Grundlagen diagnostischer Verfahren

2.1 Voraussetzungen und theoretische Basis psychometrischer Tests

2.1.1 Anforderungen an einen Test

2.1.1.1 Was versteht man unter einem »Test«?

Testbegriff

Der Begriff »Test« ist schon lange in unsere Alltagssprache und unser Alltagsleben eingedrungen. Bevor wir einen neuen Staubsauger, ein Auto, ein Fernsehgerät oder vielleicht auch nur ein Haarwaschmittel kaufen, suchen wir nach einem Testbericht über dieses Produkt. Einige Zeitschriften befassen sich allgemein mit Verbrauchertests (*Test*, *Ökotest*, *Finanztest*), andere wie Auto-, Computer- oder Fotozeitschriften berichten immer wieder über einschlägige Tests. Banken werden einem »Stresstest« unterzogen, um ihre Funktionsfähigkeit unter widrigen Randbedingungen abzuschätzen. In der Apotheke kann man Tests kaufen, die eine Schwangerschaft, hohe Blutzuckerwerte oder Eiweiß im Urin erkennen. Und dann gibt es auch Tests, die meist von Psychologen entwickelt wurden. Sie dienen dazu, die Eignung für ein Studium, Persönlichkeitseigenschaften oder etwa die Intelligenz zu messen. Im Internet findet man psychologische »Tests«, die vielleicht nicht einmal diesen Namen verdienen. Deshalb ist es sinnvoll, erst einmal zu definieren, was man unter einem Test versteht. Anschließend lässt sich beurteilen, ob ein Produkt den Namen »Test« verdient. Wir werden auch feststellen, dass es diagnostische Verfahren gibt, die nicht als Test bezeichnet werden und dennoch die Definitionskriterien erfüllen.

In der Fachliteratur finden sich zahlreiche Definitionen, die auf den ersten Blick uneinheitlich erscheinen. Zunächst wird eine Auswahl von Definitionen vorgestellt, aus der dann wesentliche gemeinsame Definitionsmerkmale herausgearbeitet werden.

Definitionen von »Test«

Testdefinitionen in der Fachliteratur

- »A **test** is a standardized procedure for sampling behavior and describing it with categories or scores. In addition, most tests have norms or standards by which the results can be used to predict other, more important behaviors« (Gregory, 2004, S. 30).
- »A test may be defined simply as a measuring device or procedure. … the term **psychological test** refers to a device or procedure designed to measure variables related to psychology (for example, intelligence, personality, …) … A psychological test almost always involves analysis of a sample of behaviour. The behaviour sample could range from responses to a pencil-and-paper questionnaire to oral responses to questions to performance of same task. The behaviour sample could be elicited by the stimulus of the test itself, or it could be naturally occurring behaviour (under observation)« (Cohen & Swerdlik, 2010, S. 5).
- »Ein [psychologischer] Test ist ein wissenschaftliches Routineverfahren zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung« (Moosbrugger & Kelava, 2007, S. 2).
- »Ein psychometrischer Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung. Rost (2004) erweitert diese Definition mit dem Hinweis, dass es nicht immer um eine quantitative Aussage geht, sondern das Ziel eines Tests auch eine qualitative Aussage sein kann (z. B. Zuordnung von Personen zu bestimmten Kategorien)« (Bühner, 2006, S. 21). Der erste Teil der Definition erfolgt mit Verweis auf eine Definition im Lehrbuch von Lienert und Raatz (1998).
- »Ein *psychologisch-diagnostisches Verfahren* (vereinfacht oft »*Test*« genannt) erhebt unter standardisierten Bedingungen eine Informationsstichprobe über einen (oder mehrere) Menschen, indem systematisch erstellte Fragen/Aufgaben interessierende Verhaltensweisen oder psychische Vorgänge auslösen; Ziel ist es, die fragliche Merkmalsausprägung zu bestimmen« (Kubinger, 2009, S. 10).

Definitionsmerkmale eines Tests

Einige Definitionsmerkmale kristallisieren sich heraus, die mehrfach erwähnt werden und denen in den übrigen Definitionen meist nicht explizit oder implizit widersprochen wird. Diese können zu einer Definition zusammengeführt werden.

Definition

Bei einem **psychologischen Test**
(a) handelt es sich um eine Messmethode,
(b) mit der ein psychologisches Merkmal (oder auch mehrere Merkmale) erfasst werden soll(en).
(c) Das Vorgehen ist standardisiert
(d) und schließt die Erhebung einer Verhaltensstichprobe ein.
(e) Das Verhalten wird durch die spezifischen im Test realisierten Bedingungen hervorgerufen.
(f) Seine Variation soll weitgehend auf die Variation des zu messenden Merkmals zurückzuführen sein.
(g) Ziel ist eine quantitative (Ausprägung des Merkmals)
(h) und/oder eine qualitative Aussage (Vorhandensein oder Art des Merkmals) über das Merkmal.

Messgegenstand

Ad b. Der **Messgegenstand** wird pauschal als psychologisches Merkmal bezeichnet. Eine Einengung auf Persönlichkeitsmerkmale (einschließlich Intelligenz, Interessen, Motivation etc.) ist nicht zwingend. Auch wenn solche Eigenschaften häufig Gegenstand von Tests sind, kann ein Test auch emotionales Erleben (Emotionen, Gefühle etc.), Beziehungen zwischen Menschen (etwa die Qualität einer Paarbeziehung) oder situative Merkmale (z. B. belastende Faktoren am Arbeitsplatz) erfassen.

standardisiertes Vorgehen

Ad c. Das **standardisierte Vorgehen** ist ein wesentliches Merkmal aller Messmethoden. Die Bedingungen für die Durchführung müssen genau spezifiziert sein, ebenso die Auswertung und Interpretation der Antworten bzw. Ergebnisse (s. dazu die Ausführungen zur Objektivität ► Abschn. 2.3).

Verhaltensstichprobe

Ad d. Die Erhebung einer **Verhaltensstichprobe** impliziert, dass der Test Verhaltensweisen erfasst (z. B. Antworten auf Fragen). Da es sich um eine Stichprobe von Verhaltensweisen handelt, sind Prinzipien der Stichprobenziehung zu beachten (repräsentative, systematische oder auch zufällige Auswahl aus einem Universum von Verhaltensweisen).

Das Verhalten wird durch den Test hervorgerufen

Ad e. Das **Verhalten wird durch den Test hervorgerufen** (vgl. Definition 5). Damit wird gewährleistet, dass nicht unkontrollierbare situative Bedingungen das Verhalten determinieren. In der Regel wird das Verhalten durch eine präzise Instruktion (z. B. »Kreuzen Sie an, ob die Aussage auf Sie zutrifft oder nicht«; Zusatz »zügig arbeiten«, »ehrlich antworten«, »nicht zu lange nachdenken«) und zusätzliche Fragen bzw. Feststellungen (z. B. »Ich bin nicht so leicht aus der Ruhe zu bringen« – »stimmt« oder »stimmt nicht«) oder Aufgaben (z. B. »Streichen Sie alle d's mit zwei Strichen durch«) hervorgerufen. Eine systematische Beobachtung von Alltagsverhalten oder die Beurteilung von Merkmalen wie »Durchsetzungsfähigkeit« im Rollenspiel eines Assessment Centers sind demnach nicht als Test anzusehen. Selbst wenn in einem Assessment Center die Rollen der Teilnehmer durch Instruktionen genau festgelegt sind, werden die einzelnen Personen ihre Anweisungen unterschiedlich umsetzen, und sie werden zudem auf das Verhalten der anderen Teilnehmer reagieren. Diese Eigendynamik führt dazu, dass die situativen Bedingungen des Verhaltens eines Teilnehmers nicht die gleichen sind wie die bei einem anderen Teilnehmer. In einem hoch strukturierten Interview werden eventuell nur Fragen vorgelesen und die Antworten wörtlich protokolliert. Hier können alle Merkmale eines Tests erfüllt sein. Es ist dennoch nicht üblich, ein Interview als einen Test zu bezeichnen. Eine systematische Verhaltensbeobachtung in genau definierten Mini-Situationen (z. B. eine Spinne in einem verschlossenen Glas einen Meter vor den Patienten stellen, tote Spinne auf die Hand des Patienten legen – bei standardisiertem »Testmaterial«) kann ebenfalls als Test gelten. Man könnte hier von einem »Verhaltenstest« sprechen. Solche Verfahren kann man wie einen Test im engeren Sinne konstruieren, analysieren und auch bewerten.

Testverhalten

Ad f. Das **Testverhalten** soll die Ausprägung oder das Vorhandensein eines Merkmals anzeigen. Unter Verhalten im Test verstehen wir die Antworten, die ein Proband auf Fragen gibt, seine Reaktionszeit auf Reize oder etwa seine Lösung einer Aufgabe. Die auf Kurt Lewin zurückgehende Verhaltensgleichung »Verhalten ist eine Funktion von Person und Umwelt« macht deutlich, dass das Verhalten im Test nur dann als Indikator eines Personenmerkmals interpretiert werden darf, wenn die Situation (Umwelt) während der Testdurchführung konstant gehalten wird.

Quantifizierung

Ad g. Die **Quantifizierung** eines Merkmals bedeutet, dass die Ausprägung üblicherweise durch einen Normwert, zumindest aber durch einen Rohwert (also immer durch eine Zahl) ausgedrückt wird. Dass die Ausprägung zwecks Interpretation auch in Kategorien wie »durchschnittlich« oder »hochbegabt« übersetzt werden kann, schränkt die Forderung nach zahlenmäßiger Abbildung nicht ein.

qualitative Aussage

Ad h. Eine **qualitative Aussage** wird in den Definitionen 1 und 4 explizit vorgesehen. Manchmal wird nur ermittelt, ob jemand einer bestimmten Klasse oder Kategorie von Menschen zugerechnet werden kann. In der klinischen Diagnostik gelten genaue

Regeln, wann eine bestimmte psychische Störung zu diagnostizieren ist. Beispielsweise kann verlangt werden, dass Symptom A, B und C voll ausgeprägt vorliegen müssen und darüber hinaus noch zwei weitere von fünf Symptomen. Auf einen Zahlenwert zur Merkmalsausprägung wird verzichtet. Aus dem Gebot der Standardisierung folgt, dass solche Regeln eindeutig festgelegt sein müssen.

2.1.1.2 Arten von Tests

Messgegenstand als Einteilungskriterium

Alleine im deutschsprachigen Raum gibt es hunderte von psychologischen Tests. Um ein konkretes Testverfahren einordnen zu können und um gezielt Alternativen aufzufinden, ist eine Systematik der Tests hilfreich. Das wichtigste Kriterium für eine Einteilung von Tests ist der **Messgegenstand** (welches Merkmal soll erfasst werden?). Die Merkmale lassen sich nach Bereichen unterteilen, wobei sich eine hierarchische Ordnung anbietet, da sich **Leistungs- und Persönlichkeitsbereich** jeweils weiter untergliedern lassen. Beispielsweise bietet sich für den Persönlichkeitsbereich eine Differenzierung in allgemeine Persönlichkeitsmerkmale (Beispiel: Extraversion), klinisch relevante Persönlichkeitsmerkmale (Beispiel: Depressivität), Motive und Interessen an.

Konstruktionsprinzipien

Die Frage, wie die Verhaltensstichproben für einen Test gewonnen werden, führt zu den **Konstruktionsprinzipien** »induktiv«, »deduktiv«, und »external« (▶ Abschn. 2.2.2). Die Annahmen, wie und warum das Testverhalten Schlussfolgerungen auf das zu messende Merkmal zulässt, können unter dem Begriff »**theoretische Modellannahmen über die Entstehung von Testantworten**« eingeordnet werden. Bei Fragebögen wird meist angenommen, dass Menschen in der Lage sind, angemessene Selbstbeschreibungen abzugeben. Dazu gehört die Fähigkeit, sich selbst zu beobachten und das Beobachtete schließlich in die richtigen Worte zu fassen bzw. festzustellen, ob eine Aussage zur Selbstbeobachtung passt. Projektiven Verfahren liegt die Annahme zugrunde, dass mehrdeutiges Material in Abhängigkeit von Persönlichkeitsmerkmalen unterschiedlich interpretiert wird; Introspektionsfähigkeit und Selbstbeurteilung spielen keine Rolle.

Anwendungsbereiche und Zielgruppen

Viele Tests wurden für bestimmte **Anwendungsbereiche** entwickelt. Wichtige Anwendungsfelder, in denen Tests häufig eingesetzt werden, sind Berufseignungsdiagnostik, Klinische Psychologie, Neuropsychologie und Schul- und Erziehungsberatung. Für Anwender stellt oft die **Zielgruppe**, für die ein Test aufgrund seiner Aufgaben und seiner Normen geeignet ist, ein wichtiges Auswahlkriterium dar. Es liegen Tests für Kinder, Jugendliche und Erwachsene vor, wobei oftmals der Altersbereich noch genauer festgelegt bzw. eingeschränkt ist.

Einzel- oder Gruppenuntersuchung

Aus pragmatischer Sicht stellt sich manchmal die Frage, ob ein Test im **Einzelversuch** durchgeführt werden muss, oder ob auch **Gruppenuntersuchungen** möglich sind. Letzteres ist bei der Untersuchung vieler Probanden äußerst ökonomisch.

Papier-und-Bleistift-Tests vs. computergestützte Tests

Anwender haben manchmal eine Präferenz für **Papier-und-Bleistift-Tests** oder **computergestützte Tests**. Letztere haben den Vorteil, dass die Auswertung automatisch erfolgt. Sie setzen aber die Verfügbarkeit von Computerarbeitsplätzen und teilweise die Anschaffung von Basissoftware für ein Testsystem voraus. Weiterführende Informationen zu den unterschiedlichen Arten von Tests finden sich in ▶ Kapitel 3.

2.1.1.3 Grundvoraussetzungen für die Konstruktion eines Tests

Merkmal ist klar definiert

Merkmal ist hinreichend klar definiert und erforscht Nicht für alle Merkmale liegen Tests vor. Neben mangelnder Nachfrage kann dafür auch eine unbefriedigende Forschungslage verantwortlich sein: Was man messen möchte, ist konzeptuell noch nicht hinreichend präzisiert worden, und oft mangelt es an empirischer Forschung, die ein theoretisches Modell auch stützt. Solche Bedenken werden manchmal beiseite geschoben. Verschärft könnte man daher auch behaupten, dass es Tests gibt, die etwas messen (sollen), über das man kaum etwas weiß. Eine stark zugespitzte Bemerkung dazu lautet: »Sie wissen nicht, was es ist – aber messen können sie es.« Zur Entlastung von

Testautoren, die etwas nebulöse Merkmale per Test erfassen wollen, muss man einräumen, dass die Konstruktion und der gezielte Einsatz von Tests auch dazu beitragen können, ein Konstrukt zu präzisieren. Solche Tests sind vorerst ausschließlich für die Forschung geeignet!

Verhalten im Test indiziert das Merkmal Wie kommt man zu der Annahme, dass jemand, der weiß, in welcher Himmelsrichtung die Sonne aufgeht, intelligenter ist als andere? Oder warum soll jemand, der einen kurzen englischen Text liest und Fragen zum Inhalt richtig beantwortet, für ein Psychologiestudium geeignet sein? Oder warum soll jemand depressiv sein, der angibt, dass er unter Appetitmangel leidet? Sämtliche Beispiele stammen aus aktuellen diagnostischen Verfahren.

Testverhalten indiziert Merkmal

Man könnte argumentieren, dass es völlig genügt, empirisch einen Zusammenhang zwischen der Antwort im Test und dem Merkmal nachzuweisen. Tatsächlich begegnen wir dieser Argumentation bei external konstruierten Tests (▶ Abschn. 2.2.2.2). Meist liegen einem Test aber bestimmte Annahmen oder Modelle zugrunde. Anhand der drei oben genannten Itembeispiele soll dies erläutert werden.

Intelligenzmodell von Cattell

Ein Intelligenzmodell, das auf den amerikanischen Chemiker und Psychologieprofessor Cattell zurückgeht, besagt folgendes: Menschen haben eine unterschiedlich stark ausgeprägte Fähigkeit, gut (schnell und richtig) zu denken. Diese **fluide Intelligenz** genannte Fähigkeit führt dazu, dass man sich in der Schule – und generell im Leben – effizient Wissen aneignen kann, sofern hinreichend Lernmöglichkeiten bestehen. Als Resultat entsteht **kristalline Intelligenz**, was nichts anderes als Wissen bedeutet. Folglich ist es angebracht, Wissensfragen (z. B. »In welcher Himmelsrichtung geht die Sonne auf?«) zu stellen, um die kristalline Intelligenz zu messen. Da die kristalline Intelligenz eine wichtige Komponente der allgemeinen Intelligenz ist (Carroll, 1996), kann man solche Fragen (in Kombination mit anderen) auch einsetzen, um die allgemeine Intelligenz zu messen.

Anforderungsanalyse

Will man die Eignung für ein bestimmtes Studium messen, beginnt man mit einer **Anforderungsanalyse**. Man versucht also herauszufinden, welche Voraussetzungen jemand beispielsweise für ein Psychologiestudium mitbringen sollte. Da an den meisten Universitäten englischsprachige Literatur zu lesen ist, sollen die Studierenden diese Texte sinnverstehend lesen können. Deshalb ist eine Aufgabe, die sinnverstehendes Lesen an einem Text prüft, grundsätzlich für einen Studierfähigkeitstest im Fach Psychologie geeignet.

Symptome kennzeichnend für Störung

Dem dritten Beispiel liegt eine Konvention zugrunde. Experten haben sich darauf geeinigt, welche und wie viele Symptome vorliegen müssen, damit man von einer bestimmten Störung sprechen kann (z. B. ICD-10; Weltgesundheitsorganisation et al., 2006). Testautoren greifen deshalb oft genau die Symptome auf, die als kennzeichnend für eine Störung gelten. Ein Depressionsfragebogen kann daher Fragen enthalten, die das Vorliegen von trauriger Stimmung, Pessimismus, mangelndem Appetit oder etwa Schuldgefühlen prüfen sollen.

An das Formulieren von Items (Aufgaben, Fragen) werden also weitaus höhere Anforderungen gestellt als nur Einfallsreichtum. Viele Nichtpsychologen trauen sich zu, einen Fragebogen zu »machen«. Wie gezeigt wurde, braucht man jedoch fundiertes Wissen über die Merkmale, die man erfassen will – und einiges mehr: Das Thema Itemkonstruktion wird in ▶ Abschnitt 2.2 vertieft.

2.1.2 Die klassische Testtheorie (KTT)

2.1.2.1 Annahmen der KTT

Von wenigen Ausnahmen abgesehen sind die heute gebräuchlichen Tests nach den Regeln der sog. klassischen Testtheorie (KTT) konzipiert worden. Gulliksen (1950)

hat frühere Forschungsarbeiten, darunter auch Arbeiten von Spearman aus den Jahren 1904 bis 1913, zusammengefasst und aufgearbeitet. Eine mathematisch fundierte Fassung haben Lord und Novick (1968) vorgelegt. Dieses Buch gilt als Grundlage der KTT (vgl. Krauth, 1996).

Annahme: Testwerte sind fehlerbehaftet

Die KTT ist eine Reliabilitätstheorie, liefert also eine theoretische Begründung der Reliabilität (Messgenauigkeit) eines Tests. Eine grundlegende Annahme ist, dass Testwerte, also die Ergebnisse, die uns Persönlichkeitsfragebögen, Intelligenztests, Konzentrationstests etc. liefern, fehlerbehaftet sind.

beobachteter Wert, wahrer Wert und Messfehler

Wenn jemand in einem Intelligenztest einen IQ von 131 erreicht, muss er nicht unbedingt hochbegabt sein; Hochbegabung ist definiert als IQ über 130. Der IQ von 131 ist nur der **beobachtete Wert**, der **wahre Wert** der Person kann tatsächlich niedriger, aber auch noch höher sein. Diese Abweichung kommt durch **Messfehler** zustande. Wir stellen uns vor, dass der Intelligenztest wiederholt würde, ohne dass Erinnerungs- und Übungseffekte auftreten. Der beobachtete IQ wäre nun 125. Da sich die Intelligenz der Person nicht verändert hat (ihr wahrer Wert ist gleich geblieben), muss der Messfehler jedes Mal unterschiedlich groß gewesen sein. Damit sind auch schon die zentralen Begriffe »beobachteter Wert«, »wahrer Wert« und »Messfehler« eingeführt.

Annahme: Der Messfehler variiert von Messung zu Messung

Auch eine wichtige Annahme über das Wesen des Messfehlers wurde angedeutet: Der Messfehler variiert von Messung zu Messung. Akzeptiert man bestimmte Grundannahmen, lassen sich Formeln zur Schätzung der Messgenauigkeit (Reliabilität) eines Tests herleiten. Wir können damit die Reliabilität eines Tests berechnen und den Bereich bestimmen, in dem der wahre Wert einer Person (mit einer frei wählbaren Sicherheitswahrscheinlichkeit) liegt. Schon diese kurzen Vorbemerkungen weisen darauf hin, dass die KTT ein sehr nützliches Handwerkszeug darstellt.

Grundannahmen der klassischen Testtheorie (KTT)

Die KTT beginnt mit einigen wenigen Grundannahmen (Axiomen). Diese werden a priori angenommen und nicht etwa empirisch durch Untersuchungen begründet. Sie stellen die Grundlage für mathematische Ableitungen dar, die schließlich zu Formeln führen, mit denen wir beispielsweise die Messgenauigkeit eines Tests berechnen. Die Auffassungen, welche Aussagen grundlegende Definitionen, welche Zusatzannahmen und welche bereits Ableitungen darstellen, gehen in der Sekundärliteratur auseinander. Steyer und Eid (2001), denen sich Bühner (2010) anschließt, gehen von nur zwei Grundannahmen aus. Für das Verständnis der KTT, wie sie im Folgenden dargestellt wird, sind diese Unterscheidungen jedoch von nachrangiger Bedeutung.

Erläuterung zu den Symbolen und Abkürzungen

Für Kennwerte der Population werden in der Statistik griechische und für Kennwerte der Stichprobe lateinische Buchstaben verwendet. Einer besseren Lesbarkeit zuliebe bleiben wir bei den vertrauten lateinischen Buchstaben. Im Folgenden eine Übersicht über die Bedeutung der im Text verwendeten Abkürzungen:

Zeichen	Bedeutung	Anmerkung
T	True score = wahrer Wert	Äquivalent mit τ (tau, griech. Buchstabe für t)
E	Error score = Messfehler	Äquivalent mit ε (epsilon, griech. Buchstabe für e)
X	beobachteter Wert	

▼

Zeichen	Bedeutung	Anmerkung
Rel	Reliabilität	r_{tt} wird manchmal allgemein für Reliabilität benutzt, manchmal auch nur für Retest-Reliabilität.
Corr	Correlation = Korrelation	r wird häufig verwendet, indiziert aber streng genommen nur die Produkt-Moment-Korrelation. Für die Korrelation in der Population wird ρ (Rho, gesprochen Roh) verwendet.
Cov	Kovarianz	Gemeinsame Varianz zweier Variablen. Werden die beiden Variablen z. B. durch z-Transformation standardisiert, entspricht deren Kovarianz der Korrelation.
E(x)	Erwartungswert einer Variablen	Stochastischer Begriff; arithmetisches Mittel einer Variablen, das sich bei unendlich vielen Wiederholungen theoretisch ergibt.

Grundvoraussetzung

Eine Grundvoraussetzung für alle weiteren Schritte ist, dass die Testwerte angemessen variieren; die Varianz darf nicht null betragen, und sie darf nicht unendlich groß sein. Diese Voraussetzung dürfte normalerweise erfüllt sein (Krauth, 1996). Die Überlegungen, die nun über Tests und die Werte von Personen in diesen Tests angestellt werden, gelten nicht nur für »komplette« Tests. Sie sind auch gültig, wenn man einen Test in zwei Hälften aufteilt und einen halben, verkürzten Test betrachtet. Man kann sogar noch einen Schritt weiter gehen: Sie gelten auch für ein einzelnes Item.

Verhältnis wahrer Wert, beobachteter Wert, Messfehler

Jeder beobachtete Wert X_i einer Person i in einem Test setzt sich zusammen aus einem wahren Wert T_i dieser Person i und einem Fehlerwert E_i:

$$X_i = T_i + E_i$$

Fehlerwert oder Messfehler

Der Fehlerwert oder Messfehler wird somit als eine Größe angesehen, die sich beim Vorgang des Messens über den »eigentlichen« oder wahren Wert legt. Das Pluszeichen bedeutet *nicht*, dass der beobachtete Wert immer größer ist als der wahre Wert. Man muss sich lediglich vorstellen, dass der Messfehler positive und negative Werte annehmen kann. Dadurch weicht der beobachtete, durch die Testanwendung erhaltene Wert mehr oder weniger stark nach oben oder unten vom wahren Wert ab. Das Ergebnis in einem Test (der beobachtete Wert) darf also nicht als absolut *genaue* Messung angesehen werden.

wahrer Wert

Der wahre Wert einer Person im Test ist unveränderlich; er ist bei jeder Durchführung des Tests gleich groß – so die Annahme. Der Begriff »wahr« ist übrigens missverständlich. Damit ist nicht die wahre Ausprägung eines Merkmals gemeint, sondern nur die Ausprägung des Merkmals, wie sie mit diesem Test gemessen wird. Man stelle sich vor, für Forschungszwecke würden zwei Forschergruppen je einen Intelligenztest entwickeln, der eine extrem hohe Messgenauigkeit erreichen soll (was übrigens technisch möglich ist). Nun untersucht man eine Person mit diesen Tests und ist erstaunt, dass der eine Test einen IQ von 120 und der andere einen von 130 ergibt! Um ganz sicher zu sein, untersucht man nun 100 Personen. Die beiden Tests, so stellt man fest, korrelieren r = .60 miteinander (Intelligenztests korrelieren in dieser Größenordnung untereinander). Die Erklärung für dieses Phänomen ist einfach: Jeder Test liefert ein anderes Ergebnis, misst also eine etwas andere Art der Intelligenz. Die wahre Intelligenz einer Person wird man nie herausfinden, da es sie nicht gibt. Intelligenz ist ein Konstrukt, und ein Konstrukt kann man auf vielfältige Weise operationalisieren (mes-

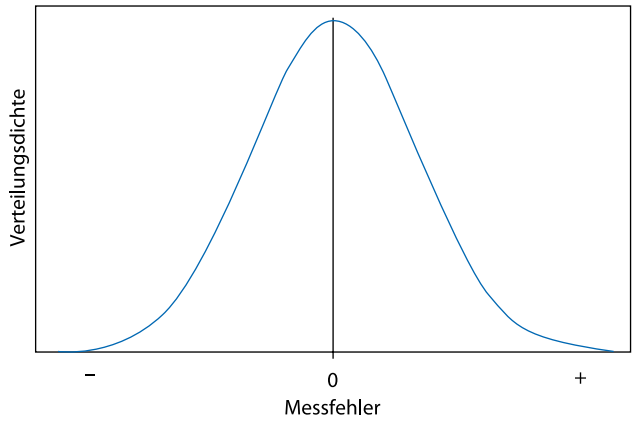

Abb. 2.1 Verteilung der Messfehler

sen). Der Zusammenhang zwischen unseren beiden Intelligenztests ist eine Frage der Validität dieser Tests!

wahrer Wert als Erwartungswert aller Messergebnisse

Für jede Person existiert in einem Test ein wahrer Wert Der wahre Wert einer Person ist konstant – zumindest über einen bestimmten Zeitraum. Er könnte theoretisch ermittelt werden, indem man den Test extrem (genau genommen unendlich) oft durchführt und dabei sicherstellt, dass keine Erinnerungs- und Übungseffekte auftreten. Der Mittelwert oder Erwartungswert aller Messergebnisse (also aller beobachteter Werte) wäre dann der wahre Wert:

$$T_i = E(X_i)$$

Der Erwartungswert des Messfehlers ist null

Der Erwartungswert des Messfehlers ist null Für jede Testperson i stellt der Messfehler E_i eine Zufallsvariable mit dem Erwartungswert (Mittelwert bei unendlich vielen Messungen) null dar (Abb. 2.1):

$$E(E_i) = 0$$

Messfehler als Gesamtheit aller unsystematischen Einflussgrößen

Theoretisch ergibt die Summe der Fehlerwerte einer Person bei unendlich häufiger Messwiederholung unter identischen Bedingungen null. Inhaltlich umfasst das Konzept des Messfehlers die Gesamtheit aller unsystematischen Einflussgrößen, die auf das Messergebnis einwirken können. *Unsystematisch* bedeutet, dass man nicht weiß, welche Fehlerquellen im konkreten Fall wie stark wirken und in welche Richtung. Die unten aufgelisteten Messfehler und viele andere mehr sind potenziell bei jeder Messung wirksam. Sie führen dazu, dass es bei einer Messung vielleicht zu einer leichten Abweichung vom wahren Wert nach unten kommt, bei einer anderen Messung zu einer starken Abweichung nach oben. Über unendlich viele Messungen hinweg gleichen sich die Messfehler aus, addieren sich zu null. Würde man (unendlich) viele Messungen an einer Person durchführen, könnte man die Messfehler völlig ignorieren. Der Mittelwert aller Messungen wäre identisch mit dem wahren Wert der Person in diesem Test.

Wie entstehen Messfehler? Grundsätzlich sind die Quellen der Fehlervarianz bekannt. Die Messfehler entstehen durch Fehler

- bei der Testkonstruktion,
- bei der Durchführung und
- bei der Auswertung des Tests.

Fehler bei der Testkonstruktion

Bei der **Testkonstruktion** besteht die Gefahr, Items aufzunehmen, die mehrdeutig sind, also von unterschiedlichen Testpersonen unterschiedlich interpretiert werden.

Ein Item wie »Ich ärgere mich gelegentlich über mich selbst« bietet gleich mehrfach die Gelegenheit für Interpretationen. Was bedeutet »sich ärgern«? Die Spanne reicht von leichter Verärgerung bis Wut. Wie oft muss man sich am Tag oder in der Woche ärgern, um von »gelegentlich« zu sprechen? Worauf soll sich der Ärger beziehen? Auf die ganze Person, auf ein Verhalten, auf Körperteile, auf die Kleidung etc.? Auch die Instruktion kann missverständlich sein. »Streichen Sie alle d′s mit zwei Strichen durch« wird normalerweise so verstanden, dass alle d′s, die mit zwei Strichen versehen sind (egal, ob oben oder unten), durchzustreichen sind. Es ist aber schon vorgekommen, dass eine Testperson alle d′s doppelt, also mit zwei Strichen, durchgestrichen hat.

Fehler bei der Durchführung

Bei der **Durchführung** eines Tests kann die *Testsituation* variieren: Lichtverhältnisse, Geräusche, Luftqualität, Raumtemperatur, Sitzkomfort, Art und Anzahl der anderen Testteilnehmer etc. sind nicht immer identisch, wenn der Test durchgeführt wird. Auch die *Testperson* selbst ist eine Quelle von Fehlervarianz: Die Motivation, ein gutes Ergebnis zu erzielen, die momentane geistige (Wachheit etc.) und emotionale Verfassung (Angst etc.) oder etwa pharmakologische Einflüsse (Einnahme von Medikamenten, Koffeingenuss etc.) können sich auf die Testleistung auswirken. Der *Testleiter* ist ebenfalls keine Konstante; Aussehen, Geschlecht, Alter, Kleidung, der Tonfall beim Vorlesen einer Instruktion, Gestik und Mimik etc. variieren und können einen Einfluss auf das Testergebnis haben.

Fehler bei der Auswertung

Die **Auswertung** kann bei Tests, die freie Antworten verlangen (dies ist etwa bei vielen projektiven Tests der Fall), nicht völlig standardisiert werden. Selbst beim Auflegen von Schablonen und dem Auszählen von Punkten sind Fehler möglich. Wenn anschließend in der Normtabelle für den Rohpunktwert der richtige Standardwert abgelesen wird, kann die falsche Tabelle aufgeschlagen oder beim Ablesen der Zahlen ein Fehler passieren.

Entstehung des Messfehlers

Diese Auflistung möglicher Fehler ist nicht vollständig. In ihrer Gesamtheit ergeben sie den Messfehler im Sinne der KTT. Die Auflistung macht plausibel, dass die Annahme von Messfehlern begründet ist. Sie hilft auch zu verstehen, warum manche Tests eine hohe und andere eine niedrige Messgenauigkeit aufweisen (je größer der Einfluss von Messfehlern auf das Testergebnis ist, desto geringer ist die Messgenauigkeit des Tests).

Die Messfehler sind unabhängig vom wahren Wert Die Fehlerwerte E_i sind unabhängig von den wahren Werten T_i der Person i im Test:

$$\text{Corr}\ (E_i, T_i) = 0$$

Der Messfehler ist unabhängig vom wahren Wert

Dass die Fehlerwerte unabhängig von den wahren Werten sind, bedeutet nichts anderes, als dass ein Test im unteren Bereich (niedrige Fähigkeit) ebenso genau misst wie im mittleren oder im oberen Bereich.

Die Messfehler zweier Tests A und B sind unkorreliert Die Messfehler in einem Test korrelieren nicht mit den Messfehlern in einem anderen Test. Zwischen den Fehlerwerten zweier Tests besteht eine Nullkorrelation:

$$\text{Corr}\ (E_A, E_B) = 0$$

Messfehler sind unkorreliert

Dieser Grundgedanke ist auch auf Testteile, bis hin zu den Items, übertragbar. Wenn die Fehlerwerte zweier Tests unkorreliert sind, wie hier angenommen wird, muss die Korrelation der beiden Testwerte alleine auf den wahren Zusammenhang der Merkmale zurückzuführen sein. Es sei daran erinnert, dass Messfehler *unsystematische* Fehler sind. Selbstverständlich kann die Korrelation zweier Tests durch systematische

http://www.springer.com/978-3-642-17000-3

Psychologische Diagnostik (Lehrbuch mit Online-Materialien)
Schmidt-Atzert, L.; Amelang, M.
2012, XVI, 624 S. 118 Abb. Mit Online-Extras., Hardcover
ISBN: 978-3-642-17000-3