Definitionen (siehe Zusammenfassung).

Dritte Folgerung Die dritte Folgerung besagt, dass der **Erwartungswert E** des **Messfehlers (E, Epsilon)** (3a) über unendlich viele wiederholte **Messungen (t)** einer **Person (v)** und (3b) über alle **Personen (v)** einer beliebigen **Population** oder **Teilpopulation** null ist.

Da der Messfehler in der Klassischen Testtheorie als unsystematisch definiert ist, muss der Messfehler über unendlich viele Messwiederholungen null ergeben:

$$\mathbf{E}(\mathrm{E}_{vt}) = \mathbf{0} \qquad \rightarrow \textbf{Folgerung 3a}$$

Beispiel 2.7 **Folgerung 3a**

Die dritte Folgerung (3a) aus den Definitionen bedeutet auf unser Hochsprungbeispiel bezogen Folgendes: Würde ein Hochspringer unendlich viele Sprünge unter identischen Bedingungen ausführen, und würden wir für jeden Sprung den Messfehler bestimmen und dann daraus einen Mittelwert bilden, würde sich der Messfehler ausmitteln und null ergeben. Das würde darüber hinaus die Konsequenz haben, dass die Sprünge unabhängig voneinander erfolgen. Das heißt ein Sprung beeinflusst nicht die Leistung im nächsten Sprung. Betrachtet man als Beispiel einen Intelligenztest, dürfte eine Aufgabe keine Hinweise auf die Lösung der nächsten Aufgabe enthalten. Weiterhin dürfte der Test nicht zeitbegrenzt sein, da dann die Bearbeitung der nicht bearbeiteten Aufgaben von der Zeit abhängen würde, die vorher für die bearbeiteten Aufgaben verwendet wurde.

Exkurs 2.3 **Systematische Messfehler/Bias**

Ein systematischer Messfehler oder **Bias** (Rost, 2004, S. 36), auf das Hochsprungbeispiel bezogen, wäre, wenn bei unendlich vielen Sprüngen immer Rückenwind herrscht. Dieser systematische Messfehler würde sich nicht ausmitteln, da er die Leistung immer in eine bestimmte Richtung beeinflussen würde. In unserem Fall würde sich dieser Bias zum wahren Wert hinzuaddieren. Betrachten wir nun zur Veranschaulichung ein weiteres Beispiel: Würde nun bei jedem Sprung der Wind von vorne anstatt von hinten kommen, würde die Leistung des Hochspringers systematisch unterschätzt und dieser Bias würde vom wahren Wert abgezogen. Das heißt **systematische Messfehler verzerren** den **wahren Wert** entweder nach oben oder unten, können vom wahren Wert jedoch nicht ohne Weiteres getrennt werden.

Auch wenn **Personen (v)** einer Population unter identischen Bedingungen getestet werden (3b), mittelt sich der **Messfehler (**E, **Epsilon)** aus und der **Erwartungswert E** ergibt null.

$$\mathbf{E\,(E) = 0}$$ **→ Folgerung 3b**

Beispiel 2.8 **Folgerung 3b**

Wenn unterschiedliche Hochspringer einer Population, beispielsweise die Hochspringer einer Olympiade, unter identischen Bedingungen nur einen Wettbewerb durchführen, mittelt sich der Messfehler über alle Leistungen der Hochspringer des Wettbewerbs aus und der Messfehler ergibt null. Damit entspricht der Mittelwert der beobachteten Werte (μ_X) dem Mittelwert der wahren Werte (μ_T) in der Population:

$$\mathbf{E\,(X) = E\,(T) = \mu_x = \mu_\tau}$$

Auch hier werden wieder die großen Buchstaben X und T (Tau) verwendet, um anzuzeigen, dass hier eine Zufallsvariable gemeint ist. Das heißt, wenn wir alle Hochspringer einer Population oder Teilpopulation in einem Wettbewerb betrachten und deren Leistungen mitteln, erhalten wir nicht nur die beobachtete mittlere Leistung der Springer, sondern deren wahre mittlere Leistung.

Vierte Folgerung Die vierte Folgerung besagt, dass es keinen **Zusammenhang (ρ)** der variierenden **Messfehler (ε)** über **wiederholte Messungen (t)** mit dem konstanten **wahren Wert (τ)** einer einzelnen **Person** (v) gibt.

$$\rho\,(E_{vt}, \tau_v) = 0,\ \text{da}\ \tau\ \text{für jede Person eine Konstante}$$ **→ Folgerung 4a**

Der griechische Buchstabe ρ (Rho) steht für einen Zusammenhang oder eine Korrelation in einer Grundgesamtheit bzw. Population bzw. Teilpopulation. In diesem Fall wird der zugrunde liegende Zusammenhang des variierenden Fehlerwerts mit dem konstanten wahren Wert über unendlich viele wiederholte Messungen einer Person ermittelt. Diese Folgerung ergibt sich daraus, dass der **wahre Wert** einer Person eine **Konstante** darstellt und damit jede Person nur einen einzigen wahren Wert besitzt, was eine Folge der Konstruktion des wahren Werts darstellt. Eine Konstante besitzt keine Varianz und kann somit auch nicht mit einem von Messung zu Messung variierenden Messfehler korrelieren. Aus dieser Konstruktion folgt auch, dass es keinen Zusammenhang der Messfehler mit den wahren Werten einer Messung über alle Personen in der Population oder Teilpopulation gibt:

$$\rho\,(E, T) = 0 \qquad \rightarrow \textbf{Folgerung 4b}$$

Lumsden (1976, S. 254) weist in diesem Zusammenhang auf ein Problem hin, das ein Problem des konkreten Zählens von Punkten in Tests mit endlichem Wertebereich darstellt: Man stelle sich vor, ein Test bestünde aus acht mit Null und Eins kodierten dichotomen Items. Der maximale Summenwert einer solchen Messung würde acht ergeben und der minimale null. Alle Personen mit einem Wert von null erhielten demnach nur positive Messfehler und alle Personen mit dem Wert acht nur negative Messfehler, da Werte unter null und über acht nicht realisiert werden können. Lumsden weist darauf hin, dass dadurch der Zusammenhang zwischen Messfehlern und wahren Werten einer konkreten Messung mit einem Test scheinbar negativ ist, nämlich $\rho\,(E, T) < 0$. Das heißt je höher die wahren Werte ausfallen, desto negativer werden die Messfehler ausfallen. Dieses Problem kann umgangen werden, würde man mit einem mindestens intervallskalierten Maß, nämlich Höhe, wie es in Hochsprungwettbewerben üblich ist, arbeiten. Dieses Maß besitzt keine endlichen Werte, ist kontinuierlich und mindestens intervallskaliert. Übertragen auf einen psychologischen Test würde dies aber dann bedeuten, dass ein Test unendlich viele Items enthalten und der Summenwert intervallskaliert sein müsste. Dies ist jedoch sehr selten der Fall, beispielsweise bei der Messung der Reaktionszeit auf Reize.

Beispiel 2.9 **Folgerung 4a/b**

Auf einen einzelnen Springer bezogen heißt dies, der konstante wahre Wert eines Springers über unendliche viele Wettbewerbe korreliert nicht mit den unsystematischen Messfehlern, die in diesen Wettbewerben auftreten. Wird diese Folgerung auf interindividuelle Unterschiede (mehrere Springer) bezogen, heißt dies, dass Messfehler mit abnehmender oder zunehmender wahrer Fähigkeit der Hochspringer weder ab- noch zunehmen. Würde eine Korrelation vorliegen, misst der Test in unterschiedlichen Bereichen der Merkmalsausprägung unterschiedlich genau.

Fünfte Folgerung Aus den Definitionen und aus der vierten Folgerung ρ **(E, T) = 0** ergibt sich eine für die Definition der Reliabilität entscheidende fünfte Folgerung:

$$\sigma^2_X = \sigma^2_T + \sigma^2_E \qquad \textbf{→ Folgerung 5}$$

Der Beweis für diese Folgerung wird in *Kapitel 4* kurz dargestellt. Es wird hier gezeigt, dass sich die Varianz (σ^2_X) der beobachteten Werte X additiv aus der Varianz der wahren Werte (σ^2_T) und der Fehlervarianz (σ^2_E) zusammensetzt. Diese Formel ist die Basis für die Definition der Messgenauigkeit in der Klassischen Testtheorie (siehe *Kapitel 4*).

Beispiel 2.10 **Folgerung 5**

Für unser Beispiel heißt dies, dass die beobachtete Variation der Sprungleistungen der Hochspringer bei einem Wettbewerb sich aus der Variation der wahren Sprungleistungen der Hochspringer und der Variation der Messfehler zusammensetzt.

Sechste Folgerung Die sechste Folgerung aus den Definitionen ist, dass in der Population die **Messfehler (ε)** eines Tests A keinen **Zusammenhang (ρ)** mit den **wahren Werten (τ)** eines Tests B aufweisen (vgl. Zimmermann, 2009):

$$\rho\ (E_A, T_B) = 0 \qquad \textbf{→ Folgerung 6}$$

Der griechische Buchstabe ρ (rho) steht auch hier für einen Zusammenhang oder eine Korrelation in einer Grundgesamtheit. In diesem Fall wird der zugrunde liegende Zusammenhang nicht über wiederholte Messungen einer Person ermittelt, sondern über alle Messungen der Personen einer Population oder Teilpopulation.

Beispiel 2.11 **Folgerung 6**

Auf unser Hochsprungbeispiel bezogen heißt das, dass die Messfehler des Hochsprungwettbewerbs nicht mit den wahren Werten des Weitsprungwettbewerbs zusammenhängen. Definiert man den Fehler als unsystematisch, ergibt sich die Unkorreliertheit von wahren Werten eines Tests B und Messfehlern eines Tests A aus den Definitionen der Klassischen Testtheorie. Weist eine Messung jedoch systematische Messfehler auf, kann die Korrelation zwischen den wahren Werten eines Tests A und den systematischen Anteilen der Messfehler eines anderen Tests B auch größer oder kleiner als null ausfallen (siehe Box Korrelationen systematischer Messfehler). Würden wir diesen Sachverhalt auf unser Beispiel übertragen, würden die durch die unsystematischen Rückenwindverhältnisse erzielten Gewinne und Verluste des Hochsprungwettbewerbs für die einzelnen Springer mit den verschiedenen wahren Leistungen im Weitsprungwettbewerb zusammenhängen.

Zusatzannahme der Klassischen Testtheorie

Weiterhin wurde folgende zusätzliche Annahme formuliert, die nicht zwingend aus den beiden Definitionen der Klassischen Testtheorie abgeleitet werden kann (Zimmermann & Williams, 1977, S. 136): Die **Messfehler (E)** eines Tests A (z.B. Gedächtnistest) weisen in der Population keinen **Zusammenhang (ρ)** mit den **Messfehlern (E)** eines anderen Tests B (z.B. Leistungsmotivationstests) auf:

$$\rho\,(E_A, E_B) = 0 \qquad \rightarrow \textbf{Zusatzannahme}$$

Beispiel 2.12 **Zusatzannahme**

Nehmen wir an, dass Personen einen Zehnkampf durchführen. Im Rahmen eines Zehnkampfs würden dann die Messfehler des Hochsprungwettbewerbs nicht mit den Messfehlern des Weitsprungwettbewerbs zusammenhängen.

Liegt Eindimensionalität vor, liegt auch lokale Unabhängigkeit vor und die Messfehler sind unkorreliert. Die Unkorreliertheit der Messfehler muss jedoch durch geeignete statistische Analysen, beispielsweise durch konfirmatorische Faktorenanalysen, nachgewiesen werden. Man kann jedoch nicht umgekehrt davon ausgehen, dass bei Vorliegen von unkorrelierten Messfehlern oder lokaler Unabhängigkeit Eindimensionalität vorliegt (Zimmermann & Williams, 1977, S. 140). Steyer und Eid (2001, S. 104) weisen darauf hin, dass die Annahme eines nicht vorhandenen Zusammenhangs von Messfehlern in der Praxis falsch sein kann. Bei der Konstruktion des wahren Werts wurde darauf hingewiesen, dass die Messungen für jede einzelne Person unabhängig voneinander sein müssen. Diese damit gemachte Annahme wird auch als lokale stochastische Unabhängigkeit bezeichnet. Liegen lokal stochastisch unabhängige Messungen vor, resultieren daraus unkorrelierte Messfehler. Aus den Definitionen von wahrem Wert und Messfehler ohne die zusätzliche Annahme lokal stochastisch unabhängiger Messungen folgt die Annahme unkorrelierter Messfehler jedoch nicht.

Durch die aufgeführte Definition und die dargestellten Annahmen lässt sich die Messgenauigkeit oder **Reliabilität** (ρ_{tt}) einer Messung herleiten (ausführlich in *Kapitel 4* beschrieben). Die Reliabilität ist ein wichtiges Gütekriterium einer Messung bzw. eines Tests. Die Reliabilität ist ein Wert, der zwischen null und eins liegt und die Messgenauigkeit eines Tests angibt. Ein Wert von eins bedeutet, dass eine Messung perfekt genau ist. Ein Wert von null bedeutet, dass keine Messung im eigentlichen Sinne vorliegt. Zunächst soll jedoch auf Kritikpunkte an der Klassischen Testtheorie eingegangen werden, bevor dann weitere Gütekriterien von Messungen genauer erläutert werden.

Exkurs 2.4 Korrelationen systematischer Messfehler

Ein anschauliches Beispiel für systematische Messfehler liefern De Gruijter und van den Kamp (2008, S. 13). Dieses Beispiel soll hier erweitert und zum besseren Verständnis systematischer Messfehler kurz dargestellt werden. Nehmen wir an, dass alle Items eines Tests zur Erfassung von Depression auch Angst messen. In diesem Fall enthält der beobachtete Messwert x einer Person *v* in einem Test folgende Komponenten:

$$x_v = \tau_{Depression,\ v} + \tau_{Angst,\ v} + \varepsilon_v$$

Da die Messung des Konstrukts Depression angestrebt wurde, sind die Komponenten des beobachteten Werts x_v, die nicht Depression erfassen, neutral formuliert unerwünscht und werden als systematischer Messfehler bezeichnet ($T_{Angst} = E_{Angst,\ systematisch}$). Es ist jedoch nur unter bestimmten Bedingungen möglich, einen systematischen Messfehler vom wahren Wert einer Messung zu trennen, denn der systematische Messfehler addiert sich auf den wahren Wert einer Messung:

$$x_v = \tau_{Depression\ +\ Angst,\ v} + \varepsilon_v$$

Dieser systematische unerwünschte Varianzanteil Angst, der sich auf den systematischen wahren Wert Depression addiert, kann nun mit den wahren Werten anderer Messungen, beispielsweise Leistungsorientierung, korrelieren. Er wird, weil er eigentlich einen Fehler bei der Messung darstellt und unerwünscht ist, als systematischer Messfehler $E_{systematisch}$ bezeichnet:

$$\rho\ (E_{Angst,\ systematisch}, T_{Leistungsorientierung}) \neq 0$$

Wir stellen uns nun weiter vor, dass alle Items eines Tests, der Leistungsorientierung erfassen soll, neben Leistungsorientierung ebenfalls Angst messen. Wir wissen auch, dass der Depressionstest Angst misst. In diesem Fall korrelieren die systematischen Messfehler von Depression und Leistungsorientierung, da die systematischen Messfehler beider Tests Unterschiede in der Angst ausdrücken:

$$\rho\ (E_{Depression}, T_{Leistungsorientierung}) =$$

$$\rho\ (E_{[Angst,\ systematisch\ |\ Depression]}, E_{[Angst,\ systematisch\ |\ Leistungsorientierung]}) \neq 0.$$

Wie entdecke ich systematische Messfehler? Ob die Messfehler von Items systematisch sind, kann mithilfe von konfirmatorischen Faktorenanalysen (siehe *Kapitel 7*) und einer gleichzeitig guten diskriminanten Validierung untersucht werden. Diskriminante Validierung meint, dass Konstrukte, die dem zu messenden Konstrukt ähnlich sind, das Konstrukt selbst aber nicht messen, ebenfalls mit Tests erfasst werden. Der neu entwickelte Test und die Tests, die angrenzende Konstrukte messen sollen, werden dann gemeinsam einer Faktorenanalyse unterzogen. Erfassen diese Tests tatsächlich Verschiedenes, sollten sie möglichst auf voneinander inhaltlich unterschiedlichen Faktoren laden. Eine andere Möglichkeit, den wahren Wert in verschiedene Komponenten zu zerlegen, bietet die Generalisierbarkeitstheorie (Cronbach, Gleser & Rajaratnam, 1963) als Erweiterung der Klassischen Testtheorie.

Welche Folgen hat es, wenn systematische Messfehler vorliegen? Liegen systematische Messfehler vor, hat dies Auswirkungen auf die Interpretation der Merkmalsausprägung: Der Test ist nicht mehr eindimensional. Es wird dann mehr als ein Konstrukt gemessen, nämlich in unserem Beispiel sowohl Depression als auch Angst. Wird bei einer Person nun eine hohe Merkmalsausprägung beobachtet, ist nicht klar, ob diese durch Angst, Depression oder beide Konstrukte bedingt ist.

2.2.2 Kritische Anmerkungen zur Klassischen Testtheorie

Was muss ich über Schwächen der Klassischen Testtheorie wissen?

KTT „nur" Messfehlertheorie Ein wichtiger Kritikpunkt an der Klassischen Testtheorie ist, dass sie keine Verbindung zwischen einem Merkmal und der Itembeantwortung herstellt. Die Klassische Testtheorie ist eine reine **Messfehlertheorie** und beschäftigt sich nur mit den Komponenten eines beobachteten **Messwerts**. Borsboom und Mellenbergh (2002) geben ein anschauliches Beispiel dafür, dass (wahre) Werte im Rahmen der Klassischen Testtheorie nicht für ein **reflektives Konstrukt** stehen müssen. Betrachten wir einen Test mit folgenden Items: „Ich wäre gerne ein militärischer Befehlshaber" und „Ich bin über 1.80 m groß". Nun wird zu beiden Itemantworten konsistent eine Zufallszahl addiert. Danach wird der Summenwert mit der Anzahl der Buchstaben des Vornamens der Person multipliziert. Das ergibt dann eine Zahl. Über verschiedene Wiederholungsmessungen hat jede Person einen „wahren" Wert in diesem Test. Möglicherweise hängen sogar die Testwerte aufeinanderfolgender Messungen in hohem Maße zusammen. Allerdings ist offensichtlich, dass hier kein „sinnvoll" definierbares Konstrukt gemessen wird.

Nicht erkannte Mehrdimensionalität als Problem Erst mit der Einführung von eindimensionalen Konstrukten und als Folge mit der Annahme **nicht korrelierter Messfehler** verliert die Definition eines wahren Werts ihre Beliebigkeit. Die Forderung der Eindimensionalität von Messungen wird in der Klassischen Testtheorie nicht überprüft. Die Klassische Testtheorie begnügt sich mit einer schwächeren Annahme, der Annahme der lokalen Unabhängigkeit. Diese schwächere Annahme reicht für die Ableitung der Reliabilität (siehe *Kapitel 4.2*) aus. Eine Überprüfung der Eindimensionalität von Messungen muss jedoch auf jeden Fall erfolgen und kann mithilfe von konfirmatorischen Faktorenanalysen durchgeführt werden (siehe *Kapitel 7*).

KTT-Annahmen nicht überprüfbar bzw. schwer haltbar Wie bereits geschildert, sind einige Annahmen der Klassischen Testtheorie nicht überprüfbar, sondern ergeben sich logisch aus der Festsetzung des beobachteten Wertes als wahrer Wert plus Messfehler. Auch wenn die Klassische Testtheorie mathematisch durchaus befriedigend formuliert ist (vgl. Fischer, 1974), sind manche der Modellannahmen in der psychologischen Praxis nur schwer haltbar (vgl. Fischer, 1974, S. 26). Wie Fischer (1974, S. 28) richtig bemerkt, können nicht alle Einflüsse auf das Testergebnis als Zufallseinflüsse abgetan werden. Übungs- und Transfereffekte wirken sich unter Umständen systematisch auf die Testleistung aus und verändern die wahre Leistungsfähigkeit einer Person. Auch systematische Wechselwirkungen zwischen Person und Situation, die als

unechte (spurious) Messfehler bezeichnet werden, verzerren die Messung (Ziegler & Bühner, 2009). In letzter Konsequenz ist damit sowohl die Annahme eines fehlenden Zusammenhangs zwischen wahrem Wert und Messfehler zu bezweifeln als auch die nur durch zufällige Einflüsse überlagerte Konstanz des wahren Wertes über verschiedene Messwiederholungen. Auch die Annahme des fehlenden Zusammenhanges der Fehlerwerte ist nicht zwingend.

Stichprobenabhängigkeit der KTT-Kennwerte Ein Problem mit großem Gewicht ist, dass die Testwerte der Klassischen Testtheorie stichprobenabhängig sind. Das heißt für Abiturienten mag ein Intelligenztest andere Testkennwerte (z.B. Messgenauigkeit) besitzen als für Hauptschüler oder Realschüler. Zu berücksichtigen ist auch die Tatsache, dass die Werte einer Person in verschiedenen Tests, die nach der Klassischen Testtheorie konstruiert wurden und dasselbe Konstrukt messen sollen, nicht direkt vergleichbar sind. Das heißt ein Summenwert von 20 gelösten Items kann in einem Test eine gute und in einem anderen Test eine schlechte Leistung bedeuten. Dies ist deshalb so, weil Tests unterschiedliche Itemanzahlen enthalten und darüber hinaus die Testitems unterschiedlich schwer sein können.

Generell ist jedoch die Frage zu stellen, ob die Anwendung der Klassischen Testtheorie auf Tests mit dichotomen Items oder Ratingskalen überhaupt angemessen ist, ist sie doch nur für intervallskalierte und kontinuierliche Merkmale ohne Einschränkung anwendbar. Problematisch ist daher die Anwendung der Klassischen Testtheorie auf dichotome Items oder auf Ratingskalen, die diese Eigenschaften nicht besitzen. Würde man, wie Rost (1999) vorschlägt, solche Items mit dem Rasch-Modell analysieren und würden diese jenem Modell genügen, erhielte man Messwerte auf dem gewünschten Skalenniveau, auf die dann die Messfehlertheorie ohne Weiteres anwendbar wäre.

Die oben genannten Punkte zu Unzulänglichkeiten der Klassischen Testtheorie, die noch problemlos erweitert werden könnten (vgl. Amelang & Zielinski, 2002, S. 62 f.), sind grundsätzlich berechtigt und schwerwiegend. Insgesamt könnte nun der Eindruck entstehen, dass die Klassische Testtheorie so unzulänglich ist, dass sie in der Praxis nicht eingesetzt werden sollte. Dennoch ist die Klassische Testtheorie nach Stumpf (1996, S. 416) nicht unbrauchbar, sondern hat sich in der Praxis bewährt (vgl. auch Amelang & Zielinski, 2002). Dies mag daran liegen, dass die Brauchbarkeit eines Tests vor allem von einer inhaltlich begründeten Konstruktion der Items und der Skalen abhängt. Solange dieser Prozess sachgemäß durchgeführt wird, überdeckt er auch Schwächen der testtheoretischen Annahmen. Nach der folgenden Zusammenfassung soll kurz und überblicksartig auf wenige Grundüberlegungen des Rasch-Modells eingegangen werden und danach auf die Haupt- und Nebengütekriterien psychologischer Tests.

ZUSAMMENFASSUNG

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

Kern der Klassischen Testtheorie Um die unterschiedlichen Werte aufeinanderfolgender Messungen zu erklären, wird im Rahmen der Klassischen Testtheorie zunächst ein wahrer Wert für eine Person konstruiert. Dieser wahre Wert ist definiert als der Mittelwert oder genauer der Erwartungswert über unendlich viele gedachte Messwiederholungen an derselben Person:

$$\tau_v = \mathbf{E}\,(\mathbf{X}_{vt}) \qquad \rightarrow \textbf{Definition 1}$$

Dabei geht die Klassische Testtheorie davon aus, dass eine Person einer Population zufällig entnommen wurde und aus ihrer intraindividuellen Werteverteilung zufällig ein Wert beobachtet wurde. Der Messfehler in der Klassischen Testtheorie wird als unsystematisch angenommen und wird wie folgt ermittelt: Der Messfehler bei der Messung einer Person v zu einem festgelegten Zeitpunkt t ist als Differenz des beobachteten Wertes einer Person abzüglich des konstanten wahren Werts dieser Person definiert.

$$\varepsilon_{vt} = \mathbf{x}_{vt} - \tau_v \qquad \rightarrow \textbf{Definition 2}$$

Folgerungen aus den Definitionen Aus den beiden Definitionen ergeben sich vier Folgerungen.

Zunächst folgt aus den Definitionen unmittelbar, dass die Varianz der beobachteten Werte einer Person der Varianz der Messfehler dieser Person entspricht.

$$\sigma^2(\mathrm{X}_{vt}) = \sigma^2(\mathrm{E}_{vt}) \qquad \rightarrow \textbf{Folgerung 1}$$

Weiterhin folgt aus den Axiomen, in welche Bestandteile sich der wahre Wert aufgliedert.

$$\mathbf{x}_{vt} = \tau_v + \varepsilon_{vt} \qquad \rightarrow \textbf{Folgerung 2}$$

Der Erwartungswert des Messfehlers über unendlich viele Messungen einer Person oder über alle Personen einer beliebigen Population oder Teilpopulation ist null.

$$\mathbf{E}\,(\mathrm{E}_{vt}) = \mathbf{0} \qquad \rightarrow \textbf{Folgerung 3a}$$

$$\mathbf{E}\,(\mathrm{E}) = \mathbf{0} \qquad \rightarrow \textbf{Folgerung 3b}$$

Es besteht kein **Zusammenhang** der Messfehler mit dem konstanten wahren Wert einer Person über verschiedene Messzeitpunkte. Daraus folgt, dass die Messfehler mit den wahren Werten eines Tests in der Population bzw. Teilpopulation ebenfalls keinen Zusammenhang $\rho\,(\varepsilon_v, \tau_v) = \mathbf{0}$ aufweisen.

$$\rho\,(\mathrm{E}_{vt}, \mathrm{T}_v) = \mathbf{0} \qquad \rightarrow \textbf{Folgerung 4a}$$

$$\rho\,(\mathrm{E}, \mathrm{T}) = \mathbf{0} \qquad \rightarrow \textbf{Folgerung 4b}$$

Mithilfe der beiden Axiome und Folgerung 4b lässt sich eine weitere Folgerung ableiten, die für die Berechnung der Messgenauigkeit von Tests wichtig ist: Die Varianz der beobachteten Werte setzt sich additiv aus der Varianz der wahren Werte und der Fehlervarianz zusammen:

$$\sigma^2{}_X = \sigma^2{}_\mathrm{T} + \sigma^2{}_\mathrm{E} \qquad \rightarrow \textbf{Folgerung 5}$$

Schließlich hängt der Messfehler eines Tests A nicht mit dem wahren Wert eines Tests B zusammen.

$$\rho\,(\mathrm{E}_A, \mathrm{T}_B) = 0 \qquad \rightarrow \textbf{Folgerung 6}$$

Zusatzannahme aus den Definitionen. Folgende zusätzliche Annahme muss darüber hinaus im Rahmen der Klassischen Testtheorie gemacht werden: Der Messfehler eines Tests A weist keinen Zusammenhang mit dem Messfehler eines Tests B auf.

$$\rho\,(\mathrm{E}_A, \mathrm{E}_B) = 0 \qquad \rightarrow \textbf{Zusatzannahme}$$

Kurzübersicht In der Zusammenschau sind hier noch einmal alle Axiome, Folgerungen und Zusatzannahmen dargestellt. Sie dienen im Prinzip dazu, die Reliabilität als ein Gütekriterium von Tests bzw. Messungen zu definieren (vgl. *Kapitel 4*):

$\tau_v = \mathbf{E}\,(\mathbf{X}_{vt})$	→ **Definition 1**
$\varepsilon_{vt} = \mathbf{x}_{vt} - \tau_v$	→ **Definition 2**
[$(\rho\,(\mathbf{X}_A, \mathbf{X}_B \mid \mathbf{v}) = 0$ **für alle v**	→ **Definition 3** → **lokale stoch. Unabhängigkeit]**
$\sigma^2(\mathrm{X}_{vt}) = \sigma^2(\mathrm{E}_{vt})$	→ **Folgerung 1**
$\mathbf{x}_{vt} = \tau_v + \varepsilon_{vt}$	→ **Folgerung 2**
$\mathbf{E}\,(\mathrm{E}_{vt}) = 0$ **und** $\mathbf{E}\,(\mathrm{E}) = 0$	→ **Folgerung 3 a/b**
$\rho\,(\mathrm{E}_{vt}, \tau_v) = 0$ **und** $\rho\,(\mathrm{E}, \mathrm{T}) = 0$	→ **Folgerung 4 a/b**
$\sigma^2{}_X = \sigma^2{}_\mathrm{T} + \sigma^2{}_\mathrm{E}$	→ **Folgerung 5**
$\rho\,(\mathrm{E}_A, \mathrm{T}_B) = 0$	→ **Folgerung 6**
$\rho\,(\mathrm{E}_A, \mathrm{E}_B) = 0$	→ **Zusatzannahme**

Führt man, wie beispielsweise Fischer (1974) oder Krauth (1995), lokale stochastische Unabhängigkeit der Messungen **[**$(\rho\,(\mathbf{X}_A, \mathbf{X}_B \mid \mathbf{v}) = 0$ **für alle v]** als dritte Definition ein, wird aus der hier dargestellten Zusatzannahme eine weitere Folgerung dieser neu eingeführten Definition. Die Annahme der lokalen stochastischen Unabhängigkeit definiert letztendlich Messfehler als zufällig. Eine genaue Definition von lokaler stochastischer Unabhängigkeit findet sich in *Kapitel 8.1.1*. Wie man an diesen Beispielen sieht, variiert die Darstellung der Klassischen Testtheorie in der Literatur. Die hier dargestellten Folgerungen und weitere Folgerungen werden bei Krauth (1995) als Sätze der Klassischen Testtheorie bezeichnet. Sätze sind aus Axiomen bzw. Definitionen abgeleitete Aussagen. Die Folgerungen 2, 3b, 4b, 6 und die Zusatzannahme werden hingegen häufig als Axiome bezeichnet, obwohl es sich streng genommen um Folgerungen plus Zusatzannahme aus den beiden Definitionen handelt. Man findet auch Darstellungen, in denen Definition 1 als Existenzaxiom, Folgerung 2 als Verknüpfungsaxiom und Folgerung 4b als Unabhängigkeitsaxiom (vgl. Moosbrugger, 2007) bezeichnet wird, da diese von den Autoren als Kern der Klassischen Testtheorie angesehen werden. Eine weitere wichtige Unterscheidung der Folgerungen kann wie folgt getroffen werden: Es gibt zum einen Folgerungen, die sich auf die Messwerte einer Person beziehen, und zum anderen Folgerungen, die sich auf Messwerte unterschiedlicher Personen beziehen. Während sich die Definitionen 1, 2 und 3 sowie die Folgerungen 2, 3a, 4a auf die wiederholte Messung (auch als Replikation bezeichnet) an einer Person beziehen, beziehen sich die Folgerungen 1, 3b, 4b, 5, 6 und die Zusatzannahme auf interindividuelle Unterschiede zwischen Personen.

Kritik an der Klassischen Testtheorie Die Klassische Testtheorie ist eine reine Messfehlertheorie. Sie stellt Definitionen zur Verfügung, mit deren Hilfe man ein Reliabilitätsmaß definieren kann. Problematisch ist vor allem, dass die Klassische Testtheorie nur unsystematische Messfehler betrachtet. In der Praxis treten jedoch systematische Messfehler auf und führen dazu, dass das Klassische Modell nicht mehr gilt und daher auch die daraus abgeleitete Reliabilität nicht präzise geschätzt wird. Ebenfalls problematisch ist, dass die Klassische Testtheorie und damit auch die Definition der Messgenauigkeit weitgehend auf Mittelwerten, Varianzen und Kovarianzen basiert. Das hat die Konsequenz, dass für jede Stichprobe die Messgenauigkeit oder Reliabilität einer Messung neu bestimmt werden muss. Die Reliabilität ist dabei ein Wert, der zwischen null und eins liegt und die Höhe der Messgenauigkeit einer Messung angibt. Dabei bedeutet ein Wert von eins, dass die Messung perfekt genau ist.

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

2.3 Kurzer Ausblick auf die Probabilistische Testtheorie

Die Grundlagen der Probabilistischen Testtheorie werden in *Kapitel 8* beschrieben. Daher wird an dieser Stelle nur ein ganz kurzer Ausblick gegeben. Da dieses Kapitel jedoch testtheoretische Grundlagen behandelt, darf eine Kurzbeschreibung des wichtigsten Probabilistischen Testmodells nicht fehlen: Die folgenden Ausführungen beziehen sich folglich auf das Rasch-Modell. Im Rasch-Modell geht es im Gegensatz zur Klassischen Testtheorie darum, wie Antworten auf Items zustande kommen. Aus diesem Grund werden Antwortmuster untersucht. Die beobachteten Antwortmuster müssen einem bestimmten Modell folgen. Dieses Modell sagt voraus, dass mit steigender Personenfähigkeit die **Wahrscheinlichkeit** einer Itemlösung zunimmt. Die **Lösungswahrscheinlichkeit** für ein bestimmtes Item hängt von zwei Parametern ab: (1) der **Fähigkeit** oder **Eigenschaftsausprägung** einer Person sowie (2) der **Schwierigkeit** eines Items. Diese **Beziehung** zwischen Personenfähigkeit und Lösungswahrscheinlichkeit eines Items ist **probabilistisch**. Das heißt auch eine Person mit geringer Fähigkeit im Vergleich zur Schwierigkeit eines Items hat eine, wenn auch geringe, Wahrscheinlichkeit, ein solches Item zu lösen. Im Rahmen der Probabilistischen Testtheorie können verschiedene Modelltests durchgeführt werden. Wird das Modell durch den Modelltest nicht abgelehnt, beinhaltet der **Summenwert** der Itemantworten auch wirklich alle Informationen über den **Ausprägungsgrad einer Person** auf der latenten Variablen. Damit ist der Summenwert auch eine **erschöpfende Statistik** der Personenfähigkeit. Erschöpfend heißt, der Summenwert einer Person liefert alle Informationen über die Fähigkeitsausprägung der Person. Das bedeutet, die Interpretation des Antwortmusters ist unnötig, weil sie keine zusätzliche Information über die Merkmalsausprägung der Personen enthält. Ein Item ist dann ein guter Indikator für eine latente Variable, wenn die Leistung in diesem Item komplett auf die Fähigkeitsausprägung auf der latenten Variablen zurückzuführen ist und nicht auf andere Variablen. Dies ist eine höchst wünschenswerte Annahme für die Testkonstruktion, da sie eine sehr präzise Definition von Itemhomogenität darstellt (vgl. Stelzl, 1993). Formalisiert wird diese Eigenschaft durch die **lokale stochastische Unabhängigkeit**. Wenn das Rasch-Modell durch den Modelltest nicht verworfen wird, liegt auch diese Eigenschaft vor. Das Rasch-Modell implementiert damit eine echte Messtheorie in die Psychologie. Weitere Ausführungen zum Rasch-Modell finden sich in *Kapitel 8.2*.

2.4 Haupt- und Nebengütekriterien

Was macht einen guten psychometrischen Test aus?

Es gibt verschiedene anerkannte Kriterien, nach denen die Güte eines Tests beurteilt werden kann. Nur wenn die nachfolgend beschriebenen Gütekriterien vollständig und nachvollziehbar im Testhandbuch aufgeführt sind, kann ein Test in seiner Güte beurteilt werden. Man unterteilt Neben- und Hauptgütekriterien (Lienert & Raatz, 1998; Kubinger & Proyer, 2005). Zu den **Hauptgütekriterien** gehören **Objektivität**, **Reliabilität, Validität** und die **Skalierbarkeit** eines Tests. Alle diese Begriffe lassen sich weiter differenzieren. Im Hinblick auf die Testauswahl und Testbeurteilung ist ein sicherer Umgang mit diesen Begriffen unumgänglich.

Objektivität

- Durchführungsobjektivität
- Auswertungsobjektivität
- Interpretationsobjektivität

Reliabilität

- Konsistenz
- Retest-Korrelation
- Paralleltest-Korrelation

Validität

- Inhaltsvalidität
- Konstruktvalidität
- Kriteriumsvalidität

Skalierbarkeit

2.4.1 Hauptgütekriterien

Was sind die wichtigsten Indikatoren für einen guten psychometrischen Test?

Objektivität

Ein psychometrischer Test sollte bei der Durchführung und der Auswertung **objektiv** sein. Das heißt die Durchführung und Auswertung des Tests sowie die Interpretation der Testleistung einer Person variieren nicht, auch, wenn unterschiedliche Testleiter den Test durchführen, auswerten oder interpretieren. Das bedeutet, die ganze Testdurchführung sollte standardisiert ablaufen. Dazu müssen die Durchführungsbedingungen beschrieben sein. Auswertung und Interpretation des Tests sollen standardisiert sein, damit jeder Untersucher die gleiche Testleistung für ein und denselben Probanden ermittelt und diese auch gleich interpretiert.

Definition: Objektivität
Unter Objektivität versteht man den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind.

Man unterscheidet drei Arten von Objektivität:

- Durchführungsobjektivität

  Die Durchführung eines Tests darf nicht von Untersuchung zu Untersuchung variieren. Dazu muss genau definiert sein, wie und unter welchen Bedingungen ein Test, Fragebogen, Interview oder eine Verhaltensbeobachtung durchzuführen ist. Zeitbegrenzung oder Hilfestellungen bei der Beantwortung der Fragen müssen vorgegeben werden. Die größte Sorgfalt sollte auf die Instruktion verwendet werden, denn dadurch können Rückfragen an den Untersucher minimiert werden. Dieser läuft wiederum nicht Gefahr, den Personen unterschiedliche Hilfestellungen zu geben. Betrachten wir beispielsweise den Test d2 (Brickenkamp, 2002). Bei diesem Test ist die Aufgabe der Versuchsperson, den Buchstaben „d“ mit zwei Strichen unter den Buchstaben „p“ und „d“ mit unterschiedlicher Anzahl an Strichen zu markieren. Hält sich nun der Testleiter nicht an die genaue Testinstruktion, beeinträchtigt dies unter Umständen das Testergebnis, und die wahre Leistungsfähigkeit des Probanden wird über- oder unterschätzt. Dies ist beispielsweise dann wahrscheinlich, wenn der Testleiter bei einer Testung folgende Instruktion gibt: „Es kommt darauf an, sorgfältig und schnell zu arbeiten“, und in der anderen Testung: „Es kommt darauf an, schnell zu arbeiten.“ Häufig muss auch festgelegt werden, welche Ausschlusskriterien zu berücksichtigen sind, damit eine Testleistung gültig ist. Bleiben wir beim Test d2. Streicht eine Person beispielsweise nur den Buchstaben p mit zwei Strichen an, ist die Testleistung offensichtlich ungültig, da die Person ja den Buchstaben d mit zwei Strichen markieren sollte. Es könnte aber sein, dass bereits zehn oder zwölf falsch markierte Buchstaben ausreichen, um die Testleistung nicht mehr interpretieren zu können. Dies muss im Handbuch klar festgelegt sein.

- Auswertungsobjektivität

  Jeder Auswerter muss die gleichen Punkt- oder Leistungswerte eines Probanden ermitteln. Dazu sind genaue Auswertungsvorschriften nötig. Hilfreich sind Schablonen und Auswertungsblätter bei Tests und Fragebögen bzw. verhaltensverankerte Ratingskalen bei Interviews und Verhaltensbeobachtungen. In diesen Hilfsmitteln sind die für die Auswertung relevanten Daten enthalten. Aber auch Schablonen garantieren nicht immer eine ausreichende Auswertungsobjektivität. So kann das Auflegen von Schablonen selbst wieder fehlerträchtig sein. Bleiben wir bei dem Beispiel des Tests d2. Hier müssen über 600 Zeichen mithilfe der Schablonen überprüft werden. Würden zwei Auswerter unabhängig voneinander denselben Test auswerten, ist nicht gesagt, dass jeder Auswerter auch alle falsch durchgestrichenen Zeichen entdeckt. Es könnte sein, dass der Auswerter Markus zwölf falsch markierte Zeichen mithilfe der Schablonen entdeckt und der Auswerter Matthias 15 falsch markierte Zeichen. Ebenso können auch bei der Verwendung von verhaltensverankerten Ratingskalen Fehler auftreten. In solchen Skalen ist in Form von Verhaltensankern aufgeführt, für welche Antwort bzw. welches Verhalten wie viele Punkte vergeben werden. Fehler treten auf, wenn die Auswerter nicht das gleiche Verständnis der Verhaltensanker haben. In solchen Fällen wäre es hilfreich, die Auswertungsobjektivität empirisch zu überprüfen. Dazu ist eine Reihe von Indizes geeignet, wie beispielsweise Cohens Kappa, Scotts Pi oder die Intraklassenkorrelation (vgl. Wirtz & Caspar, 2002). Die lapidare Anmerkung, dass die Auswertungsobjektivität durch das Bereitstellen von Schablonen bzw. verhaltensverankerten Ratingskalen zur Auswertung vorliegt, ist nicht ausreichend, um eine hohe Auswertungsobjektivität zu

belegen. Die Auswertungsobjektivität hängt nicht zuletzt von der Art und Weise ab, wie gefragt wird. Wird die Frage offen, ohne festgelegte Antwortmöglichkeiten gestellt, muss sehr exakt definiert werden, was als „richtig" zu bewerten ist. Bei manchen Tests ist es notwendig, Regeln zu definieren, ab wann ein Ergebnis zu werten ist und ab wann nicht. Auch dies lässt sich am Beispiel des Tests d2 zeigen. Im Test d2 lässt sich ein so genanntes Ü-Syndrom diagnostizieren. Ein Ü-Syndrom liegt dann vor, wenn die Mengenleistung (Gesamtzahl bearbeiteter Zeichen, GZ) über einem Prozentrang von PR = 90 liegt und die Sorgfaltsleistung (Fehlerprozentwert, F%) unter PR = 10. Dann liegt der Verdacht nahe, dass die Testleistung instruktionswidrig zustande gekommen ist. In einem solchen Fall sollte das Ergebnis mit Vorsicht interpretiert werden. In manchen Fällen kann es vorkommen, dass ein Ergebnis gar nicht interpretiert werden kann. Um Auswertungsobjektivität zu gewährleisten, müssen diese Richtlinien im Handbuch festgelegt werden.

- Interpretationsobjektivität

  Jeder Auswerter sollte möglichst zur gleichen Beurteilung oder Interpretation der Testergebnisse kommen, wie etwa, ob der Probandenwert als durchschnittlich oder über- bzw. unterdurchschnittlich im Vergleich zu einer Normgruppe in einem bestimmten Test einzuordnen ist. Interpretationsobjektivität schließt ausreichend große Normstichproben und ausreichend geprüfte Gütekriterien mit ein, so dass man davon ausgehen kann, dass jede Person mit dem gleichen Maßstab beurteilt wird. Allerdings ist dies alleine nicht ausreichend. Häufig fehlen standardisierte Interpretationen. Manchmal wird dies damit entschuldigt, dass man den Testleiter nicht an ein bestimmtes Schema binden möchte. Als Begründung wird herangezogen, dass der Test weitaus mehr Interpretationsmöglichkeiten bietet, als durch eine standardisierte Interpretation zur Verfügung gestellt werden können. Standardisierte Interpretationsmöglichkeiten sollten jedoch für jeden Test vorliegen.

Reliabilität

Unter Reliabilität versteht man den Grad der Genauigkeit, mit dem ein Test ein bestimmtes Merkmal misst, unabhängig davon, was er zu messen beansprucht. Das heißt, lässt man die Frage, was ein Test misst, außen vor, kann man trotzdem bestimmen, wie genau der Test eine Eigenschaft oder Fähigkeit misst. Dabei kann die Reliabilität einer Messung auf unterschiedliche Arten geschätzt werden (siehe auch *Kapitel 4.4*). Basis dafür sind Korrelationskoeffizienten. Dabei wird derselbe oder ein paralleler Test zu zwei Testzeitpunkten wiederholt vorgegeben oder in Testteile aufgeteilt. Dann wird die Korrelation der Messwerte der beiden Testteile berechnet oder die Messwerte der wiederholten Testungen miteinander korreliert. Dies wird im Folgenden noch etwas ausführlicher dargestellt.

Definition: Reliabilität
Die Reliabilität gibt den Grad der Messgenauigkeit eines Messwerts an.

- Innere bzw. interne Konsistenz/aufgewertete Halbierungskorrelation

  Unter aufgewerteter Halbierungskorrelation versteht man Folgendes: Der Test wird in möglichst „gleiche" Testhälften unterteilt und die Messwerte dieser Testhälften werden miteinander korreliert. Dabei wird als zusätzlicher Faktor die Testlänge

berücksichtigt: Man kann zunächst grob sagen, dass mit steigender Testlänge die Reliabilität monoton zunimmt. Unter innerer Konsistenz wird Folgendes verstanden: Jedes einzelne Item wird als eigenständiger Testteil angesehen, und die innere Konsistenz berücksichtigt den Zusammenhang zwischen den Items und die Testlänge. Sie steht dabei für die Messgenauigkeit des Tests zu einem bestimmten Messzeitpunkt gemessenen Messwerts.

- Retest-Korrelation (oder Stabilität)

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

- Paralleltestkorrelation

  Es wird die Korrelation zwischen zwei Tests, die dieselbe Eigenschaft oder Fähigkeit mittels verschiedener Items erfassen sollen, berechnet. Mit der Paralleltestkorrelation wird erfasst, wie invariant die Testergebnisse gegenüber einer Variation der Itemauswahl und zeitlicher Variation sind. Sie ist somit ein Kennwert der Bedingungsstabilität.

Validität

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

Definition: Validität
Die Validität gibt an, ob der Test das auch wirklich misst, was er zu messen beansprucht.

- Inhaltsvalidität

  Von **Inhaltsvalidität** spricht man, wenn ein Test (bzw. seine Testitems im Gesamten) und auch jedes einzelne Item das zu messende Merkmal wirklich bzw. hinrei-

chend präzise erfasst. Präzise meint hier nicht den Aspekt der Messgenauigkeit, sondern präzise bezieht sich auf die Abbildung des Konstrukts durch das Item. Das Item muss präzise das angestrebte Konstrukt messen und eben kein weiteres Konstrukt bzw. einen Überschneidungsbereich mit einem anderen Konstrukt. Man geht dabei nach Michel und Conrad (1982) von einem Repräsentationsschluss aus. Das heißt, dass die Testitems eine repräsentative Itemmenge aus dem „Universum" von Items bilden, die das interessierende Merkmal abbilden. Die Gewährleistung einer hohen Inhaltsvalidität ist der wichtigste Schritt der Testkonstruktion.

Nach Michel und Conrad (1982, S. 57) wird die Inhaltsvalidität in der Regel nicht numerisch anhand eines Kennwerts, sondern „aufgrund logischer und fachlicher Überlegungen" bestimmt und „mit oder ohne Einschränkung akzeptiert oder verworfen". Die Autoren verweisen darauf, dass auch die Begriffe **logische Validität** oder **Augenscheinvalidität** (die eigentlich kein wissenschaftliches Konzept darstellen) mit der Inhaltsvalidität eng verbunden sind. Während die logische Validität in etwa der Inhaltsvalidität entspricht, wird unter Augenscheinvalidität verstanden, dass selbst ein Laie unmittelbar den Zusammenhang zwischen Testaufgaben und gemessenem Verhalten erkennt. Die Augenscheinvalidität kann jedoch kein ausreichendes Validitätskriterium für die Güte eines Tests darstellen.

Es ist in der Praxis sehr schwierig zu beurteilen, ob ein Test eine repräsentative Itemmenge enthält. Es gibt sehr viele Verhaltensweisen, die eine Fähigkeit kennzeichnen – bei breit gemessenen Fähigkeiten nahezu unendlich viele. Ob eine Auswahl in einem solchen Fall repräsentativ oder zumindest in irgendeiner Form geeignet ist, kann nur schwer entschieden werden. Dennoch sollte der Versuch unternommen werden, dies zu tun. Beispielsweise resultieren die Big 5 aus einem psycholexikalischen Ansatz. Das heißt, dass sich Persönlichkeitseigenschaften durch die Sprache ausdrücken. Dabei wurden ursprünglich über 10.000 Wörter faktorenanalytisch untersucht. Man sieht an diesem Beispiel, dass man sehr wohl versuchen kann, die Inhaltsvalidität eines Tests zu sichern. Ein anderes Beispiel ist die automatische Itemgenerierung bei Intelligenztests (Arendasy & Sommer, 2010). Hier werden Aufgabenmerkmale definiert und mithilfe von Regeln computergestützt ein Itemuniversum generiert. Selbst in einem Fall, in dem ein solches Vorgehen nicht möglich ist, muss zu diesem Punkt im Testhandbuch Stellung genommen werden.

Murphy und Davidshofer (2001) geben ein hilfreiches Vorgehen an, wie man Inhaltsvalidität erfassen kann (S. 150): (1) Beschreibung der Inhaltsebene des Konstrukts (Fähigkeit, Eigenschaft); (2) Festlegung, welcher Inhaltsbereich durch welches Item erfasst wird; (3) Vergleich der Teststruktur mit der Struktur des Konstrukts. Da viele Konstrukte nur ungenau definiert sind, ist gerade der erste Schritt nicht einfach. Man behilft sich hier mit Arbeitsdefinitionen oder betrachtet nur Teilausschnitte eines Konstrukts.

In der Praxis wird der Inhaltsvalidität oft nicht die Aufmerksamkeit geschenkt, die sie eigentlich verdient hat. Es gibt Argumente, man könnte die Inhaltsvalidität von Konstrukten gar nicht bestimmen, weil man das Itemuniversum nicht bestimmen könne. Zumindest Letzteres trifft oft zu. Das ist jedoch keine Entschuldigung, es erst gar nicht zu versuchen. Häufig krankt die Testentwicklung an der mäßigen Inhaltsvalidität der psychologischen Tests. Dies ist insofern verständlich, da eben die Sicherung der Inhaltsvalidität, im Vergleich zur Berechnung von Koeffizienten für die anderen Validitätsarten, relativ schwerfällt. Daher sind Tests meist das Ergebnis eines statistischen Homogenisierungsprozesses, der mit theoretischer Fundierung

nichts mehr zu tun hat. Guttman (1977) kritisierte dieses Vorgehen treffend: „To throw away items that do not ‚fit' unidimensionality is like throwing away evidence that the world is round." Mangelnde Überlegungen am Anfang des Konstruktionsprozesses führen schon in der Entwicklungsphase zu unzureichenden Verfahren.

- Kriteriumsvalidität

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

  – Vorhersagevalidität (prognostische Validität, prädiktive Validität). Es werden Zusammenhänge (Korrelationen) mit zeitlich später erhobenen Kriterien ermittelt. Beispielsweise wird die Intelligenztestleistung vor Beginn der Lehre ermittelt und mit der Abschlussnote der Ausbildung korreliert.
  – Übereinstimmungsvalidität *oder* konkurrente Validität. Korrelationen mit zeitlich (fast) gleich erhobenen Kriterien. Beispielsweise könnte die Konzentrationstestleistung vor Beginn einer Klausur ermittelt und dann die Korrelation mit der Klausurnote berechnet werden.
  – Retrospektive Validität. Es werden Zusammenhänge (Korrelationen) mit zeitlich vorher ermittelten Kriterien berechnet. Beispielsweise wird die Intelligenztestleistung während des Studiums erhoben und mit den Schulnoten des zurückliegenden Abiturs korreliert.
  – Inkrementelle Validität. Sie bezeichnet den Beitrag eines Tests zur Verbesserung der Vorhersage eines Kriteriums über einen anderen Test hinaus. Durch Intelligenztests lässt sich beispielsweise besonders gut Berufserfolg vorhersagen. Jede andere diagnostische Methode muss sich nun daran messen lassen, ob sie über die Intelligenz hinaus noch etwas zur Vorhersage von Berufserfolg beitragen kann. Eine der Methoden, die das leisten kann, ist beispielsweise das strukturierte Interview (Schmidt & Hunter, 1998). Zur Feststellung der inkrementellen Validität werden hierarchische Regressionsanalysen verwendet (vgl. Bühner & Ziegler, 2009, Kapitel 7.2, und zur Durchführung mit SPSS Kapitel 7.6).

- Konstruktvalidität

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

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

Innerhalb dieses Ansatzes kann also zwischen konvergenter Validität und diskriminanter oder synonym auch divergenter Validität unterschieden werden:

– Konvergente Validität

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

– Diskriminante/divergente Validität

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

– Faktorielle Validität

  Ebenfalls sehr häufig werden die Zusammenhänge zwischen verschiedenen Tests mit Faktorenanalysen untersucht. Die so genannte *faktorielle Validität* dient zum einen dazu, homogene konstruktnahe Inhaltsbereiche zusammenzufassen, und zum anderen, diese von konstruktfremden Bereichen zu trennen (siehe *Kapitel 6*). Zudem kann hierzu auch die Prüfung des vor der Testkonstruktion aufgestellten Testmodells mithilfe von konfirmatorischen Faktorenanalysen gezählt werden. Das bedeutet, die festgelegte Zugehörigkeit bestimmter Items zu bestimmten Konstruktbereichen oder Facetten kann ebenso getestet werden wie die Annahme unkorrelierter Messfehler. Wichtig ist hier, dass eine konfirmatorische Faktorenanalyse durchgeführt wird und nicht eine explorative. Schließlich wurde die Zugehörigkeit der Items zu einzelnen Facetten oder Faktoren bereits festgelegt und muss nicht mehr ermittelt werden.

Es gibt verschiedene Methoden, um konvergente und diskriminante Validität zu bestimmen. Die drei häufigsten seien hier kurz aufgeführt. An erster Stelle sind Korrelationen zu nennen. In Testhandbüchern werden häufig Korrelationen mit konstruktnahen und konstruktfremden Verfahren angegeben. Liegen viele Korrelationen vor, kann eine Darstellung schnell unübersichtlich werden. Ein Ansatz, der zur Systematisierung von Korrelationen vorgeschlagen wurde, ist der Multitrait-Multimethod-Ansatz von Campbell und Fiske (1959). Dieser ist im Beispielkasten dargestellt.