2 Methodische Grundlagen der Psychologie

2.1 Untersuchungsplanung

Die Psychologie will menschliches Erleben und Verhalten beschreiben, erklären und vorhersagen können. Dazu bedient sie sich naturwissenschaftlicher, empirischer Methoden.

2.1.1 Hypothesenbildung

Hypothesen sind klar definierte Annahmen über Mittelwertsunterschiede (z. B. die Patientengruppe mit Medikament A ist im Mittel weniger depressiv als die Gruppe mit Medikament B) oder über Zusammenhänge von Variablen (z. B. je höher die Dosis von Medikament A ist, desto besser ist dessen antidepressive Wirkung). Will man in der Psychologie oder Soziologie etwas untersuchen, so stellt man als Erstes eine Hypothese auf, die dann wissenschaftlich geprüft wird.

Für wissenschaftliche Hypothesen gilt das **Falsifikationsprinzip** nach Karl Popper: Es muss grundsätzlich möglich sein, die Hypothese zu widerlegen (Falsifikation). Die endgültige Verifikation einer Hypothese ist nicht möglich, da die Hypothese dann unter allen nur denkbaren Bedingungen zutreffen müsste – unzutreffende Annahmen können allerdings ausgeschlossen werden. Dies ist aber nicht überprüfbar. Deshalb kann sich eine Hypothese nur **bewähren**.

Deterministische Hypothese. Die deterministische Hypothese fordert, dass eine Aussage unter bestimmten Bedingungen **immer zutrifft**.

LERNTIPP !

Das IMPP wollte im Herbst 2015 wissen, welche Art von Hypothese über die Wirksamkeit einer Maßnahme in der medizinischen Forschung **nicht** aufgestellt wird. Deterministische Hypothesen sind absolute Tatsachenbehauptungen und daher in diesem Fall ungeeignet.

Probabilistische Hypothese. Eine Aussage, die nur mit einer bestimmten **Wahrscheinlichkeit** zutrifft, ist eine probabilistische Hypothese.

Beispiel: Der Risokofaktor löst bei einem Teil der Betroffenen die Krankheit aus.

Null- und Alternativhypothesen. Das Prinzip der Falsifikation findet in der Formulierung einer Nullhypothese Anwendung: Die Nullhypothese (H0) bestreitet das Vorliegen eines Effekts, der mit der Alternativ- oder Untersuchungshypothese (H1) angenommen wird.

- Die Alternativ- oder Untersuchungshypothese (H1) ist die Hypothese, die der Forscher belegen möchte. Sie postuliert, dass es den vermuteten bzw. gesuchten Effekt wirklich gibt.
- In Abhängigkeit der Untersuchungshypothese wird eine der Alternativhypothese entgegengesetzte Hypothese, die **Nullhypothese** (H0), aufgestellt. Sie postuliert, dass es den vermuteten bzw. gesuchten Effekt nicht gibt.

Die meisten Hypothesen in einem psychologischen Experiment sind probabilistische Hypothesen. Sie treffen also nur mit einer bestimmten Wahrscheinlichkeit zu. Das Wahrscheinlichkeitsniveau, auf dem die Alternativhypothese mindestens zutreffen muss, damit sie angenommen wird (und die Nullhypothese entsprechend abgelehnt), wird vorher festgelegt.

Wenn die **Irrtumswahrscheinlichkeit (Signifikanzniveau) α** 0,05 beträgt und die **Nullhypothese stimmt**, dann beträgt die Wahrscheinlichkeit 0,05 (5 %), fälschlicherweise die Nullhypothese abzulehnen. Dies wird auch durch die sog. **Signifikanz** ausgedrückt. Bei einer Irrtumswahrscheinlichkeit von 0,05 kann man auch sagen, das Ergebnis sei auf dem 5 %-Niveau signifikant. Durch eine Reduzierung des Signifikanzniveaus, z. B. auf $\alpha = 0{,}01$ statt auf $\alpha = 0{,}05$, sinkt die Wahrscheinlichkeit, dass das Ergebnis statistisch signifikant ist. Wenn ein Testergebnis statistisch signifikant ist, dann ist es nur mit einer geringen Wahrscheinlichkeit zufällig entstanden.

Anders ausgedrückt: Wenn die Nullhypothese mit einer Signifikanz von 0,05 (0,01) stimmt, wird in ca. 5 (1) von 100 gleichartigen Studien ein Unterschied im Ergebnis zwischen experimenteller und Kontrollbedingung gefunden werden.

Statistische Tests geben außerdem einen Signifikanzwert (**p-Wert**) an. Er gibt die Glaubwürdigkeit der Nullhypothese an bzw. anders ausgedrückt: Der p-Wert gibt an, mit welcher Wahrscheinlichkeit das Studienergebnis durch Zufall zustande kam. Dabei gilt: Die Nullhypothese wird dann verworfen, wenn $p <$ Irrtumswahrscheinlichkeit α, und sie wird angenommen, wenn $p > \alpha$.

α- und β-Fehler. Bei der Annahme bzw. der Ablehnung der Alternativ- oder/und Nullhypothese werden zwei Arten von Fehlern unterschieden. Der Fehler, einen Effekt als bedeutsam anzunehmen, obwohl er es eigentlich nicht ist, heißt α-Fehler (Fehler erster Art). Sein Gegenstück ist der β-Fehler (Fehler zweiter Art), bei dem ein Effekt für zufällig gehalten wird, obwohl er tatsächlich besteht.

- **α-Fehler** (Fehler 1. Art): Die Alternativhypothese wird fälschlicherweise für richtig gehalten. Per Konvention wird der akzeptierte α-Fehler auf 5 %, 1 % oder 1 ‰ bzw. 0,05, 0,01 oder 0,001 festgelegt (Signifikanzniveau). (Je mehr Tests in einer Stichprobe durchgeführt werden und je höher das Signifikanzniveau festgelegt wird, desto größer ist die Wahrscheinlichkeit der Erhöhung des Fehlers 1. Art.
- **β-Fehler** (Fehler 2. Art): Die Nullhypothese wird fälschlicherweise für richtig gehalten. Hier wird ein β-Fehler von 20 % toleriert.

FAZIT – DAS MÜSSEN SIE WISSEN ✘

- ! **Psychologie** ist die Wissenschaft vom Erleben und Verhalten des Menschen.
- !! **Falsifikationsprinzip:** Eine wissenschaftliche Hypothese muss grundsätzlich widerlegbar sein.
- !! Unzutreffende Annahmen können beim Falsifikationsprinzip ausgeschlossen werden.
- ! **Deterministische Hypothesen** über die Wirksamkeit einer Maßnahme werden in der medizinischen Forschung eher **nicht** aufgestellt.

- ! **Probabilistische Hypothese:** Eine Aussage, die nur mit einer bestimmten Wahrscheinlichkeit zutrifft, ist eine probabilistische Hypothese.
- ! **Beispiel für einen probabilistischen Zusammenhang:** Ein bestimmter Risikofaktor löst bei einem Teil der Betroffenen die Krankheit aus.
- !! **Signifikanz:** Ist ein Testergebnis statistisch signifikant, dann ist es nur mit einer geringen Wahrscheinlichkeit zufällig entstanden.
- ! Beträgt die **Irrtumswahrscheinlichkeit** (Signifikanzniveau) 0,05 und **stimmt** die **Alternativhypothese**, dann beträgt die Wahrscheinlichkeit, fälschlicherweise die Nullhypothese anzunehmen, 0,05 (5 von 100 Fällen). Man kann auch sagen, das Ergebnis sei auf dem **5 %-Niveau signifikant**.
- ! Durch eine Reduzierung des Signifikanzniveaus, z. B. auf $\alpha = 0{,}01$ statt auf $\alpha = 0{,}05$, sinkt die Wahrscheinlichkeit, dass das Ergebnis statistisch signifikant ist.
- !! Die **Nullhypothese** wird dann verworfen, wenn p-Wert < Irrtumswahrscheinlichkeit α, und sie wird angenommen, wenn $p > \alpha$.
- ! Der **p-Wert** gibt die Wahrscheinlichkeit an, mit der ein Studienergebnis durch Zufall zustande kam.
- ! Je mehr Tests in einer Stichprobe durchgeführt und je höher das Signifikanzniveau festgelegt wird, desto größer ist die Wahrscheinlichkeit der **Erhöhung des Fehlers 1. Art**.
- ! Bei einem **β-Fehler** (Fehler 2. Art) wird die Nullhypothese fälschlicherweise für richtig gehalten.

2.1.2 Operationalisieren, Beobachten und Messen

Operationalisierung. Die meisten Phänomene, mit denen sich die Sozialwissenschaften auseinandersetzen, sind nicht direkt zu beobachten. So sind **Lebensqualität**, **Gesundheit**, **Intelligenz**, **Resilienz**, **Introversion**, **Neurotizismus** und **Depressivität** z. B. **latente Konstrukte** oder **latente Variablen**, da man sie nicht direkt beobachten kann.

APROPOS

Verhaltensweisen hingegen sind direkt zu beobachtende Phänomene. Einen weinenden Menschen kann man direkt beobachten, weil ihm die Tränen über das Gesicht laufen. Wenn wir also jemanden weinen sehen, und wir sagen, dieser Mensch sei depressiv, so schließen wir auf das Konstrukt Depressivität. Dies kann unter Umständen auch falsch sein. Manchen Menschen laufen die Tränen über das Gesicht, obwohl sie nicht depressiv sind, sondern weil sie gerade eine Zwiebel geschält haben.

Operationalisierung bezeichnet den Vorgang, nicht direkt beobachtbare Phänomene (= **latente Variablen**) für die Beobachtung und Messung zugänglich zu machen. Dazu werden Variablen (S. 20) herangezogen, die beobachtet und somit gemessen werden können (= **manifeste Variablen**). Intelligenz ist z. B. solch ein nicht direkt beobachtbares Phänomen. Diese **latente Variable** (Intelligenz) kann durch **manifeste Variablen** messbar gemacht werden. Zum Beispiel durch die Anzahl gelöster Aufgaben in einem IQ-Test. Die Operationalisierung umfasst sowohl die Beschreibung der **Vorgehensweise bei der Messung** als auch die Beschreibung der eingesetzten Messinstrumente.

Arten der Beobachtung. Es werden nicht immer Experimente durchgeführt, um ein psychologisches oder soziologisches Phänomen zu erforschen. Die operationalisierten Kriterien werden häufig auch lediglich durch Beobachtung gewonnen.

- Bei einer **offenen Beobachtung** ist bekannt, wer und wo der Beobachter ist. Setzt sich ein Arzt beispielsweise zu seinen Patienten, um sie einfach besser kennenzulernen, handelt es sich um eine offene Beobachtung.
- Bei der **verdeckten Beobachtung** ist der Beobachter nicht zu sehen.

Die Beobachtungsformen lassen sich weiter in **teilnehmend** und **nicht teilnehmend** unterteilen.

Messen. Dieser Begriff ist von der Operationalisierung zu unterscheiden. „Messen" meint die Zuordnung von **empirischen Sachverhalten** zu **Zahlen** nach einer bestimmten Regel (Stevens, 1959).

2.1.3 Skalierungsmethoden

Testergebnisse lassen sich anhand verschiedener Skalenniveaus abbilden. Diese erlauben unterschiedliche Rechenoperationen.

Verhältnisskala (Rational- oder Absolutskala). Auf diesem Skalenniveau sind die meisten **Rechenoperationen** möglich. Hier weiß man, dass die Verhältnisse, die hier abgebildet werden, einen absoluten Nullpunkt haben. Dies sind Größen wie Körpergewicht, Temperatur in Kelvin, Reaktionszeiten etc. Erlaubte Rechenoperationen sind Multiplikation und Division (A ist doppelt so groß wie B) sowie Addition und Subtraktion. Als Maß der zentralen Tendenz (allgemein: Gipfel einer Häufigkeitsverteilung) kann das **geometrische Mittel** berechnet werden. Auf einer Verhältnisskala lassen sich in Zahlen ausgedrückt z. B. Serum-Enzymaktivitäten oder Reaktionszeiten abbilden.

Intervallskala. Diese Skala hat keinen absoluten Nullpunkt mehr. Erlaubte Rechenoperationen sind daher nur noch Addition und Subtraktion. Die Abstände zwischen den Merkmalsausprägungen entsprechen sich (typisches Beispiel: Temperaturskala nach Celsius – die Temperaturdifferenz von –12 Grad zu –10 Grad entspricht der von 14 Grad zu 16 Grad). Die Berechnung von **arithmetischem Mittel** (Summe der Einzelwerte, geteilt durch ihre Anzahl n; ist meist gemeint, wenn vom **„Mittelwert"** gesprochen wird) und der Abweichung von diesem Mittelwert (**Standardabweichung**) kann erfolgen. Die meisten psychologischen Testverfahren (Intelligenzquotient, Ängstlichkeit etc.) messen das Merkmal auf Intervallskalenniveau.

Ordinalskala (Rangskala). Die Merkmale, die hier abgebildet werden, lassen eine Anordnung nach bestimmten Kriterien (größer/kleiner, schlechter/besser, schöner/hässlicher) bzw. nach ihrer Ausprägungsstärke zu. Beispiel: Probanden können die Häufigkeit angeben, mit der bestimmte körperliche Beschwerden in den letzten Wochen aufgetreten sind: 1 (nie), 2 (selten), 3 (gelegentlich), 4 (oft), 5 (immer). Die Werte lassen sich auch in Prozent ausdrücken. Erlaubte Rechenoperationen sind $a < b$, $a > b$. Auf diesem Niveau werden auch Krankheitsstadien oder als intervenierende Variablen auch Schichtzugehörigkeit, Bildungsabschluss, Schulnote und sozialer Status abgebildet.

Die **zentrale Tendenz einer Ordinalskala** beschreibt der **Median** (Zentralwert = Wert, der in der Mitte steht, wenn alle vorhandenen Ausprägungswerte hierarchisch nebeneinander aufgereiht werden). Als Maß für die Streuung eignet sich der **Interquartilabstand** (Abb. 2.1).

Nominal- oder Kategorieskala. Auf diesem Skalenniveau lassen sich nur noch Kategorien bzw. kategoriale Variablen abbilden. Hier kann man also die wenigsten Aussagen machen. Kategorien

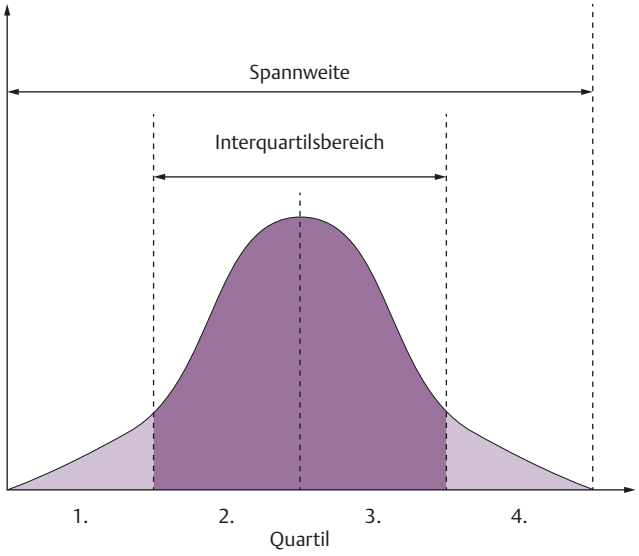

Abb. 2.1 **Streuung und Interquartilabstand.**

sind klar zuzuordnende Merkmale wie z. B. viele anamnestische Daten wie verheiratet – ledig, Mann – Frau, oder auch Diagnosen (wie der ICD-10). Als Maß der **zentralen Tendenz einer Nominalskala** kann der **Modus** (**Modalwert**) angegeben werden. Er bezeichnet das Merkmal, das am häufigsten ausgeprägt ist.

LERNTIPP !

Die Reihenfolge der Skalen geordnet nach ihren Niveaus vom niedrigsten zum höchsten lautet:
Nominalskala – Ordinalskala – Intervallskala – Verhältnisskala.

Absolute Beurteilungsskalen. Hierzu zählen Skalen, bei denen Merkmale auf einer **mehrstufigen Skala** direkt eingeschätzt werden (Ordinalskala-Niveau).

Hierher gehört auch die **dichotome Beurteilung** „trifft zu", „trifft nicht zu" (**Nominal- bzw. Kategorieskala**). Einige Beispiele zu Skalierungsmethoden werden im Folgenden besprochen:

- **Likert-Skala:** Hier geben die Probanden ihre Zustimmung auf einer meist fünfstufigen Skala an. Dabei werden die Antwortmöglichkeiten verbal beschrieben (z. B. „stimme gar nicht zu" und „stimme völlig zu"). Die Besonderheit dieser Skala ist, dass der Gesamttestwert eines Probanden berechnet wird, indem die angekreuzten Skalenwerte einfach zusammengezählt werden. Likert-Skalen sind zur Indexbildung geeignet.
- **Thurstone-Skala:** Hier liegt ein dichotomes Format von „stimme zu" und „stimme nicht zu" vor.
- **Numerische Analogskala:** Hier wird ein Merkmal auf einer Zahlenreihe (wie ein Lineal) zwischen zwei Extremwerten eingeschätzt. Patienten können beispielsweise das Ausmaß ihrer Schmerzen auf einer Zahlenreihe zwischen den Extremwerten „keine Schmerzen" bis „sehr starke Schmerzen" auftragen.
- **Visuelle Analogskala:** Es sind nur die Endpunkte der Skala markiert, dazwischen finden sich keine Zahlenwerte und auch sonst keine Einträge – sie ist völlig unbeschriftet. Auf der Rückseite können die Markierungen des Patienten dann in Zahlen abgelesen werden oder die Abstände zu den Endpunkten werden ausgemessen.

LERNTIPP !

Seien Sie sich dessen bewusst, dass bei einer Ordinal- oder Kategorieskala keine quantifizierbare Aussage wie „Prima, Ihre Schmerzen haben sich ja halbiert!" gemacht werden kann.

Relative Beurteilungsskalen. Bei relativen Beurteilungsskalen stellt man einen Vergleich an. „Sind Ihre Schmerzen heute stärker als gestern?" erfordert einen Vergleich mit dem Vortag. „Haben Sie Schmerzen?" ist eine absolute Frage.

Zu den relativen Beurteilungen zählen der **Rangvergleich**, der **Paarvergleich** und das **Soziogramm**.

FAZIT – DAS MÜSSEN SIE WISSEN ✘

- !!! **Latente Konstrukte:** z. B. Lebensqualität, Gesundheit, Intelligenz, Resilienz, Introversion, Neurotizismus und Depressivität, da man diese Parameter **nicht direkt beobachten** kann.
- ! **Manifeste Variablen:** z. B. Körpergröße oder Anzahl gelöster Aufgaben in einem IQ-Test, können direkt beobachtet und gemessen werden.
- ! **Latente Variablen** bzw. Konstrukte können durch **manifeste Variablen** messbar werden.
- !! **Operationalisierung:** Vorgang, bei dem man nicht direkt beobachtbare Phänomene (latente Konstrukte) für die Beobachtung und **Messung zugänglich** macht. Dazu werden **Variablen** herangezogen, die beobachtet und somit gemessen werden können. Die Operationalisierung umfasst sowohl die Beschreibung der **Vorgehensweise** bei der Messung als auch die Beschreibung der eingesetzten **Messinstrumente**.
- ! Bei der **verdeckten Beobachtung** ist der Beobachter nicht zu sehen.
- !!! **Verhältnisskala (Rational-, Absolutskala):** Auf diesem Skalenniveau sind die meisten Rechenoperationen möglich. Hier weiß man am meisten über die Wirklichkeit. Man weiß, dass die Verhältnisse, die hier abgebildet werden, einen absoluten Nullpunkt haben. Dies sind Größen wie Körpergewicht, Temperatur in Kelvin, Reaktionszeiten etc.
- ! **Intervallskala:** Hier gibt es eine hierarchische Abstufung mit gleichen Abständen zwischen den Merkmalsausprägungen, jedoch keinen absoluten Nullpunkt. Beispiele: Temperaturskala nach Celsius oder Intelligenzquotient.
- !!! Der **Mittelwert** (arithmetisches Mittel) von (mindestens intervallskalierten) Messwerten errechnet sich aus der Summe der Einzelwerte, geteilt durch ihre Anzahl.
- ! **Standardabweichung:** statistische Abweichung vom Mittelwert.
- ! **Median:** Wert, der in der Mitte steht, wenn alle vorhandenen Werte hierarchisch nebeneinander aufgereiht werden.
- ! **Modalwert:** Wert, der am häufigsten in einer Stichprobe vorkommt.
- !!! **Ordinalskala (Rangskala):** Die Merkmale, die hier abgebildet werden, lassen eine Zuordnung nach bestimmten Kriterien (größer/kleiner, schlechter/besser) zu. Beispiele sind Krankheitsstadien oder als intervenierende Variablen auch Schichtzugehörigkeit und Bildungsabschluss.
- ! Die zentrale Tendenz einer Ordinalskala beschreibt der **Median**. Er entspricht dem Wert, der direkt **in der Mitte der Verteilung** liegt.
- ! Als Maß für die Streuung von auf Ordinalskalenniveau erfassten Werten eignet sich der **Interquartilabstand**.

- !! **Nominal- oder Kategorieskala:** Auf diesem Skalenniveau lassen sich nur noch Kategorien bzw. kategoriale Variablen abbilden, wie z. B. verheiratet, ledig, geschieden oder Diagnosen nach ICD-10 etc.
- !!! Der **Modus** (**Modalwert**) ist ein Maß der zentralen Tendenz, das auf allen Skalenniveaus (also auch, aber nicht ausschließlich bei Nominalskalen) ermittelt werden kann. Er gibt die am häufigsten vorkommende Merkmalsausprägung an.
- !! Die Reihenfolge der **Skalen** geordnet nach ihren **Niveaus** vom niedrigsten zum höchsten lautet: **Nominalskala** – Ordinalskala – Intervallskala – **Verhältnisskala**.
- ! **Likert-Skala:** Hier geben die Probanden ihre Zustimmung auf einer meist fünfstufigen Skala an. So kann 1 „stimme gar nicht zu" und 5 „stimme völlig zu" bedeuten.
- ! **Numerische Analogskala:** Hier wird ein Merkmal auf einer **Zahlenreihe** (z. B. 1, 2, 3, 4, 5) zwischen zwei Extremwerten eingeschätzt.
- ! Eine **visuelle Analogskala** beinhaltete zwei gegensätzliche Bezeichnungen auf einer **völlig unbeschrifteten Skala**. Sie kann der Messung des **subjektiven Schmerzempfindens** dienen.

2.1.4 Testdiagnostik

Im folgenden Kapitel wird beschrieben, wie man einen Test konstruiert und wie man seine Tauglichkeit überprüft.

Testkonstruktion

Ein psychologischer Test ist ein Verfahren, mit dem quantifizierbare Aussagen über psychische Merkmale gemacht werden können. Es sollen also hypothetische Konstrukte gemessen werden. Psychologische Tests unterteilen sich in Leistungs-, Persönlichkeits- und Intelligenztests.

Zunächst werden **Testaufgaben** (Items) ausgewählt. Durch eine Itemselektion wird entschieden, welche Aufgaben in die Endform kommen. Kriterien hierfür sind Itemschwierigkeit, Trennschärfekoeffizient und Itemhomogenität.

- Die **Itemschwierigkeit** besagt, wie viele Probanden die Frage richtig gelöst haben.
- Mit dem **Trennschärfekoeffizienten** wird beurteilt, wie die Beantwortung eines Items mit dem Gesamttestergebnis (über alle Items) zusammenhängt. Wenn man davon ausgeht, dass das Physikum ein Test ist, der das medizinische Wissen testet, dann sind die einzelnen Aufgaben die Items. Hat eine Aufgabe eine hohe Trennschärfe, wird bei richtiger Lösung dieser Aufgabe auch das Gesamtergebnis des Physikums gut sein. Das Lösen einer Aufgabe mit niedriger Trennschärfe wiederum ist weniger mit dem Gesamtergebnis korreliert.
- Die **Itemhomogenität** besagt, wie sehr sich die einzelnen Items in Schwierigkeit und Trennschärfe gleichen.

Eine gute Trennschärfe liegt bei etwa 50 %, dies gilt auch für die Itemhomogenität. Nun wird der Test auf seine Güte geprüft (s. u.).

Die entstandene Testendform wird dann an einer **Normstichprobe** normiert. Man spricht auch von Eichung und einer Eichstichprobe. Für die Normierung benötigt man eine möglichst große und repräsentative Stichprobe. Aus diesen Ergebnissen werden Normen gewonnen. Anhand dieser Normen lassen sich individuelle Testergebnisse interpretieren.

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

Testgütekriterien

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

Objektivität. Objektivität meint die Unabhängigkeit des Tests von der Person des Testleiters und besagt, dass **jeder** Testleiter, der den Test durchführt, auswertet und interpretiert, zu **demselben Ergebnis** kommen muss. Ein Maß für die Objektivität ist die **Interrater-Reliabilität** (IR). Die **IR** beschreibt das Ausmaß der **Übereinstimmung** der Einschätzungen von unterschiedlichen Beobachtern. Wird z. B. ein Proband von zwei verschiedenen Ärzten mit dem gleichen Verfahren getestet und die Ergebnisse beider Ärzte stimmen überein, dann ist die IR hoch und der Test erfüllt das Gütekriterium der Objektivität.

Reliabilität (Zuverlässigkeit). Sie macht eine Aussage über die **Genauigkeit**, mit der ein Test ein bestimmtes Merkmal misst, egal, ob er dieses Merkmal auch messen soll. Es gibt mehrere Arten, die Reliabilität eines Tests zu bestimmen.

- **Retest-Reliabilität** (auch Test-Retest-Verfahren): Der Test wird mit ein und derselben Versuchsperson zweimal durchgeführt und die Testergebnisse miteinander verglichen (korreliert, s. u.). Ein hoher Reliabilitätskoeffizient liegt vor, wenn beide Male ähnliche Ergebnisse erzielt werden.
- **Split-Half-Reliabilität:** Ein Test wird in zwei Teile geteilt. Nun lässt man Probanden beide Testteile ausfüllen und vergleicht dann die Ergebnisse der Testteile (Splithalf = in die Hälfte geteilt).
- **Innere Konsistenz (Inter-Item-Konsistenz-Analyse):** Die Methode ähnelt der Testhalbierungsmethode, jedoch wird hier jede einzelne Testaufgabe mit allen Testaufgaben in Beziehung gesetzt. Ein Korrelationskoeffizient (s. u.) gibt demnach Auskunft über die innere Konsistenz.
- **Paralleltest-Reliabilität:** Ein vergleichbares Testverfahren sollte zu gleichen Ergebnissen kommen. Der Korrelationskoeffizient wäre in einem solchen Fall hoch.

Durch **Verlängerung** eines Tests verbessert sich seine Reliabilität (aus statistischen Gründen), durch Verkürzung verschlechtert sie sich.

Korrelationskoeffizient. Je höher die Reliabilität, desto besser der Test. Der Reliabilitätswert wird durch den Korrelationskoeffizienten (r) ausgedrückt. Er sagt aus, ob und wie 2 Variablen miteinander zusammenhängen.

- Ein **positiver linearer Korrelationskoeffizient** (z. B. r = 0,7) gibt an, dass ein hoher Wert einer Variablen mit einem hohen Wert einer anderen Variablen einhergeht (z. B. Gedächtnisleistung und Intelligenz). Der Zusammenhang ist gleichgerichtet.
  Korrelieren Testergebnisse zur **sozialen Intelligenz** mit jenen zur allgemeinen Intelligenz, bedeutet dies für den Versuch selbst, dass nur wenige zusätzliche Informationen durch die Testung der sozialen Intelligenz gefunden werden.
- Ein **negativer linearer Korrelationskoeffizient** (z. B. r = –0,79) gibt an, dass ein hoher Wert der einen Variablen mit einem niedrigen Wert der anderen Variablen einhergeht (entgegengerichteter Zusammenhang).
- Ein eher **niedriger Korrelationskoeffizient** (z. B. r = 0,15) gibt an, dass die beiden Variablen eher einen geringen Zusammenhang zeigen.

LERNTIPP !

Der Korrelationskoeffizient macht keine Aussage über die Kausalität, also welche Variable die jeweils andere beeinflusst. Weiterhin kann aus einem bekannten Testwert nicht allein mithilfe des Korrelationskoeffizienten der Wert der anderen Variablen berechnet werden.

Arten von Korrelationskoeffizienten. Welche Art Korrelationskoeffizient gewählt wird, hängt von dem Skalenniveau der Variablen ab.

Bei intervallskalierten und normalverteilten Variablen wird die **Produkt-Moment-Korrelation** nach **Pearson** berechnet. Der Korrelationskoeffizient r liegt dabei immer im Bereich von **–1** bis **1**. **–1** bedeutet, es liegt ein absoluter Gegensatz zwischen den beiden Variablen vor, bei einem Wert von 0 besteht kein linearer Zusammenhang zwischen den Variablen. Man erreicht nie eine Reliabilität von 1, jedoch strebt man Werte an, die um 0,8 oder 0,9 liegen. Von einem **starken Zusammenhang** wird konventionsgemäß gesprochen wenn **r > 0,5**. Eine gewisse Unzuverlässigkeit muss also in Kauf genommen werden. Diese Ungenauigkeit wird durch den Standardmessfehler ausgedrückt.

Ist eine der Variablen nicht normalverteilt oder ordinalskaliert, wird auf den Rangkorrelationskoeffizienten nach **Spearman** (S. 38) zurückgegriffen.

BEISPIEL

Eine Studie beschreibt den Zusammenhang zwischen dem starken Rauchen und dem Auftreten einer Bronchitis im vergangenen Jahr mit einer Produkt-Moment-Korrelation von r = 0,8 und einem Signifikanzwert von p = 0,001 bei einem Signifikanzniveau α von 0,05. Da p < α, ist die Korrelation signifikant. Demnach wird die Nullhypothese falsifiziert und die Alternativhypothese (S. 13) angenommen. Der positive (hohe) Korrelationskoeffizient von 0,8 gibt an, dass der Zusammenhang zwischen beiden Variablen (Rauchen und Bronchitis) stark ist.

LERNTIPP !

Hier noch ein praktisches Beispiel, das vom IMPP so auch schon abgefragt wurde: Wenn bei einem Intelligenztest die Retest-Reliabilität r = 0,50 beträgt, hat dieser Test eine vergleichsweise schlechte Reliabilität. Er liefert bei mehrfacher Durchführung bei derselben Versuchsperson nicht sehr zuverlässig ähnliche Ergebnisse: Ein Proband, der beim 1. Durchgang ein prima Ergebnis hatte (hier: hoher IQ), schneidet ggf. bei der Wiederholung deutlich schlechter ab.

Standardabweichung (SD). Die Standardabweichung ist ein Maß für die Streuung von Testwerten. Sie bezeichnet einen Bereich um den Mittelwert (M) eines Tests, in dem mit einer bestimmten Wahrscheinlichkeit der tatsächliche Wert des Tests liegt. Es gilt:

- 68,3 % aller Werte eines Tests liegen zwischen M ± 1 SD.
- 95,4 % aller Werte eines Tests liegen zwischen M ± 2 SD.

Beispiel: Über einen entsprechenden Test wird herausgefunden, dass die Lebensqualität eines bestimmten Schmerzpatienten 2 SD unter dem Mittelwert liegt. Von den oben genannten 95,4 % liegt die Hälfte der Werte oberhalb und die andere Hälfte unterhalb des Mittelwertes, also jeweils 47,7 %. Dies bedeutet, dass nur 2,3 % der Gesamtbevölkerung einen noch niedrigeren Wert aufweisen.

13,55 % der Gesamtbevölkerung weisen bei normalverteilten Daten Werte auf, die mindestens eine, aber nicht mehr als zwei Standardabweichungen vom Mittelwert nach oben abweichen: (95,4 % – 68,3 %) / 2 = 13, 55 %.

Standardmessfehler (SM). Der Messfehler, der durch die mangelnde Reliabilität eines Tests zustande kommt, wird als **Standardmessfehler (SM)** bezeichnet. Er errechnet sich aus dem **Reliabilitätskoeffizienten** (r) und der **Standardabweichung** (SD) der Testwerteverteilung. Die Standardabweichung gehört zu den Kennwerten der Intervallskala:

$$SM = SD \sqrt{(1 - r)}$$

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 (z. B. zu 95 %) liegt. Je reliabler der Test, desto geringer ist der Standardmessfehler und desto enger ist das Konfidenzintervall.

Varianz. Einzelne Testwerte können dem Mittelwert entsprechen oder mehr oder weniger von ihm abweichen (gestreute Daten). Zur Formulierung des Ausmaßes der Streuung dienen die Variabilitätsmaße **Varianz** und **Standardabweichung**. Die Varianz bezeichnet die Abweichung der Messergebnisse von ihrem Mittelwert. Sie wird berechnet aus der Summe der **quadrierten Abweichungen** vom Mittelwert, geteilt durch die Anzahl der Messwerte.

Validität. Sie ist die **Gültigkeit**. Ein Test ist dann valide, wenn er auch das misst, was er zu messen vorgibt. Ein Test, der Angst misst, sollte also das Konstrukt Angst erfassen und nicht etwa das latente Konstrukt Introversion. Die Validität kann auf mehrere Arten bestimmt werden:

- Bei der **Kriteriumsvalidität** wird die Validität gemessen, indem das Testergebnis mit einem Außenkriterium in Beziehung gesetzt (korreliert) wird. 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 (vorhersage-)valide, wenn er den späteren Berufserfolg gut vorhersagen kann. Übereinstimmungsvalidität wird z. B. getestet, indem ein neuartiger Computer-Persönlichkeitstest parallel zu dem entsprechenden älteren und bewährten Papiertest durchgeführt wird. Der Computertest ist dann übereinstimmungsvalide, wenn die Ergebnisse ähnlich denen des Papiertests sind.
- Häufig gibt es für ein komplexes Konstrukt nicht ein einzelnes Merkmal. Bei der **Konstruktvalidität** wird deshalb überprüft, inwieweit das Testergebnis mit anderen Indikatoren desselben Konstrukts zusammenhängt. Die Konstruktvalidität lässt sich weiter in eine **konvergente Validität** und eine **diskriminante Validität** unterteilen. Die konvergente Validität stellt die Korrelation zwischen unterschiedlichen Tests dar, die das selbe Konstrukt messen. Die Korrelation sollte hierbei hoch ausfallen (positiver linearer Korrelationskoeffizient). Wenn jemand bei einem Angsttest sehr hohe Werte erzielt, dann sollte er z. B. eher schüchtern sein und kein extrovertierter „Draufgänger". Die diskriminante Validität indes stellt die Korrelation unterschiedlicher Tests dar, die zudem unterschiedliche Konstrukte