

3.4 Die Untersuchungskriterien



Lerncoach

Im folgenden Abschnitt wird erläutert, was psychologische Tests von „Tests“ abhebt, die Sie häufig in Illustrierten (unter Titeln wie „Sind Sie eine gute Freundin? Testen Sie hier!“) finden. Sie müssen nicht in allen Einzelheiten wissen, wie ein Test entwickelt wird, sollten sich aber die wichtigsten Merkmale eines wissenschaftlichen Tests und vor allem die Definition der Gütekriterien einprägen.

3.4.1 Der Überblick

Die Erstellung eines wissenschaftlich fundierten psychologischen Tests beginnt mit einer sorgfältigen Auswahl und Analyse von Items, an deren Ende die Testendform steht (Testkonstruktion). Die anschließende Eichung des Tests an einer repräsentativen Stichprobe wird Testnormierung genannt. Wie „gut“ ein Test ist, wird anhand dreier Kriterien – der Testgütekriterien Objektivität, Reliabilität und Validität – bestimmt. Auch für diagnostische Entscheidungsstrategien gibt es Beurteilungskriterien, die am Ende dieses Kapitels erläutert werden.

3.4.2 Die Testkonstruktion

Unter einem **psychologischen Test** (z. B. Intelligenztest) wird ein Verfahren verstanden, mit dem quantitative Aussagen über den Ausprägungsgrad individueller Merkmale (Leistungs- oder Persönlichkeitsmerkmale) gemacht werden können.

Die **Konstruktion eines Tests** ist ein aufwändiges Verfahren, das mit der Auswahl von Items (= einzelne Testaufgaben) beginnt, die das zu messende hypothetische Konstrukt abbilden. Mit einer anschließenden **Itemanalyse** auf der Basis von Daten, die an einer ersten Stichprobe gewonnen werden, wird anhand statistischer Kennwerte entschieden, welche Items in die Testendform aufgenommen werden (**Itemselektion**). Zu den statistischen Kennwerten zählen zum Beispiel der **Schwierigkeitsindex**, der die Lösungswahrscheinlichkeit eines Items angibt, und der **Trennschärfekoeffizient**, aus dem abzulesen ist, wie gut ein Item Probanden mit hoher und niedriger Merkmalsausprägung von-

einander trennen kann. Die Testendform wird anschließend normiert und der Test auf seine Gütekriterien überprüft.

3.4.3 Die Testnormierung



Für viele Studenten wirken Testnormierungen und vor allem der Umgang mit den dabei notwendigen Zahlenwerten abschreckend. Versuchen Sie dennoch, die Absicht, die hinter der Normierung eines Tests steht, zu verstehen. Zudem ist es gut, einige der in Abb. 3.1 angeführten Standardwerte auswendig zu kennen (z. B. IQ).

Eine **Testnormierung** ist die **Eichung** eines Tests an einer repräsentativen Stichprobe. Um eine solche Eichung vorzunehmen, werden Testdaten an einer möglichst großen Stichprobe unter standardisierten Bedingungen erhoben. Aus diesen werden **Normen** gewonnen, zu denen die individuellen Testergebnisse in Beziehung gesetzt werden können. Erst wenn der durchschnittlich erzielte Wert – der **Mittelwert** (s. u.) – und ein Maß für die Streuung der Testwerte – die **Standardabweichung** (s. u.) – bekannt sind, kann beurteilt werden, was ein einzelnes Testergebnis eigentlich bedeutet.

Merke

Die Normierung eines Tests schafft ein Bezugssystem, in das individuelle Testergebnisse eingeordnet werden können und diese miteinander vergleichbar macht.

Selbstverständlich sollte die Stichprobe, anhand derer die Normen gewonnen werden, den Personen, für die der Test bestimmt ist, so ähnlich wie möglich sein. Für psychologische Tests liegen häufig gesonderte Normtabellen zum Beispiel für Männer und Frauen und für unterschiedliche Altersklassen vor.

Um verschiedene Tests miteinander vergleichen zu können, müssen **Normskalen** entwickelt werden. Sie standardisieren die Vergleichsmaßstäbe (Normen) von Tests. Einige seien im Folgenden genannt.

Die Äquivalenznormen

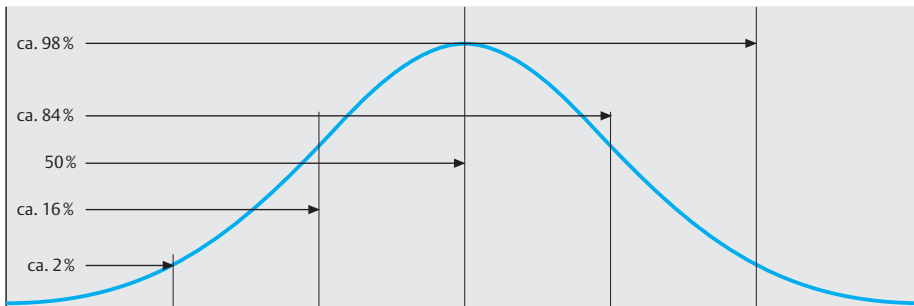
Bei **Äquivalenznormen** wird ein Rohwert (Testwert, der sich direkt aus dem Test ergibt) einer Altersgruppe zugeordnet, für die er besonders typisch ist. Es wird sozusagen ein Altersäquivalent zur individuellen Leistung geschaffen.

Ein Beispiel sind die Staffeltests zur Intelligenzmessung von Binet (19. Jh.): Das Lösen von Aufgaben, die typisch für eine bestimmte Altersklasse sind, bestimmt das Intelligenzalter (IA), das dem tatsächlichen Lebensalter (LA) entsprechen oder über oder unter ihm liegen kann. (vgl. S. 107).

Die Abweichungsnormen

Bei Abweichungsnormen (= Variabilitätsnormen) geschieht der Vergleich der Testwerte anhand der Abweichungen vom Mittelwert der Referenzstichprobe. Damit dies klarer wird, muss zunächst eine Verteilungsform erläutert werden, der die Messwerte entsprechen müssen:

Wenn Daten einer biologischen (z.B. Körpergröße) oder psychologischen Variablen (z.B. Intelligenz) an einer großen Anzahl von Probanden gesammelt werden, ähnelt die Häufigkeitsverteilung der erhobenen Messwerte der einer **Normalverteilung** (Abb. 3.1). Auf der Abszisse ist die Ausprägung des zu messenden Merkmals und auf der Ordinate die Anzahl der Probanden abgetragen. Der höchste Punkt der Verteilung kennzeichnet also den Ausprägungsgrad eines Merkmals, den die meisten Probanden erzielen. Er ist bei dieser glockenförmigen Verteilung zugleich Mittelwert (M), Median und Modalwert. Die Standardabweichungen (SD) markieren die Wendepunkte der Verteilung und kennzeichnen Abschnitte, für die bekannt ist, wie viel Prozent der Fälle in sie entfallen. So liegen beispielsweise in dem Bereich zwischen dem Mittelwert und einer Standardabweichung ca. 34% der Fälle. Da die Verteilung symmetrisch ist, liegen entsprechend im Bereich zwischen einer Standard-



Fälle pro Intervall [in %] 2.14 13.59 34.13 34.13 13.59 2.14

Differenz vom Mittelwert (M) in Standardabweichungen (SD) -2 SD -1 SD M 1 SD 2 SD

Test-normen	M	SD						Test-Bsp.
z-Werte	0	1	-2	-1	0	1	2	
Stanine	5	2	1	3	5	7	9	FPI
T-Werte	50	10	30	40	50	60	70	MMPI
IQ	100	15	70	85	100	115	130	HAWIE HAWIK
Z-Werte	100	10	80	90	100	110	120	IST

Abb. 3.1 Normalverteilung und einige Testnormen

abweichung über und unter dem Mittelwert 68 % der Stichprobe.

Die Verteilungen von Variablen wie Körpergröße oder Intelligenz sind je nach Maßstab auf der Abszisse unterschiedlich breit (z. B. cm, Testpunktwerte). Um dasselbe Aussehen der Verteilung zu erlangen, können die Daten **z-transformiert** werden, also in eine **Standardnorm** umgewandelt (= standardisiert) werden. Jeder z-Wert macht eine eindeutige Aussage darüber, wie weit der dazugehörige Rohwert vom Mittelwert entfernt ist. Die Verteilung, die entsteht, heißt **Standardnormalverteilung** und wird mit einem Mittelwert von 0 und einer Standardabweichung von 1 hinreichend beschrieben.

Abb. 3.1 zeigt einige typische Testnormen. Sie lassen sich ineinander überführen, sodass auch ein Vergleich von Testergebnissen verschiedener Tests möglich wird. Ein Beispiel: Ein IQ im Hamburg-Wechsler-Intelligenztest (HAWIE) von 130 und ein Wert von 120 im Intelligenz-Struktur-Test (IST) bedeuten beide, dass ein Wert erreicht wurde, der zwei Standardabweichungen über dem Mittelwert liegt und dass nur ca. 2 % der Bezugsgruppe einen höheren Wert erzielten (oder: 98 % erzielten einen geringeren Wert). Über einige in der Tabelle angeführten Testbeispiele erfahren Sie auf S. 110 mehr.

Die Prozentränge

Prozentränge sind ebenfalls eine Form der Normierung, bei der den Testergebnissen die relative Position auf der Messwertskala der Referenzgruppe zugeordnet werden. Ein Prozentrang sagt aus, wieviel Prozent der Referenzgruppe unterhalb des ermittelten Testwertes liegen. Ein Prozentrang von 90 bedeutet also, dass 90 % der Referenzstichprobe einen niedrigeren Wert erreicht haben. Das Bilden von Prozenträngen stellt keine Voraussetzung an die Verteilungsform der Messwerte, sie müssen also nicht normalverteilt sein.

3.4.4 Die Testgütekriterien

Ein psychologischer Test muss gewisse Qualitätsmerkmale aufweisen, um als gut zu gelten. Die Hauptgütekriterien sind die **Objektivität**, die **Reliabilität** und die **Validität**. Im weiteren Sinne können auch Ökonomie und Änderungssensitivität als Güte Merkmale verstanden werden.

Die Objektivität

Objektivität meint die **Unabhängigkeit des Tests von der Person des Testleiters**. Die Objektivität kann folglich erhöht werden, wenn der Einfluss des Testleiters auf die Durchführung, Auswertung und Interpretation des Tests minimiert wird. Dies kann geschehen, indem der Test standardisiert vorgegeben wird, also zum Beispiel schriftliche Testinstruktionen anstelle mündlicher Erläuterungen gewählt werden und indem der Spielraum bei der Testauswertung begrenzt wird (z. B. Schablone, Computerprogramm).

Um das Ausmaß der Objektivität zu bestimmen, wird überprüft, inwieweit die Ergebnisse, die unterschiedliche Versuchsleiter bei der Auswertung desselben Tests ermitteln, übereinstimmen. Ist die Übereinstimmung gering, ist dies ein Hinweis darauf, dass die Testergebnisse der Versuchspersonen nicht unabhängig vom Einfluss der Versuchsleiter sind. Die Höhe der Übereinstimmung wird durch den **Korrelationskoeffizienten** ausgedrückt, auf den bei der Datenauswertung noch ausführlich eingegangen wird.

Die Reliabilität

Die verschiedenen Formen der Reliabilität

Die Reliabilität oder **Zuverlässigkeit** meint die **Messgenauigkeit** eines Tests. Wenn ein Test wiederholt bei derselben Versuchsperson unter gleichen Bedingungen angewendet wird und die Ergebnisse identisch oder zumindest sehr ähnlich sind, misst der Test präzise. Es gibt unterschiedliche Möglichkeiten, die Reliabilität eines Tests zu überprüfen:

- Die **Retest-Reliabilität** (Testwiederholungs-Reliabilität) wird ermittelt, indem ein Merkmal (z. B. Angst, Intelligenz) mit demselben Test an denselben Probanden wiederholt gemessen wird. Je höher die Messergebnisse übereinstimmen, desto reliabler ist der Test.
- Bei der Bestimmung der **Paralleltest-Reliabilität** wird nicht derselbe Test wiederholt, sondern parallele Formen eines Tests (Version A und B) eingesetzt. Dies schließt zum Beispiel Erinnerungseffekte aus.
- Bei einer **Konsistenzanalyse** wird der Test nur einmal durchgeführt. Zur Bestimmung der Konsistenz wird der Test entweder in zwei Hälften

geteilt (z. B. nach geraden und ungeraden Items) und die Testhälften miteinander verglichen – hier spricht man von **Testhalbierungsreliabilität („Split-half“-Reliabilität)** – oder jede einzelne Testaufgabe wird mit allen übrigen in Beziehung gesetzt (**innere Konsistenz**) und so ein Maß für die Homogenität (Gleichartigkeit) der Testteile gewonnen.

Die Reliabilität eines Tests kann durch eine **Testverlängerung** erhöht werden. Mit zunehmender Zahl an Items zu demselben Merkmal findet eine Art Fehlerausgleich statt und der Test wird messgenauer (in der klassischen Testtheorie erklärt sich dies durch das stärkere Anwachsen der wahren Varianz im Vergleich zur Fehlervarianz; Bortz, 1999).

Der Standardmessfehler

Die Messgenauigkeit psychologischer Tests ist nie maximal, sodass eine gewisse Unzuverlässigkeit in Kauf genommen werden muss. Der Messfehler, der durch die **mangelnde Reliabilität** eines Tests zustande kommt, wird als Standardmessfehler (SM) bezeichnet. Zur Bestimmung des Standardmessfehlers wird ein Maß für die Messgenauigkeit (der **Reliabilitätskoeffizient** r) und ein Maß für die Streuung der Testwerte (**Standardabweichung** SD) berücksichtigt ($SM = SD \sqrt{1-r}$).

Jeder individuelle Wert, der mit dem Test erhoben wird, ist also mit einem Fehler behaftet. Rechnet man zu dem Testwert eines Probanden einen Bereich hinzu, der vom Ausmaß des Standardmessfehlers abhängt, ergibt sich ein **Konfidenzintervall** (Vertrauensintervall), in dem der „wahre“ (also fehlerfreie) Wert sehr wahrscheinlich liegt. Je reliabler der Test, desto geringer ist der Standardmessfehler und desto enger das Konfidenzintervall.

Die Validität

Die Validität (**Gültigkeit**) ist das dritte Hauptgütekriterium. Sie gibt an, ob der Test das Merkmal, das er zu messen vorgibt, auch tatsächlich misst. Ein valider Test sollte Probanden mit hohen Merkmalsausprägungen von denen mit niedrigen trennen können.

Ein Beispiel: Ein Angsttest ist dann valide, wenn er Angst misst und nicht etwa ein anderes Merkmal, wie zum Beispiel Schüchternheit. Trifft dies zu,

dann müssen Ängstliche und Nichtängstliche auch unterschiedliche Testwerte erhalten. Zur Bestimmung der Validität bestehen mehrere Möglichkeiten:

- Bei der **Kriteriumsvalidität** wird der Test mit einem oder mehreren **Außenkriterien verglichen** (Korrelation mit Außenkriterien, s. u.). Ein solches Kriterium könnte zum Beispiel die Einschätzung eines langjährig erfahrenen Therapeuten sein oder aber ein anderer Angstfragebogen, der sich bereits bewährt hat.
- Werden das Testergebnis und das Außenkriterium zur gleichen Zeit erhoben, spricht man von **Übereinstimmungsvalidität**.
- Soll das Testergebnis das Kriterium zu einem späteren Zeitpunkt vorhersagen, spricht man von **Vorhersagevalidität** (prädiktiver Validität). Ein Berufseignungstest beispielsweise ist dann (vohersage)valide, wenn er den späteren Berufserfolg gut vorhersagen kann.

Häufig gibt es nicht ein einzelnes Kriterium für ein komplexes hypothetisches Konstrukt. Um die **Konstruktvalidität** zu überprüfen, wird das Ausmaß bestimmt, wie eng der Test mit anderen gültigen Indikatoren des Konstrukts zusammenhängt.

Wenn die Testaufgaben selbst das zu messende Merkmal repräsentieren, spricht man von **Inhaltsvalidität**. Beispielsweise ist ein Rechentest ein inhaltsvalider Test, wenn es um die Erfassung von Rechenfähigkeit geht.

Es wird weiterhin in **interne** und **externe Validität** unterschieden. Eine Untersuchung ist dann intern valide, wenn die erzielten Ergebnisse eindeutig für (oder gegen) die Hypothese sprechen, alternative Erklärungen für deren Zustandekommen also ausgeschlossen werden können. Externe Validität meint, dass die Ergebnisse auch für andere vergleichbare Probandengruppen, Orte und Situationen gültig sind.

Der Zusammenhang der Testgütekriterien

Nach der klassischen Testtheorie sind die Testgütekriterien voneinander abhängig: Hohe Objektivität ist die Voraussetzung für hohe Reliabilität und diese ist wiederum die Voraussetzung für hohe Validität. Ein Test kann also nicht valider sein als er reliabel oder objektiv ist.

Die Ökonomie und die Änderungssensitivität Ob ein Test eingesetzt wird oder nicht, hängt neben den Hauptgütekriterien auch davon ab, ob er **ökonomisch** durchzuführen, auszuwerten und zu interpretieren ist. Der Aufwand, zum Beispiel hinsichtlich Zeit, Kosten oder Arbeitskraft, sollte zum Nutzen, der durch das Testergebnis erzielt wird, in angemessener Beziehung stehen.

Ein Test ist dann **änderungssensitiv**, wenn er sensibel gegenüber Veränderungen eines Merkmals ist. Ein Beispiel ist die Messung von Unterschieden der Angst vor und nach einer Angsttherapie, um deren Wirksamkeit zu überprüfen. Hierbei besteht das methodische Problem, nur schwer unterscheiden zu können, ob unterschiedliche Testergebnisse tatsächlich die Veränderungen des Merkmals wiedergeben oder aber die Folge der mangelnden Reliabilität des Tests sind (und das Merkmal in Wirklichkeit stabil bleibt).

Die Gütekriterien einer Entscheidungsstrategie Psychologische Tests sollen Aussagen über den tatsächlichen Zustand von Individuen machen. Allgemeiner gesprochen soll mit ihrer Hilfe eine Entscheidung getroffen werden, welcher **Merkmalsklasse** ein Individuum zugeordnet werden kann. Nichts anderes ist bei medizinischen Diagnosen der Fall: in der klinischen Praxis stehen **diagnostische Entscheidungen** an, bei denen mindestens zwei Alternativen (z. B. krank/gesund) zur Auswahl

stehen. Egal, ob eine Entscheidung durch das Urteil eines Experten, aufgrund eines Tests oder anhand anderer Entscheidungsstrategien getroffen wird – sie ist immer mit Risiko behaftet, da die Informationen, die zur Entscheidung herangezogen werden können, begrenzt sind. Mit der **Entscheidungstheorie** wird versucht, häufig intuitiv getroffene Entscheidungen explizit und durchschaubar zu machen und diagnostische Strategien hinsichtlich ihres Nutzens zu beurteilen.

Um einige weitere Begriffe zu erläutern, wird das einfache Beispiel herangezogen, dass sich die zu treffende Entscheidung auf die beiden Zustände krank/gesund bezieht. Die Kombinationen der Diagnosen und der tatsächlichen Zustände können in einem **Vier-Felder-Schema** angeordnet werden (**Abb. 3.2**). Anhand dieses Schemas lassen sich Kennwerte zur Güte der Diagnosestrategie (allgemeiner: der Entscheidungsstrategie) berechnen: die Sensitivität, die Spezifität und der positive und negative Prädiktionswert.

Mit „positiv“ wird das Vorhandensein eines kritischen Merkmals – in diesem Fall das Vorliegen einer Krankheit – bezeichnet, „negativ“ meint das Nichtvorhandensein.

Die **Sensitivität** ist die Wahrscheinlichkeit, mit der ein bestehender *positiver* Zustand auch tatsächlich erkannt wird. Sie errechnet sich aus dem Anteil der richtig als krank Klassifizierten an den Kranken insgesamt.

Diagnose	tatsächlicher Zustand		insgesamt
	positiv (krank)	negativ (gesund)	
positiv (krank)	Entscheidung richtig positiv A	Entscheidung falsch positiv B	positiver Prädiktionswert A + B A / (A + B)
negativ (gesund)	Entscheidung falsch negativ C	Entscheidung richtig negativ D	negativer Prädiktionswert C + D D / (C + D)
insgesamt	A + C Sensitivität A / (A + C)	B + D Spezifität D / (B + D)	

Abb. 3.2 Vier-Felder-Schema der Entscheidungsmöglichkeiten

Die **Spezifität** ist die Wahrscheinlichkeit, mit der ein bestehender *negativer* Zustand erkannt wird. Die richtig als gesund Klassifizierten werden zu allen Gesunden in Beziehung gesetzt.

Die Prädiktionswerte werden über die Zeilen, also aus der Sicht der Diagnosestrategie errechnet. Der **positive Prädiktionswert** meint die Wahrscheinlichkeit, mit der eine positive Diagnose zutreffend ist. Er errechnet sich aus dem Anteil der mittels Diagnose richtig als krank Klassifizierten an den als krank Klassifizierten insgesamt.

Der **negative Prädiktionswert** meint die Wahrscheinlichkeit, mit der eine negative Diagnose zutreffend ist. Er bestimmt sich durch den Anteil der mittels Diagnose richtig als gesund Klassifizierten an den als gesund Klassifizierten insgesamt.

3.4.5 Klinische Bezüge

Messung otoakustischer Emissionen

Ein Screeningtest (auch Filtertest) wird in einer größeren Bevölkerungsgruppe eingesetzt und dient dazu, das Vorliegen einer Erkrankung im Frühstadium zu erkennen. Ein Beispiel ist die Messung der otoakustischen Emission als Screeningmethode, um Hörstörungen bei Säuglingen zu identifizieren. Hierbei werden Schallwellen gemessen, die in der Cochlea entstehen und über das Mittelohr in den Gehörgang abgestrahlt werden.

Ein solcher Screeningtest sollte sich durch eine hohe Sensitivität kennzeichnen, um sicherzugehen, dass Kranke auf jeden Fall als solche erkannt werden, damit ihnen rechtzeitig Unterstützung und Förderung zukommen kann. Gleichzeitig sollte er eine hohe Spezifität besitzen, d.h. Gesunde auch als gesund erkennen, damit sie nicht fälschlicherweise behandelt und zu Unrecht verunsichert werden. Den Screeningtests folgen spezifische Nachweisverfahren, um die Verdachtsdiagnose zu überprüfen und sie gegebenenfalls noch zu präzisieren.



Check-up

- ✓ **Rekapitulieren Sie, wozu die Normierung eines Tests dient und wie sie erfolgt.**
- ✓ **Wiederholen Sie die Hauptgütekriterien eines Tests.**
- ✓ **Stellen Sie sich vor, ein Lehrer würfelt die Noten seiner Schüler aus. Ist sein „Messin-**

strument“ (das Würfeln) objektiv? Achten Sie beim Beantworten der Frage genau auf die Definition von „Objektivität“. Ist das Messinstrument reliabel oder sogar valide?

3.5 Die Untersuchungsplanung



Lerncoach

Im folgenden Kapitel werden Sie verschiedene Untersuchungsarten kennenlernen. Bei der Planung einer Untersuchung muss man sich für eine dieser Untersuchungsarten entscheiden und den Ablauf genau festlegen. Die wohl wichtigste Art einer wissenschaftlichen Untersuchung ist das Experiment. Versuchen Sie das Prinzip genau zu verstehen, dann erklären sich viele der anderen Begriffe von selbst.

3.5.1 Der Überblick

Wissenschaftliche Fragestellungen können auf unterschiedliche Art und Weise untersucht werden. In diesem Abschnitt werden einige Untersuchungsarten, wie die Feldstudie, die Längs- und die Querschnittsstudie erläutert. Die bedeutendste psychologische Untersuchungsmethode ist das Experiment, das in diesem Abschnitt detaillierter dargestellt wird. Das Experiment deckt Ursache-Wirkungs-Beziehungen auf, indem die Ursache vom Versuchsleiter variiert wird und die Auswirkungen der Variationen (Versuchsbedingungen) registriert und verglichen werden. Ist eine für die Fragestellung geeignete Untersuchungsart ausgewählt, muss eine repräsentative Stichprobe gewonnen werden. Hierzu macht man sich entweder die zufällige Gleichverteilung von Merkmalen in einer großen Stichprobe zunutze (Zufallsstichprobe) oder wählt die Probanden gezielt aus (Quotenstichprobe).

3.5.2 Die Arten von Untersuchungen

Das Experiment

Die abhängige und die unabhängige Variable

Das Experiment ist die Methode um **Ursache-Wirkungs-Beziehungen** aufzudecken. Um die Kausalität einer Beziehung zu überprüfen, muss die vermutete Ursache manipuliert werden und die Auswirkungen dieser Manipulation betrachtet werden.