Peter Sedlmeier
Frank Renkewitz

Forschungsmethoden und Statistik

Ein Lehrbuch für Psychologen und Sozialwissenschaftler

2., aktualisierte und erweitertete Auflage

Bibliografische Information der Deutschen Nationalbibliothek

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

10 9 8 7 6 5 4 3 2 1

15 14 13

ISBN 978-3-8689-4131-9 (Print); 978-3-86326-683-7 (PDF)

Lektorat: Kathrin Mönch, kmoench@pearson.de
Tim Schönemann, München
Korrektorat: Ruth Schneider, München
Einbandgestaltung: Thomas Arlt, tarlt@adesso21.net
Herstellung: Claudia Bäurle, cbaeurle@pearson.de
Satz: mediaService, Siegen (www.mediaservice.tv)
Druck und Verarbeitung: Drukarnia Dimograf, Bielsko-Biala

Printed in Poland

Messen und Testen

3

ÜBERBLICK

» Im vorangegangenen Kapitel haben wir uns mit der Frage befasst, wie wissenschaftliche Erkenntnisse gewonnen werden können. Insbesondere sind wir dabei darauf eingegangen, wie sozialwissenschaftliche Theorien und Hypothesen überprüft werden. Wie wir gesehen haben, erfolgen solche Überprüfungen empirisch: Aus einer Theorie abgeleitete Vorhersagen werden mit der Realität verglichen. Zu diesem Zweck werden in empirischen Untersuchungen Daten erhoben – dies bedeutet nichts anderes, als dass Messungen durchgeführt werden. Bevor wir in den nachfolgenden Kapiteln spezifische, in den Sozialwissenschaften häufig verwendete Verfahren der Datenerhebung behandeln, werden wir uns in diesem Kapitel mit den Grundlagen des Messens beschäftigen. Was genau ist Messen? Können die Ergebnisse von Messungen des Körpergewichts ebenso interpretiert werden wie die Ergebnisse von Messungen der Intelligenz? Schon hier sei erwähnt, dass die Antwort auf diese Frage „Nein" lautet. Stellen wir z.B. fest, dass die Herren Jäger und Kunze 140 kg und 70 kg wiegen, so ist Herr Jäger offensichtlich doppelt so schwer wie Herr Kunze. Messen wir dagegen bei denselben Personen IQ-Werte von 140 und 70, so ist die Aussage, Herr Jäger sei doppelt so intelligent wie Herr Kunze, dennoch *nicht* gerechtfertigt. Der Grund dafür besteht darin, dass verschiedene Messungen zu Messwerten mit unterschiedlichem Informationsgehalt führen können. Anhand des Informationsgehalts der Messwerte werden verschiedene Skalenniveaus unterschieden. Welche Informationen können wir einer Messung auf einem bestimmten Skalenniveau entnehmen? Wie lässt sich feststellen, auf welchem Skalenniveau eine Messung erfolgt? Welche Konsequenzen hat das Skalenniveau für die weitere Auswertung der Daten? Ist es z.B. bei allen Messungen sinnvoll, aus den Messwerten einen Mittelwert zu berechnen? Treten bei Messungen in der Psychologie und den Sozialwissenschaften spezifische Probleme auf, die etwa bei physikalischen Messungen nicht bestehen? Gibt es Kriterien, anhand derer beurteilt werden kann, ob eine Messung „gelungen" ist? Nach der Lektüre dieses Kapitels sollte die Beantwortung dieser Fragen kein Problem mehr darstellen! «

3.1 Was ist Messen?

In allen empirischen Untersuchungen werden Daten erhoben. Diese Daten sind das Ergebnis von Messungen. Eine Messung bezieht sich stets auf eine *Variable*. Mit dem Begriff Variable werden beliebige Merkmale oder Eigenschaften eines Objekts oder einer Person bezeichnet, die mindestens zwei Ausprägungen annehmen können. Eine Variable kann durch die Menge der möglichen Merkmalsausprägungen beschrieben werden. Beispiele für Variablen, die genau zwei Ausprägungen annehmen können, sind das Geschlecht oder die Teilnahme an einem Fahrsicherheitstraining (mit den Ausprägungen „teilgenommen“ und „nicht teilgenommen“). Bei anderen Variablen entsprechen die Ausprägungen nicht zwei, sondern mehreren möglichen Kategorien. Solche Variablen sind etwa die Partei- und Religionszugehörigkeit, das Studienfach oder die psychiatrische Diagnose. Schließlich können die möglichen Ausprägungen einer Variable in vielen (oder auch unendlich vielen) unterschiedlichen Intensitätsgraden eines Merkmals bestehen. Dies ist bei den Variablen Länge, Temperatur, Windstärke, Alter, Intelligenz, Reaktionszeit bei einer bestimmten Aufgabe, momentane Zufriedenheit, Ausmaß des Therapiefortschritts usw. der Fall.

Das Ziel des Messens besteht nun darin, die Ausprägung eines Merkmals, die bei einem bestimmten Objekt (oder einer Person) zu einem bestimmten Zeitpunkt gegeben ist, zu ermitteln. Dabei soll die jeweilige Merkmalsausprägung durch eine Zahl ausgedrückt werden. Eine erste vorläufige Definition von „Messen“ könnte also lauten: Messen besteht in der Zuordnung von Zahlen zu Objekten oder Personen.

Anders als physikalische Messungen treffen Messungen in der Psychologie relativ häufig auf öffentliche Kritik. Diese Kritik äußert sich zum Teil in Aussagen wie „Man kann die Seele des Menschen nicht in Zahlen fassen“. In dieser Form greift die ablehnende Beurteilung psychologischer Messungen allerdings schon deswegen ins Leere, da niemand – auch kein Psychologe – beabsichtigt, die Seele zu messen. Ebenso wenig wie ein Objekt als Ganzes gemessen werden kann, kann der gesamte Mensch (oder seine Seele) gemessen werden. Gemessen werden stets nur einzelne, definierte Eigenschaften von Objekten oder Menschen, also etwa die Länge eines Tisches oder die Intelligenz einer Person. Allerdings löst auch der Gedanke, spezifische psychische Phänomene in Zahlen zu fassen, vielfach Unbehagen aus. Tatsächlich erscheint es uns im Alltag vereinfachend und unangemessen, beispielsweise das Ausmaß unserer aktuellen Verärgerung oder unserer Zufriedenheit in einer Zahl auszudrücken. Ist es also überhaupt möglich oder vernünftig derartige Phänomene zu messen? Dem Unbehagen am Messvorgang steht die Beobachtung gegenüber, dass wir keine Schwierigkeiten haben, sprachliche Aussagen zu treffen, die sich auf die Ausprägung von psychischen Merkmalen beziehen. In vielen Fällen wird uns die Aussage, dass sich unsere Zufriedenheit seit gestern nicht geändert hat, keine Probleme bereiten. Ebenso können wir überzeugt feststellen, dass uns der Streit mit dem Vorgesetzten noch mehr verärgert hat als die Unfreundlichkeit des Schuhverkäufers. Auch der Aussage, dass Frau A intelligenter ist als Herr B, können wir ohne Zögern zustimmen. Wenn diese Aussagen aber einen Sinn haben, also eine gültige Behauptung über die Realität darstellen, dann lassen sich den fraglichen Merkmalen auch entsprechende Zahlen zuordnen. Wir könnten also etwa unserer heutigen Zufriedenheit die gleiche Zahl zuweisen wie der gestrigen oder die beschriebenen Intelligenzunterschiede durch eine höhere Zahl für Frau A als für Herrn B ausdrücken.

Könnten wir auch darauf verzichten, Ausprägungen psychischer Merkmale durch Zahlen zu bezeichnen und es bei verbalen Beschreibungen belassen? Für die Psychologie – und jede andere empirische Wissenschaft – bietet die Verwendung von Zahlen äußerst wichtige Vorteile. Zunächst ist die Bedeutung von Zahlen viel präziser festgelegt als die Bedeutung von sprachlichen Beschreibungen. So wird die Aussage „Herr X ist sehr groß" zu deutlich unterschiedlicheren Interpretationen führen als die Aussage „Herr X ist 2 Meter groß". Zahlen erlauben damit auch feinere Differenzierungen zwischen verschiedenen Merkmalsausprägungen als einfache sprachliche Beschreibungen. Schließlich besteht ein wichtiges Ziel der Psychologie darin, Beziehungen zwischen Variablen zu ermitteln. Nur wenn die Ausprägungen dieser Variablen in Zahlen gefasst werden, ist auch eine mathematische Beschreibung der Beziehung zwischen ihnen möglich. Erst die Messung von Variablen erlaubt uns also Aussagen wie: „Bei einer Änderung der Variable A um eine Einheit ist eine Änderung der Variable B um 5 Einheiten zu erwarten".

Nun ist selbstverständlich nicht jede beliebige Zuordnung von Zahlen zu Merkmalsausprägungen eine Messung. Offensichtlich wäre es völlig sinnlos, den Teilnehmern eines Statistikkurses bezüglich ihrer Intelligenz irgendwelche Zahlen zufällig zuzuweisen. Von einer Messung kann erst dann gesprochen werden, wenn es eine Zuordnungsregel gibt. Diese Zuordnungsregel muss gewährleisten, dass bestimmte Relationen (Beziehungen) zwischen den Zahlen analoge empirische Relationen zwischen den Messobjekten abbilden. Wenn wir also mittels beobachtbarer Indikatoren feststellen können, dass zwischen Frau A und Herrn B die Relation „ist intelligenter als" besteht, so muss Frau A hinsichtlich ihrer Intelligenz auch eine größere Zahl (ein größerer Messwert) zugeordnet werden als Herrn B.

Damit eine Zuordnung von Zahlen zu Objekten (oder Personen) als Messung gelten kann, müssen allerdings nicht alle denkbaren Relationen zwischen den Zahlen auch entsprechende empirische Beziehungen zwischen den Objekten zum Ausdruck bringen. Relationen zwischen Zahlen sind z.B. „gleich", „größer als" oder „doppelt so viel wie". Nicht bei jeder Messung enthalten alle diese Relationen zwischen Messwerten tatsächlich Informationen über die Messobjekte. Welche Relationen zwischen Messwerten informationshaltig sind und somit sinnvoll interpretiert werden können, hängt davon ab, welche Beziehungen zwischen den Messobjekten empirisch festgestellt werden können und bei der Messung auch berücksichtigt wurden.

Betrachten wir hierzu einige Beispiele: Nehmen wir an, wir wollen das Geschlecht verschiedener Personen messen. Empirisch ist hinsichtlich des Geschlechts ausschließlich die Relation „gleich" bzw. „ungleich" bedeutungsvoll. Eine Messung des Geschlechts könnte also darin bestehen, jedem Mann eine 1 und jeder Frau eine 2 zuzuordnen. Ebenso gut könnten wir jedem Mann eine 0,5 und jeder Frau eine 3157 zuordnen, da es bei dieser Messung ausschließlich darauf ankommt, dass alle Personen gleichen Geschlechts den gleichen Messwert erhalten. Demgemäß liefern auch die Messwerte ausschließlich Information über die Gleichheit oder Ungleichheit des Geschlechts. Die Tatsache, dass die 2 größer ist als die 1 (oder die 3157 größer als die 0,5) ist dagegen bedeutungslos, da es inhaltlich sinnlos ist zu behaupten, eine Frau sei „mehr" als ein Mann.

Zu informativeren Messwerten sollten wir z.B. dann gelangen, wenn wir versuchen, die Präferenz einer Kundin für vier verschiedene Handymodelle zu erfassen. Eine einfache Möglichkeit, dies zu tun, bestünde darin, die Kundin zu bitten, die Handys in eine Rangreihe zu bringen. Dem Handy, das der Kundin am besten gefällt, könnten wir dann

eine 4 zuordnen. Das Handy, das der Kundin am wenigsten gefällt, erhält hingegen eine 1. In diesem Fall ist es keineswegs bedeutungslos, dass beispielsweise der Messwert 4 größer ist als der Messwert 2: Diese Messwerte zweier Handys bringen hier die empirische Tatsache zum Ausdruck, dass die Kundin angegeben hat, das eine Handy stärker zu bevorzugen als das andere. Dass der Messwert 4 doppelt so groß ist wie der Messwert 2, erlaubt uns allerdings immer noch keine entsprechende Aussage über die Präferenzen für die beiden Handys. Wir wissen nicht, ob die Präferenz der Kundin für das eine Handy doppelt so groß ist wie ihre Präferenz für das andere Handy. Möglicherweise gefällt der Kundin das Handy mit dem Messwert 4 deutlich besser als alle übrigen Handys, zwischen denen sie nur geringe Unterschiede ausmachen kann. In diesem Fall wäre der Unterschied in ihrer Präferenz zwischen den Handys mit den Messwerten 4 und 3 deutlich größer als der Unterschied zwischen den Handys mit den Messwerten 3 und 2. Gleiche zahlenmäßige Unterschiede zwischen den Messwerten für verschiedene Handys bedeuten hier also nicht, dass auch zwischen den Präferenzen für die entsprechenden Handys gleiche Unterschiede bestehen. Demgemäß können auch Relationen wie „doppelt so groß wie“ zwischen den Messwerten nicht sinnvoll interpretiert werden. Da wir empirisch ausschließlich festgestellt haben, welches Handy der Kundin am besten, am zweitbesten usw. gefällt, hätten wir den Handys auch keineswegs die Messwerte 4, 3, 2 und 1 zuordnen müssen. Die Messwerte 22, 20, 11 und 5 wären genauso angemessen gewesen. Entscheidend ist nur, dass die Rangordnung der Zahlen der Rangordnung der Handys, die die Kundin vorgenommen hat, entspricht. Alle anderen Relationen zwischen den Zahlen sind bedeutungslos.

Gänzlich anders ist die Situation z.B. bei der Messung einiger physikalischer Variablen wie Länge und Gewicht. Hier können wir empirisch leicht ermitteln, dass der Größenunterschied zwischen Person A und Person B genauso groß ist wie der Größenunterschied zwischen den Personen B und C. Ebenso können wir mit sehr einfachen Mitteln feststellen, dass Person A doppelt so viel wiegt wie Person C. Eine geeignete Messung sollte derartige empirische Relationen auch in den Messwerten abbilden. Die Messwerte 100, 75 und 50 für das Gewicht dreier Personen informieren uns dann auch über die Gewichtsunterschiede zwischen den Personen und darüber, dass die schwerste der Personen doppelt so viel wiegt wie die leichteste. Allerdings sind die Zahlen, die wir den Messobjekten zuordnen, auch hier nicht eindeutig durch die empirischen Relationen zwischen den Objekten festgelegt. Die Messwerte 200, 150 und 100 würden die von uns ermittelten Relationen ebenso zum Ausdruck bringen wie die Messwerte 100, 75 und 50. Bei Messungen von Merkmalen wie Gewicht oder Länge sind uns solche *Transformationen* von Messwerten sehr vertraut: Selbstverständlich können wir das Gewicht sowohl in Kilogramm als auch in Pfund angeben.

Wie die vorangegangenen Beispiele zeigen, kommt es beim Messen zunächst darauf an, empirische Relationen zwischen den zu messenden Objekten zu ermitteln. Mithilfe einer geeigneten Zuordnungsregel sollen den Objekten dann Zahlen zugewiesen werden, deren Relationen die empirischen Relationen widerspiegeln. Die Messtheorie beschäftigt sich nun zunächst mit der Frage, welche Voraussetzungen die empirischen Relationen erfüllen müssen, damit es überhaupt möglich ist, geeignete Zuordnungsregeln zu finden. Darüber hinaus besteht die Aufgabe der Messtheorie darin, spezifische Zuordnungsregeln zu erarbeiten.

3.2 Messtheorie

In der Messtheorie wird eine formale, zunächst vielleicht etwas gewöhnungsbedürftige Sprache verwendet. Um nachvollziehen zu können, welche Probleme sich bei der Erarbeitung von Zuordnungsregeln stellen, benötigen wir einige Begriffe aus dieser Sprache:

Ein *empirisches Relativ* besteht aus einer Menge von Objekten und einer oder mehreren beobachtbaren Relationen zwischen diesen Objekten. Die Menge von Objekten enthält jeweils diejenigen Objekte (oder Personen), die gemessen werden sollen. Beispiele könnten also vier verschiedene Handymodelle, die Teilnehmer eines Statistikkurses, die Schüler einer Klasse oder auch die Bretter auf einer Baustelle sein. Wichtige Arten von Relationen sind die *Äquivalenzrelation* (die mit ~ gekennzeichnet wird) und die *Ordnungsrelation* (für die man auch $\succ$ schreibt). Die Äquivalenzrelation besagt, dass verschiedene Objekte hinsichtlich eines Merkmals die gleiche Ausprägung aufweisen. Die Äquivalenzrelation könnte also etwa eine Gruppe von Studierenden in Psychologie-, Soziologie- und Pädagogikstudenten unterteilen (innerhalb jeder dieser Untergruppen weisen die Studierenden die gleiche Ausprägung auf dem Merkmal „Studienfach" auf; Studierende in verschiedenen Untergruppen sind hingegen hinsichtlich des Merkmals Studienfach nicht äquivalent). Die Ordnungsrelation bringt zum Ausdruck, dass ein Merkmal bei einem Objekt stärker ausgeprägt ist als bei einem anderen. Besteht zwischen den Objekten in einem empirischen Relativ eine Ordnungsrelation, so bringt diese die Messobjekte in eine Rangreihe. Zu beachten ist, dass es sich bei der Äquivalenz- und der Ordnungsrelation um Arten von Relationen handelt. Konkrete empirische Relationen beinhalten immer auch das zu messende Merkmal. Empirische Äquivalenzrelationen sind also z.B. „hat das gleiche Geschlecht", „gehört der gleichen Partei an" oder „hat die gleiche Intelligenz". Empirische Ordnungsrelationen wären etwa „ist länger", „ist zufriedener", „gefällt besser" oder „ist depressiver".

Ein *numerisches Relativ* besteht aus einer Menge von Zahlen und einer bestimmten Anzahl von definierten Relationen zwischen diesen Zahlen. Beispiele für solche Zahlenmengen sind alle natürlichen Zahlen oder alle reellen Zahlen. Im Kontext des Messens sind wichtige Relationen zwischen Zahlen die Gleichheitsrelation (=) und die Größer-Kleiner-Relation (>).

Die Zuordnung von Objekten und Zahlen wird in der Messtheorie als *Abbildung* bezeichnet. Beim Messen wird ein empirisches Relativ in ein numerisches Relativ abgebildet. Dabei muss jedem Objekt aus dem empirischen Relativ genau eine Zahl aus dem numerischen Relativ zugeordnet werden. Die Regel, nach der die Zuordnung erfolgt, bezeichnen wir als (Abbildungs-)Funktion.

Eine solche Abbildung kann durch eine Menge von Pfeilen dargestellt werden (▶Abbildung 3.1). Nehmen wir an, wir wollten das Gewicht von fünf Personen messen. Durch die Abbildungsfunktion wird nun jeder Person eine Zahl zugewiesen (man sagt auch: Jedes Objekt wird in eine Zahl abgebildet). Demgemäß geht von jedem Objekt im empirischen Relativ genau ein Pfeil aus. Dies ist beim Messen sicherlich vernünftig: Bliebe ein Objekt ohne Pfeil, so erhielte es keinen Messwert. Gingen von einem Objekt dagegen zwei Pfeile aus, so würden ihm zwei Messwerte zugeordnet – dies wäre offensichtlich sinnlos, da eine Person nicht zugleich zwei „Gewichte" haben kann. Andererseits ist es durchaus möglich, dass mehrere Pfeile auf dieselbe Zahl im numerischen Relativ verweisen oder dass kein Pfeil bei einer bestimmten

Zahl endet. Auch dies ist im Kontext des Messens sinnvoll: Haben zwei Personen das gleiche Gewicht, so sollte ihnen natürlich auch die gleiche Zahl zugeordnet werden. Zudem ist es offensichtlich nicht erforderlich, dass jede beliebige Zahl im numerischen Relativ auch der Merkmalsausprägung eines Objekts im empirischen Relativ entspricht. Würden wir etwa das Geschlecht der fünf Personen messen, so könnte die Abbildungsfunktion besagen: Ordne jedem Mann eine 1 und jeder Frau eine 2 zu. In diesem Fall ginge von jedem Objekt im empirischen Relativ ein Pfeil aus, der entweder bei der 1 oder bei der 2 endet. Alle anderen Zahlen würden keine existierende Merkmalsausprägung repräsentieren.

Wie wir bereits gesehen haben, kann eine Abbildung eines empirischen Relativs in ein numerisches Relativ nur dann als Messung gelten, wenn die Relationen zwischen den Messobjekten auch durch die Relationen zwischen den zugeordneten Zahlen zum Ausdruck gebracht werden. Eine Abbildung, die diese Bedingung erfüllt, wird als *homomorphe Abbildung* bezeichnet. Besteht zwischen den Objekten eines empirischen Relativs ausschließlich eine Äquivalenzrelation (wie dies etwa bei der Messung des Geschlechts der Fall ist), so würde eine homomorphe Abbildung sicherstellen, dass zwei Objekten genau dann der gleiche Messwert zugeordnet wird, wenn sie die gleiche Merkmalsausprägung haben. Ist in einem empirischen Relativ zusätzlich eine Ordnungsrelation gegeben (wie wir dies bei der Messung der Präferenz für verschiedene Handys angenommen haben), so führt eine homomorphe Abbildung dazu, dass ein Objekt A genau dann einen höheren Messwert erhält als ein Objekt B, wenn es auch die größere Merkmalsausprägung aufweist. In den formalen Begriffen der Messtheorie ausgedrückt ist Messen also nichts anderes als die homomorphe Abbildung eines empirischen Relativs in ein numerisches Relativ.

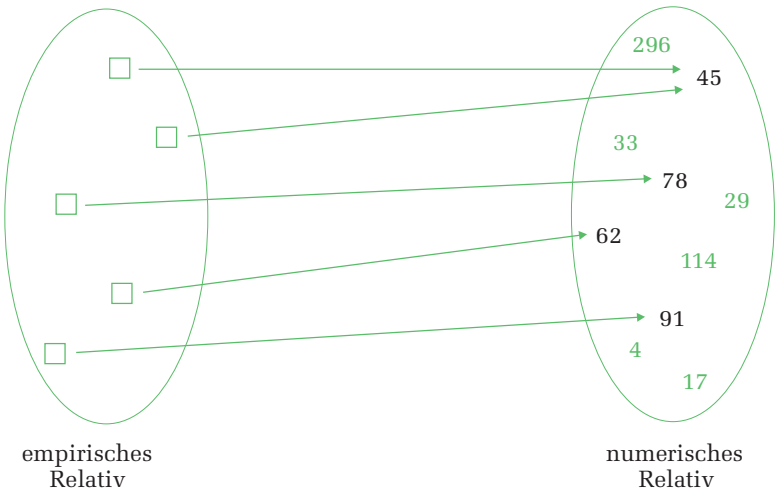


Abbildung 3.1: Abbildung eines empirischen Relativs in ein numerisches Relativ beim Messen.

Als *Skala* bezeichnet man das numerische Relativ (also eine Menge von Zahlen mit *bestimmten, definierten Relationen* zwischen diesen Zahlen), das aus einer homomorphen Abbildung resultiert. Aufgrund der Relationen, die im empirischen Relativ bestimmt werden können und die bei der Messung auch berücksichtigt werden, unterscheidet man verschiedene *Skalenniveaus*. Besteht im empirischen Relativ z.B. ledig-

lich eine Äquivalenzrelation und ist somit auch im numerischen Relativ nur die Gleichheitsrelation definiert, so misst man auf Nominalskalenniveau. Das Geschlecht von Personen wird also auf einer *Nominalskala* gemessen. Besteht im empirischen Relativ zusätzlich eine Ordnungsrelation, so ist im numerischen Relativ auch die Größer-Kleiner-Relation definiert. In diesem Fall misst man auf einer *Ordinalskala*.

3.2.1 Messtheoretische Probleme

Bei der Erarbeitung von homomorphen Abbildungen stellen sich der Messtheorie nun drei sogenannte Kardinalprobleme. Diese Probleme sind auch für die Einteilung der Skalenniveaus entscheidend. Sie werden im Folgenden kurz erläutert.

Das Repräsentationsproblem

Das Repräsentationsproblem betrifft die Frage, ob ein bestimmtes Merkmal überhaupt messbar ist. Diese Frage können wir in den Begriffen der Messtheorie auch folgendermaßen formulieren: Kann für ein bestimmtes empirisches Relativ eine homomorphe Abbildung in ein numerisches Relativ gefunden werden? Ein Merkmal ist dann messbar, wenn im empirischen Relativ bestimmte Axiome (Grundannahmen) erfüllt sind. Diese Axiome beziehen sich stets auf Eigenschaften der empirischen Relationen. Ein Beispiel für eine Eigenschaft einer empirischen Relation ist Transitivität. Diese Eigenschaft muss gegeben sein, damit ein Merkmal (mindestens) auf einer Ordinalskala messbar ist. Eine empirische Relation verfügt über die Eigenschaft der Transitivität, wenn Folgendes gilt: wenn $a \succ b$ und $b \succ c$ dann auch $a \succ c$. Solange wir an einfache physikalische Messungen denken, ist kaum einzusehen, wie dieses Axiom nicht erfüllt sein könnte. Messen wir etwa die Körpergröße dreier Personen, so wird niemand bezweifeln, dass Person A größer ist als Person C, wenn wir bereits wissen, dass Person A größer als Person B und Person B größer als Person C ist. Nehmen wir aber an, wir wollten die Spielstärke dreier Fußballteams messen. Zu diesem Zweck betrachten wir die Ergebnisse von Spielen zwischen diesen Teams. Das Team A hat das Team B geschlagen. Zudem hat Team B gegen das Team C gewonnen. Nun wäre es aller Erfahrung nach durchaus möglich, dass das Team A dennoch gegen Team C verliert. Augenscheinlich bestünde in diesem Fall also keine „echte“ Ordnungsrelation zwischen den drei Teams hinsichtlich ihrer Spielstärke. Demgemäß kann diese Relation auch nicht ins numerische Relativ abgebildet werden. Bei einer Messung der Spielstärke der drei Teams wäre die Größer-Kleiner-Relation im numerischen Relativ also nicht definiert und das Merkmal könnte nicht auf einer Ordinalskala gemessen werden.

Dasselbe Problem könnte auch bei unserer Messung der Präferenz einer Kundin für verschiedene Handymodelle auftreten. Im Beispiel hatten wir die Kundin gebeten, die Handys in eine Rangreihe zu bringen. Dass die Kundin dieser Bitte gefolgt ist und angegeben hat, welches Handy ihr am besten, am zweitbesten usw. gefällt, zeigt aber noch nicht, dass tatsächlich eine empirische Relation besteht, die die Eigenschaft der Transitivität aufweist. Um dies zu überprüfen, könnten wir auch hier Paarvergleiche vornehmen. Wir müssten der Kundin also jeweils Paare mit zwei Handys vorlegen und sie auffordern anzugeben, welches der beiden Handys ihr besser gefällt. Bei solchen Paarvergleichen geben Menschen nun unter Umständen durchaus intransitive Urteile ab. Die Kundin könnte uns also mitteilen, dass sie Handy A gegenüber Handy B bevorzugt und dass ihr Handy B besser gefällt als Handy C. Dennoch würde sie mög-

licherweise Handy C wählen, wenn es mit Handy A verglichen wird. Eine Erklärung für ein solches intransitives Urteil könnte etwa darin bestehen, dass die Kundin ihre Entscheidung bei den Paaren A und B sowie B und C hauptsächlich aufgrund des Preises der Handys traf, beim Vergleich der Handys A und C ihr Augenmerk aber auf das Design der Handys legte. Fänden wir tatsächlich intransitive Urteile im Zuge dieser Paarvergleiche, so wäre auch das Merkmal Präferenz für verschiedene Handymodelle nicht auf einer Ordinalskala messbar.

Zur Lösung des Repräsentationsproblems werden also zunächst Axiome formuliert, die im empirischen Relativ gelten sollen. Es sollte dann empirisch überprüft werden, ob diese Axiome tatsächlich erfüllt sind. Verläuft diese Überprüfung erfolgreich, so existiert eine homomorphe Abbildung des empirischen Relativs in ein numerisches Relativ. Das entsprechende Merkmal ist also (auf einem bestimmten Skalenniveau) messbar.

Allerdings wird die Forderung der Messtheorie nach einer empirischen Überprüfung der Axiome in der sozialwissenschaftlichen Forschungspraxis oftmals nicht erfüllt. Dies hängt damit zusammen, dass eine empirische Prüfung der Axiome in aller Regel sehr aufwendig ist. Bei zahlreichen psychologischen Messungen ist sie zudem auch kaum möglich. Psychologische Messungen beziehen sich oftmals auf latente Variablen, die nicht direkt beobachtbar sind (*Abschnitt 2.3.1*) – Beispiele wären etwa die Variablen „Intelligenz“ und „Extraversion“. Anhand welcher Kriterien könnten wir eindeutig und unstrittig entscheiden, welche von zwei Personen extravertierter ist? Wenn wir keine sichere Antwort auf diese Frage haben, lässt sich natürlich auch nicht verbindlich prüfen, ob hinsichtlich des Merkmals Extraversion Transitivität besteht. An die Stelle von empirischen Prüfungen der mit einem bestimmten Skalenniveau verbundenen Axiome treten daher häufig Plausibilitätsüberlegungen. Letztlich sind zahlreiche Messungen in der Psychologie *„per-fiat“* Messungen: Man „vertraut“ darauf, dass ein Messinstrument das jeweilige Merkmal auf einem bestimmten Skalenniveau erfasst. Findet man auf diese Weise konsistente und plausible Forschungsergebnisse, so spricht dies dafür, dass auch das Vertrauen in die Messprozedur gerechtfertigt war. Allerdings führt das Fehlen einer empirischen Prüfung der von der Messtheorie formulierten Axiome dazu, dass das Skalenniveau einer Messung vielfach nicht unzweifelhaft bestimmt werden kann. Entsprechend gibt es in der Psychologie durchaus immer wieder Diskussionen darüber, welches Skalenniveau durch eine bestimmte Messprozedur erreicht wird.

Das Eindeutigkeitsproblem

Mit der Lösung des Repräsentationsproblems wird zunächst nur ausgesagt, dass *mindestens* eine Möglichkeit besteht, eine Variable zu messen. In der Regel gibt es aber viele verschiedene Möglichkeiten, den Messobjekten so Zahlen zuzuordnen, dass die empirischen Relationen auch in den Messwerten zum Ausdruck kommen. Es stellt sich also die Frage, wie Messwerte verändert (transformiert) werden können, ohne dass die in ihnen enthaltene Information verloren geht. Diese Frage umreißt das sogenannte Eindeutigkeitsproblem. Beispielsweise informiert uns eine Messung der Variable Länge auch über Verhältnisse zwischen den Messobjekten. Stellen wir fest, dass die Körpergröße zweier Personen 2,00 m und 1,60 m beträgt, so wissen wir, dass die eine Person 1,25 mal so groß ist wie die andere. Nun können wir die Messwerte natürlich auch mit 100 multiplizieren und die Körpergröße in cm ausdrücken. Zwi-

schen den Messwerten 200 cm und 160 cm besteht dann nach wie vor ein Verhältnis von 1,25. Eine Multiplikation mit einer konstanten (positiven) Zahl ist im Falle der Längenmessung also eine zulässige *Transformation*. Unzulässig wäre dagegen die Addition einer Zahl. Fügen wir etwa zu beiden Messwerten 100 hinzu, so ändert sich offensichtlich das Verhältnis zwischen den resultierenden Zahlen. Anders stellt sich die Situation bei Messungen einer Variablen auf Ordinalskalenniveau dar. Hier informieren uns die Messwerte lediglich darüber, bei welchem der Messobjekte das fragliche Merkmal stärker ausgeprägt ist. Erhalten wir bei einer solchen Messung die Werte 4, 3, 2 und 1, so können wir die Werte quadrieren, sie mit irgendeiner positiven Zahl multiplizieren oder zu ihnen irgendeine positive Zahl addieren. Stets wird die Rangordnung der Messwerte auch in den resultierenden Zahlen erhalten bleiben (prüfen Sie selbst!). Die Menge der zulässigen Transformationen ist also bei Messungen auf Ordinalskalenniveau größer als bei Messungen der Länge (diese Messungen erfolgen auf dem *Verhältnisskalenniveau*). Skalen zur Messung der Länge sind daher „eindeutiger" als Skalen, die lediglich ordinale Informationen über ein Merkmal erfassen.

Das Bedeutsamkeitsproblem

Beim Bedeutsamkeitsproblem geht es um die Frage, welche mathematischen Operationen mit Messwerten zu empirisch sinnvollen Aussagen führen. Haben wir für verschiedene Messobjekte einmal Messwerte bestimmt, so hindert uns zunächst nichts daran, diese Messwerte beliebig zu verrechnen. Messen wir etwa die Variable Geschlecht, indem wir Männern eine 1 und Frauen eine 2 zuordnen, so besteht rein arithmetisch selbstverständlich die Möglichkeit, die Messwerte eines Mannes und einer Frau zu addieren. Allerdings ist dieser Vorgang offensichtlich sinnlos, da ihm empirisch nichts entspricht. Das Ergebnis dieser Addition erlaubt uns daher auch keinerlei Aussage über die Messobjekte. Die Addition nominalskalierter Messwerte führt also nicht zu *bedeutsamen* Aussagen.

Generell ist eine bestimmte Verrechnung von Messwerten dann sinnvoll, wenn sie unter allen zulässigen Transformationen der Messwerte zu derselben Aussage führt. Die Lösung des Bedeutsamkeitsproblems hängt also eng mit dem Eindeutigkeitsproblem zusammen. Betrachten wir hierzu ein Beispiel: Nehmen wir noch einmal an, es wäre uns gelungen, die Präferenz unserer Kundin für verschiedene Handymodelle auf Ordinalskalenniveau zu messen. Nehmen wir zudem an, wir würden die Messwerte 1 und 2 zweier Handys addieren. Diese Addition der Messwerte könnte uns zu der Aussage verleiten, dass die gemeinsame Präferenz für die beiden Handys ebenso stark ist wie die Präferenz für das Handy mit dem Messwert 3. Diese Aussage klingt etwas merkwürdig und sie ist tatsächlich auch nicht gerechtfertigt. Wie wir bereits gesehen haben, bestünde eine zulässige Transformation dieser ordinalskalierten Messwerte darin, sie zu quadrieren. Nach dieser Transformation erhalten wir für die drei Handys die Messwerte 1, 4 und 9. Addieren wir nun nochmals die Messwerte der ersten beiden Handys, so kommen wir auf 5. Demnach wäre die gemeinsame Präferenz für diese Handys jetzt schwächer als die Präferenz für das dritte Handy. Offensichtlich können nicht beide Aussagen korrekt sein. Auch bei ordinalskalierten Messwerten führt eine Addition also nicht zu bedeutsamen Aussagen. Dies ist beispielsweise bei verhältnisskalierten Daten anders. Offensichtlich sind zwei Bretter der Länge 1 m und 2 m gemeinsam ebenso lang wie ein Brett der Länge 3 m. Auf dem Verhältnisskalenniveau ist die Addition von Messwerten also sinnvoll.

Das Bedeutsamkeitsproblem hat wichtige Konsequenzen für die Frage, welche statistischen Verfahren bei der Analyse der in einer empirischen Untersuchung erhobenen Daten angewandt werden können. Dies ist darin begründet, dass die Verrechnung von Messwerten innerhalb von statistischen Verfahren auch zu empirisch sinnvollen Aussagen führen muss. Jedes statistische Verfahren setzt daher ein bestimmtes Skalenniveau voraus. Die meisten der Verfahren, die in diesem Buch behandelt werden und denen in der sozialwissenschaftlichen Forschung die größte Bedeutung zukommt, erfordern Messwerte, die zumindest Intervallskalenniveau (siehe unten) erreichen. Verfahren, die auch bei nominal- und ordinalskalierten Daten eingesetzt werden können, werden insbesondere in den *Kapiteln 17* und *18* beschrieben.

3.3 Skalenniveaus

In den Sozialwissenschaften werden zumeist fünf Skalenniveaus unterschieden: Nominal-, Ordinal-, Intervall-, Verhältnis- und Absolutskala. Diese Klassifikation der Skalenarten geht auf Stevens (1951) zurück. Im Folgenden werden die verschiedenen Skalentypen beschrieben und ihre wichtigsten Eigenschaften zusammengefasst. Wir beginnen dabei mit dem niedrigsten Skalenniveau, der Nominalskala, und schreiten fort bis zum höchsten Skalenniveau, der Absolutskala. Die Skalenniveaus sind dabei nach ihrem Informationsgehalt geordnet. Messwerte auf einem höheren Skalenniveau erlauben also mehr sinnvolle Aussagen über die Messobjekte als Messwerte auf niedrigeren Niveaus.

3.3.1 Nominalskala

Messungen auf diesem Niveau setzen lediglich voraus, dass im empirischen Relativ eine Äquivalenzrelation besteht. Entsprechend beinhalten nominalskalierte Messwerte auch ausschließlich Information über die Gleichheit oder Verschiedenheit von Merkmalsausprägungen. Beispiele für Merkmale, die auf Nominalskalenniveau gemessen werden, sind die Blutgruppe, der Beruf, die Nationalität, das Geschlecht, die psychiatrische Diagnose und generell alle weiteren Kategorisierungen. Die Zuordnung von Zahlen zu Merkmalsausprägungen geschieht willkürlich. Messen wir etwa die Parteizugehörigkeit, so können wir den Ausprägungen SPD, CDU und FDP die Zahlen 1, 2 und 3 oder auch die Werte 27, 9 und 41 zuweisen, da es ausschließlich auf die Gleichheit und Ungleichheit der Messwerte ankommt. Entsprechend können nominalskalierte Daten fast beliebig transformiert werden. Wichtig ist lediglich, dass gleiche Merkmalsausprägungen erneut gleiche Messwerte erhalten und dass unterschiedlichen Ausprägungen abermals unterschiedliche Messwerte zugeordnet werden. Transformationen, die dieser Bedingung genügen, werden als *ein-eindeutige* Transformationen bezeichnet.

Da auf dem Nominalskalenniveau den unterschiedlichen Merkmalsausprägungen beliebige Zahlen zugeordnet werden können, ist es sinnlos, die entsprechenden Messwerte in irgendeiner Form zu verrechnen. Beispielsweise hätte es keinen Sinn aus nominalskalierten Daten einen Mittelwert zu berechnen (dass die Missachtung dieser Regel zu erstaunlichen Fehlern führen kann, illustriert das Beispiel im Kasten „Der ‚durchschnittliche' Unfallverursacher"). Statistische Verfahren zur Analyse solcher Daten nutzen daher auch ausschließlich Informationen über die Häufigkeit, mit der verschiedene Merkmalsausprägungen aufgetreten sind. So könnten wir etwa denjeni-

gen Messwert ermitteln, der in einem Datensatz am häufigsten enthalten ist – den *Modalwert* (siehe auch *Kapitel 6*). Wir könnten also z.B. festhalten, dass wir in einer Studentenkneipe bei der Messung des Studienfachs am häufigsten den Messwert 3 beobachtet haben, der vielleicht das Fach Jura anzeigt.

Bei Merkmalen, die höchstens auf einer Nominalskala gemessen werden können, zeigen unterschiedliche Messwerte keine quantitativen Unterschiede zwischen den Messobjekten an (bei einer Messung des Merkmals Geschlecht ist ein Mann natürlich nicht mehr oder weniger als eine Frau). Derartige Merkmale werden daher auch als *qualitative Variablen* bezeichnet. Merkmale, die auf einem höheren Skalenniveau gemessen werden können, werden auch *quantitative Variablen* genannt.

Der „durchschnittliche" Unfallverursacher

Ein in Statistiklehrbüchern vielfach zitiertes Beispiel, das demonstriert, dass das Skalenniveau von Daten bei der statistischen Analyse unbedingt beachtet werden sollte, stammt aus einer US-amerikanischen Studie über Unfallursachen. In dieser Untersuchung wurde erfasst, wer an einem Unfall schuld war, wobei ausschließlich die Merkmale Hautfarbe und Geschlecht der Unfallverursacher berücksichtigt wurden. Die verschiedenen Merkmalskombinationen wurden folgendermaßen kodiert: 0 = männlich und weiß, 1 = männlich und farbig, 2 = weiblich und weiß, 3 = weiblich und farbig. Aufgrund dieser Kodierung wurde aus allen Messwerten in der Stichprobe der Mittelwert berechnet. Dieser lag im Bereich von 1,0. Daraus wurde der Schluss gezogen, dass es sich bei dem typischen Unfallverursacher um einen männlichen Farbigen handelt.

Diese Schlussfolgerung entsprach zwar möglicherweise den Erwartungen des Autors der Studie, es sollte jedoch klar sein, dass sie keinesfalls gerechtfertigt ist. Die genaue Häufigkeit, mit der Männer und Frauen sowie Weiße und Farbige in dieser Untersuchung als Unfallverursacher identifiziert wurden, ist nicht bekannt. Einen Mittelwert von exakt 1,0 würden wir jedoch beispielsweise erhalten, wenn in der Studie 100 Unfälle erfasst wurden, von denen 40 von männlichen Weißen, 30 von männlichen Farbigen, 20 von weiblichen Weißen und 10 von weiblichen Farbigen verursacht wurden ($(40 \cdot 0 + 30 \cdot 1 + 20 \cdot 2 + 10 \cdot 3) : 100 = 1{,}0$). Nun sind die Merkmale Geschlecht und Hautfarbe nominalskaliert. Alle eindeutigen Transformationen der Messwerte sind also zulässig. Wie würde sich der Mittelwert der Messwerte ändern, wenn wir beispielsweise die Kodierung für männliche Weiße und Farbige vertauschen würden (0 = männlich und farbig, 1 = männlich und weiß)? In diesem Fall betrüge der Mittelwert 1,1 ($(30 \cdot 0 + 40 \cdot 1 + 20 \cdot 2 + 10 \cdot 3) : 100 = 1{,}1$). Mit dieser Kodierung kämen wir anhand des Mittelwerts also zu dem Schluss, dass es sich bei dem typischen Unfallverursacher um einen männlichen Weißen handelt. Offensichtlich führt die Berechnung des Mittelwerts aus nominalskalierten Daten also nicht zu bedeutsamen Aussagen.

3.3.2 Ordinalskala

Messungen auf diesem Niveau erfordern, dass im empirischen Relativ eine (schwache) Ordnungsrelation besteht.[1] Wir müssen also empirisch feststellen können, ob ein bestimmtes Messobjekt eine stärkere, schwächere oder genauso große Merkmalsausprägung hat wie ein anderes Messobjekt. Genau diese Information wird dann auch

1 Die schwache Ordnungsrelation beinhaltet die Möglichkeit, dass zwei Objekte die gleiche Merkmalsausprägung aufweisen.

durch ordinalskalierte Messwerte zum Ausdruck gebracht. Ordinalskalierte Daten erlauben somit noch keine Aussage über die Größe des Unterschieds zwischen zwei Messobjekten.

Ein Beispiel sind die Single-Charts, mit denen der Verkaufserfolg von CDs auf Ordinalskalenniveau gemessen wird. Dabei wird der CD mit dem größten Verkaufserfolg bekanntermaßen nicht der größte, sondern der kleinste Messwert (die 1) zugewiesen. Dies ist unerheblich, solange bekannt ist, ob kleinere oder größere Zahlen stärkere Merkmalsausprägungen anzeigen. Weitere Beispiele sind alle Arten von Rangreihen, wie etwa militärische Ränge oder Tabellenplätze im Sport. Auch Schulnoten werden häufig als ein Beispiel für eine Ordinalskala angeführt.[2] Demnach würden uns die Mathematiknoten 1, 2 und 3 dreier Schüler darüber informieren, dass der Schüler mit der 1 über die größten Mathematikkenntnisse verfügt. Wir wüssten aber nicht, ob der Unterschied zwischen dem Schüler mit der Note 1 und dem Schüler mit der Note 2 ebenso groß ist wie der Unterschied zwischen den Schülern mit den Noten 2 und 3.

Ordinalskalen können auch entstehen, indem quantitativ geordnete Merkmalsausprägungen zu (unterschiedlich großen) Klassen zusammengefasst werden, denen jeweils der gleiche Messwert zugeordnet wird. Ein Beispiel für diese Vorgehensweise liefert die Beaufort-Skala zur Messung der Windstärke. Hier wird Windgeschwindigkeiten von 6 km/h – 11 km/h der Messwert 2 zugewiesen, Windgeschwindigkeiten von 20 km/h – 28 km/h erhalten den Messwert 4 und Geschwindigkeiten zwischen 39 km/h und 49 km/h entspricht die Windstärke 6.[3] Messen wir an drei aufeinander folgenden Tagen die Windstärken 2, 4 und 6, so bedeutet auch dies nicht, dass der Unterschied zwischen der Windgeschwindigkeit am ersten und am zweiten Tag ebenso groß ist wie der Unterschied zwischen der Geschwindigkeit am zweiten und am dritten Tag. Das Beispiel illustriert auch, dass das Skalenniveau, das bei einer Messung erreicht wird, nicht nur davon abhängt, welche empirische Relationen zwischen den Messobjekten bestehen, sondern auch davon, welche Relationen bei der Messprozedur tatsächlich festgestellt und ins numerische Relativ abgebildet werden. Offensichtlich könnten wir bei der Messung des Windes anhand der Windgeschwindigkeit in km/h auch Aussagen über Größenunterschiede treffen. Die Beaufort-Skala berücksichtigt diese Information über Größenunterschiede allerdings nicht.

Da Ordinalskalen lediglich Informationen über die Rangordnung der Messobjekte liefern, sind alle Transformationen zulässig, die die Rangreihe der Messwerte erhalten. Dies sind alle monoton steigenden Transformationen. Genau wie bei nominalskalierten Daten ist es auch bei ordinalskalierten Daten nicht sinnvoll, einen Mittelwert zu berechnen. Rechnerisch beträgt die mittlere Windstärke an drei Tagen mit den Windstärken 2, 4 und 9 offensichtlich 5 ((2 + 4 + 9) : 3 = 5). Dies heißt allerdings *nicht*, dass die durchschnittliche Windgeschwindigkeit an diesen Tagen ebenso groß war wie die Windgeschwindigkeit an einem Tag mit dem Messwert 5. Eine sinnvolle Aus-

2 Das Skalenniveau von Schulnoten ist allerdings umstritten. In der Praxis geht man meist davon aus, dass Schulnoten Intervallskalenniveau (siehe *Abschnitt 3.3.3*) erreichen.

3 Selbstverständlich muss man nicht zunächst Windgeschwindigkeiten in km/h ermitteln, um die Windstärke auf der Beaufort-Skala angeben zu können – andernfalls wäre diese Skala nutzlos. Tatsächlich sind die verschiedenen Windstärken durch „Erscheinungsbilder" charakterisiert, anhand derer die Messwerte bestimmt werden. Dem Erscheinungsbild „Wind im Gesicht fühlbar" wird z.B. der Messwert 2 zugeordnet, dem Erscheinungsbild "Staub und Papier werden verweht" entspricht die Windstärke 4.

sage über ordinalskalierte Daten kann aber beispielsweise getroffen werden, indem man den *Median* bestimmt. Der Median ist derjenige Wert, für den gilt, dass 50% aller Messwerte kleiner (oder gleich) und 50% aller Messwerte größer (oder gleich) sind (siehe auch *Kapitel 6*). Im obigen Beispiel mit den Windstärken beträgt der Median demnach 4. Ebenso wie alle anderen statistischen Verfahren, die zur Analyse von ordinalskalierten Daten geeignet sind, nutzt der Median also ausschließlich Rangsinformationen.

3.3.3 Intervallskala

Messungen auf dem Intervallskalenniveau erfordern, dass die Größe des Unterschieds zwischen verschiedenen Merkmalsausprägungen empirisch ermittelt werden kann. Die Messwerte werden den Merkmalsausprägungen dann so zugeordnet, dass gleich große Unterschiede zwischen Messwerten auch gleich große Unterschiede zwischen Merkmalsausprägungen anzeigen. Bei Messungen auf dem Intervallskalenniveau wird also eine Maßeinheit definiert. Intervallskalierte Messwerte erlauben jedoch noch keine Aussage über Verhältnisse zwischen Messwerten. Dies liegt daran, dass Intervallskalen über keinen absoluten Nullpunkt verfügen. Der Messwert 0 wird also willkürlich festgelegt und besagt nicht, dass ein Merkmal nicht vorhanden ist.

Das klassische Beispiel für eine Intervallskala ist die Celsius-Temperaturskala. Hier ist der Temperaturunterschied zwischen 5 °C und 10 °C genau so groß wie derjenige zwischen 20 °C und 25 °C. Allerdings bedeuten 0 °C nicht, dass keine Temperatur vorhanden ist, und der Messwert 0 könnte auch irgendeiner anderen Temperatur zugeordnet werden. Aufgrund dieser Beliebigkeit des Nullpunktes ist es falsch zu behaupten, 20 °C seien doppelt so warm wie 10 °C.

Für intervallskalierte Daten sind alle linearen Transformationen zulässig. Dies sind Transformationen der Form $y = a \cdot x + b$. Beispielsweise können wir Messwerte in Celsius nach folgender Formel in Messwerte in Fahrenheit umrechnen:

$$F = 1{,}8 \cdot C + 32$$

Dabei wird durch die Multiplikation mit 1,8 die Einheit der Skala verändert: Ein Temperaturzuwachs von 1 °C entspricht einem Temperaturzuwachs von 1,8 °F. Die Addition des zweiten Terms (+ 32) verändert den Nullpunkt der Skala. 0 °C entsprechen also 32 °F.

Auf dem Intervallskalenniveau ist die Berechnung eines Mittelwerts sinnvoll. Berechnen wir etwa die durchschnittliche Höchsttemperatur einiger Sommertage, so entspricht der Mittelwert der Höchsttemperatur an diesen Tagen tatsächlich der Temperatur an einem Tag mit demselben Messwert. Generell können auf dem Intervallskalenniveau (und den höheren Skalenniveaus) alle in der Psychologie und den Sozialwissenschaften gängigen statistischen Verfahren sinnvoll angewandt werden.

In den Sozialwissenschaften wird für zahlreiche „typische“ Messungen angenommen, dass sie das Niveau einer Intervallskala erreichen. So gelten IQ-Werte (die Ergebnisse von Intelligenztests) ebenso als intervallskaliert wie die Messwerte vieler anderer psychologischer Tests. Eine andere, in den Sozialwissenschaften häufig genutzte Technik der Datenerhebung besteht in der Verwendung sogenannter Rating-Skalen

(▶Abbildung 3.2). Auf solchen Rating-Skalen können Probanden beispielsweise angeben, ob und wie stark sie eine vorgegebene Aussage für zutreffend halten oder wie sehr sie einer bestimmten Meinung zustimmen.

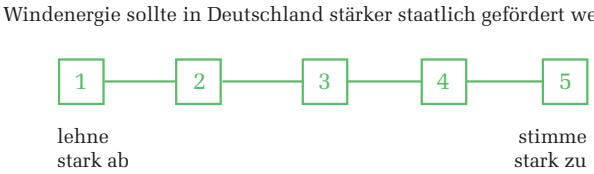


Abbildung 3.2: Ein Beispiel für eine Rating-Skala.

Auch Messungen mit solchen Rating-Skalen werden zumeist als intervallskaliert angesehen. Demnach würde zwischen den Messwerten 2 und 3 ein ebenso großer Unterschied in der Zustimmung zu der Aussage über die Windenergie bestehen wie zwischen den Messwerten 4 und 5. Nun kann man natürlich bezweifeln, dass Probanden ihren subjektiven Eindruck vom Ausmaß ihrer Ablehnung oder Zustimmung zu der Aussage tatsächlich in intervallskalierte Urteile umsetzen können. Entsprechend gab und gibt es heftige Debatten um die Frage, ob Messungen mit Rating-Skalen intervallskaliert sind oder doch nur das Niveau einer Ordinalskala erreichen. Dies zeigt, dass es schwierig und problematisch sein kann, das Skalenniveau einer Messung zu bestimmen. Dass sich in den Sozialwissenschaften überwiegend die Auffassung durchgesetzt hat, Messungen mit Rating-Skalen seien intervallskaliert, hat wohl hauptsächlich pragmatische Gründe. Zum einen stehen für die Auswertung von intervallskalierten Daten mehr und aussagekräftigere statistische Verfahren zur Verfügung. Zum anderen gelangt man in der Forschung mit Rating-Skalen oftmals auch dann zu sinnvollen Ergebnissen, die sich in der Praxis bewähren, wenn man das (höhere) Intervallskalenniveau unterstellt.

Messungen auf einem höheren Skalenniveau als dem der Intervallskala sind in der Psychologie eher selten. Ein Grund dafür besteht darin, dass sich bei psychischen Merkmalen in der Regel kein inhaltlich sinnvoller Nullpunkt angeben lässt. So können wir über eine Person, die in einem Intelligenztest keine Aufgabe löst, natürlich nicht sagen, dass sie über keine Intelligenz verfügt. Folglich ist ein Testteilnehmer, der zehn Aufgaben löst, auch nicht doppelt so intelligent wie ein Teilnehmer der fünf Aufgaben löst.

3.3.4 Verhältnisskala

Messungen auf dem Verhältnisskalenniveau setzen voraus, dass nicht nur die Größe des Unterschieds zwischen verschiedenen Merkmalsausprägungen empirisch ermittelt werden kann, sondern dass auch ein inhaltlich bedeutungsvoller Nullpunkt bestimmbar ist. Verhältnisskalen ordnen diesem Nullpunkt dann auch den Messwert Null zu (anders als etwa die Celsius-Skala bei der Temperaturmessung). Damit erlauben Messwerte auf diesem Skalenniveau auch Aussagen über Verhältnisse zwischen verschiedenen Merkmalsausprägungen.

Beispiele für Verhältnisskalen finden sich vielfach in der Physik. Länge, Zeit, Gewicht werden auf Verhältnisskalen gemessen. Eine Verhältnisskala zur Messung der Temperatur ist die Kelvinskala. Ein anderes Beispiel für ein Merkmal, das auf einer Verhältnisskala gemessen wird, ist das Monatseinkommen.

Die Einheiten einer Verhältnisskala sind nicht festgelegt. Wir können die Länge verschiedener Tische in Meter, Zentimeter oder auch Inch angeben, ohne dass sich das Verhältnis zwischen den Messwerten der Tische ändert. Verhältnisskalen sind somit eindeutig bis auf hier zulässige Ähnlichkeitstransformationen der Form $y = a \cdot x$. Ein Beispiel für eine solche Transformation ist die Umrechnung einer Längenangabe in Inch in eine Angabe in Zentimeter nach der Formel:

$$\text{cm} = 2{,}54 \cdot \text{in}$$

Psychische Merkmale wie Intelligenz, Konzentrationsfähigkeit oder Neurotizismus können nicht auf Verhältnisskalen gemessen werden. Dies bedeutet aber nicht, dass verhältnisskalierte Merkmale in der Psychologie grundsätzlich keine Rolle spielen. Insbesondere die Variable Zeit wird häufig in psychologischen Untersuchungen erfasst. Psychologen könnten sich etwa für die Dauer von Therapien, die Reaktionszeit bei verschiedenen Warnsignalen oder die Bearbeitungsdauer bei einer bestimmten Aufgabe interessieren. Allerdings ist bei Größen wie Zeit, Länge oder Einkommen zu unterscheiden, ob tatsächlich diese Variablen selbst gemessen werden sollen oder ob sie lediglich als Indikatoren für andere Merkmale dienen. Nehmen wir an, wir wollten den sozio-ökonomischen Status verschiedener Personen messen. Der sozio-ökonomische Status ist nicht direkt beobachtbar. Zur Messung dieses Merkmals müssen wir also zunächst eine beobachtbare Variable finden, die als Indikator verwendet werden kann. Eine gängige Operationalisierung für den sozio-ökonomischen Status ist das Jahreseinkommen, das auf einer Verhältnisskala gemessen werden kann. Allerdings wäre es sicher falsch zu behaupten, dass eine Person, die ein Jahreseinkommen von Null hat, über keinen sozio-ökonomischen Status verfügt. Somit messen wir den sozio-ökonomischen Status mithilfe des Jahreseinkommens nicht auf Verhältnisskalenniveau. Es muss auch bezweifelt werden, dass wir den Status auf einer Intervallskala erfassen. Gleiche Einkommensunterschiede zeigen nämlich nicht zwangsläufig gleiche Statusunterschiede an. Der Unterschied zwischen 10.000 € und 30.000 € Jahreseinkommen indiziert sicherlich einen bedeutsamen Statusunterschied. Hingegen wird sich der Status zweier Großverdiener mit Jahreseinkommen von 500.000 € und 520.000 € kaum unterscheiden. Der sozio-ökonomische Status wird durch das Jahreseinkommen also wohl nur auf Ordinalskalenniveau erfasst.

Da für die Psychologie hauptsächlich nicht beobachtbare Merkmale relevant sind, stellt sich dieses „Indikator-Problem“ regelmäßig. Bei der Bestimmung des Skalenniveaus einer Messung ist also stets danach zu fragen, ob die direkt beobachtete Variable selbst von Interesse war, oder ob sie lediglich als Operationalisierung eines anderen Merkmals verwendet wurde. Für die Bearbeitungsdauer bei einer bestimmten Aufgabe könnten wir uns etwa deswegen interessieren, weil wir feststellen wollen, wie sich die Bearbeitungsdauer mit zunehmender Übung verändert. In diesem Fall messen wir auf einer Verhältnisskala. Denkbar wäre aber auch, dass wir die Bearbeitungsdauer als Indikator für die Ausprägung einer intellektuellen Leistungskomponente verwenden. In diesem Fall erreicht unsere Messung dieser Leistungskomponente mithilfe der Bearbeitungsdauer sicher nicht das Niveau einer Verhältnisskala.

3.3.5 Absolutskala

Eine Absolutskala hat neben einem natürlichen Nullpunkt auch eine natürliche Maßeinheit. Dies ist immer dann der Fall, wenn Häufigkeiten erfasst werden. In der Psychologie begegnen uns Absolutskalen vor allem dann, wenn die Häufigkeit des Auftretens bestimmter Verhaltensweisen von Interesse ist. Die Häufigkeit, mit der sich ein Schulkind am Unterricht beteiligt, die Häufigkeit des Blickkontakts zwischen frisch Verliebten, die Anzahl der gerauchten Zigaretten oder auch die Zahl der Mitglieder einer Gruppe sind also Beispiele für Variablen, die auf einer Absolutskala gemessen werden.

Bei einer Absolutskala sind keine Transformationen zulässig, da hier sowohl der Nullpunkt als auch die Maßeinheit eindeutig festgelegt sind. Interessieren wir uns etwa für die Menge der täglich konsumierten Zigaretten, so liefert uns ausschließlich die konkrete Zahl der Zigaretten die Information, die wir benötigen.

▶Tabelle 3.1 zeigt die wichtigsten Eigenschaften der Skalenniveaus noch einmal im Überblick.

Tabelle 3.1

Eigenschaften der wichtigsten Skalenniveaus

Skala	Mögliche Aussage	Zulässige Transformationen	Beispiele	Lagemaße
Nominal	Gleichheit / Ungleichheit	ein-eindeutige	Studienort, Parteizugehörigkeit, Geschlecht	Modus
Ordinal	Größer-Kleiner-Relationen	monoton steigende	Single-Charts, Windstärke	+ Median
Intervall	Gleichheit von Differenzen	lineare $y = a \cdot x + b$	Temperatur in Celsius, IQ-Werte	+ arithmetisches Mittel
Verhältnis	Gleichheit von Verhältnissen	proportionale $y = a \cdot x$	Längenmaße, Temperatur in Kelvin, Einkommen	+ geometrisches Mittel
Absolut	zusätzlich: natürliche Maßeinheit	keine	Häufigkeiten	

3.4 Tests

Nachdem wir die Grundlagen des Messens erörtert haben, wollen wir nun einen Blick auf ein spezifisch psychologisches Messinstrument werfen, mit dem vermutlich die meisten von uns schon einmal in Berührung gekommen sind: psychometrische Tests. Derartige Tests sind standardisierte Verfahren zur Erfassung latenter Variablen (*Abschnitt 2.3.1*). Mit ihnen sollen also nicht direkt beobachtbare Merkmale von Personen gemessen werden. Psychometrische Tests bestehen stets aus einer Reihe von

Aufgaben oder Fragen, die häufig als „Items“ bezeichnet werden. Aus dem Antwortverhalten eines Probanden bei diesen Items wird dann auf die Ausprägung desjenigen Merkmals geschlossen, das gemessen werden soll. Das Antwortverhalten bildet hier also den beobachtbaren Indikator der interessierenden latenten Variablen. Die vielfältigen verschiedenen Testverfahren können in zwei große Gruppen unterteilt werden: Leistungstests und Persönlichkeitstests. Leistungstests, zu denen auch alle Intelligenztests zählen, bestehen aus Aufgaben, bei denen objektiv festgestellt werden kann, ob die Antwort richtig oder falsch ist. Zwei Beispiele für solche Items können wir dem Intelligenz-Struktur-Test (I-S-T 2000 R; Amthauer et al., 2001) entnehmen.

Beispiel-Items aus einem Leistungstest

- Item 1:

  Treppe : Leiter = Haus : ?

  a) Dach b) Hof c) Aufzug d) Wand e) Zelt

- Item 2:

  18 16 19 15 20 14 21

Bei dem ersten Item sollen die Teilnehmer eine Analogie finden. Aus den Antwortvorgaben in der zweiten Reihe soll dasjenige Wort ausgewählt werden, zu dem sich „Haus“ ebenso verhält, wie sich „Treppe“ zu „Leiter“ verhält. Die richtige Lösung ist also „Zelt“. Bei dem zweiten Item sollen die Teilnehmer diejenige Zahl angeben, mit der die Zahlenreihe sinnvoll fortgesetzt werden kann. Die richtige Lösung ist hier „13“.

Mit Persönlichkeitstests werden Merkmale wie Verträglichkeit, Offenheit oder Neurotizismus gemessen. Bei der Messung derartiger Merkmale spielt die objektiv richtige oder falsche Lösung von Aufgaben keine Rolle. Stattdessen geben die Probanden hier Selbstbeschreibungen ab. Dazu werden ihnen Fragen oder Aussagen vorgelegt, die sie bejahen oder verneinen oder denen sie in unterschiedlich starkem Ausmaß zustimmen. Zwei Beispiele:

Beispiel-Items aus Persönlichkeitstests

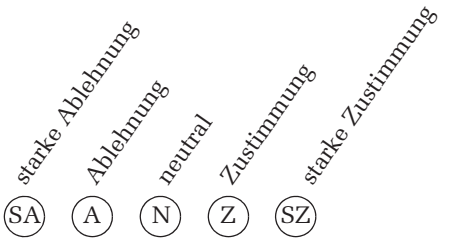
- Item 1:

  Ich bin im Grunde eher ein ängstlicher Mensch...

  stimmt ○ stimmt nicht ○

- Item 2:

  Ich habe Schwierigkeiten, meinen Begierden zu widerstehen...

Das erste Item stammt aus dem Freiburger-Persönlichkeits-Inventar (FPI-R; Fahrenberg, Hampel & Selg, 2001) und wird zur Messung des Merkmals Gehemmtheit eingesetzt. Das zweite Item gehört zu einer Gruppe von Items, mit denen im NEO-Persönlichkeitsinventar (NEO-PI-R; Ostendorf & Angleitner, 2004) das Merkmal Neurotizismus gemessen wird.

Unabhängig davon, ob es sich um einen Leistungs- oder Persönlichkeitstest handelt, wird bei der Auswertung eines Tests zunächst anhand der Antworten eines Probanden ein Rohwert ermittelt. Dieser Rohwert entspricht zumeist der Anzahl der richtigen Lösungen bzw. der „Ja"- oder „Stimmt"-Antworten bei allen Items, die dasselbe Merkmal messen sollen. Werden in einem Test Rating-Skalen verwendet, wie bei dem obigen Beispiel-Item aus dem NEO-PI-R, so sind den Antworten gegebenenfalls zunächst nach einer bestimmten Vorschrift Zahlen zuzuordnen. Der Rohwert eines Probanden ergibt sich dann in aller Regel, indem diese Zahlen über mehrere Items summiert werden. Der Rohwert wird dann wiederum mithilfe von Normen in einen sogenannten Testwert – z.B. einen IQ-Wert – umgerechnet. Diese Normen resultieren aus Untersuchungen mit sogenannten Eichstichproben, in denen eine große Anzahl von Teilnehmern den Test bearbeitet. Aus diesen Untersuchungen mit Eichstichproben ist unter anderem bekannt, wie viele Items in einem Test durchschnittlich richtig gelöst oder mit „Ja" beantwortet werden. Der Testwert eines Probanden ergibt sich nun (zumindest bei der überwiegenden Mehrzahl der Tests) aus einem Vergleich seines Rohwerts mit der durchschnittlichen Anzahl richtiger Lösungen oder „Ja"-Antworten. Ein IQ-Wert von 100 besagt beispielsweise nichts anderes, als dass der Proband mit diesem Testwert eine durchschnittliche Anzahl richtiger Lösungen erzielt hat. Hat der Proband die durchschnittliche Anzahl richtiger Lösungen mehr oder weniger deutlich übertroffen, so erhält er einen Testwert, der entsprechend deutlich über 100 liegt. Löste er eine unterdurchschnittliche Anzahl an Aufgaben, so wird ihm natürlich ein Testwert unter 100 zugeordnet.

Nun ist es offensichtlich, dass nicht jede beliebige Zusammenstellung von Aufgaben geeignet sein kann, um Intelligenz zu messen, und dass nicht alle denkbaren Sammlungen von Selbstbeschreibungen zu einer brauchbaren Messung des Merkmals Neurotizismus führen. Die Konstruktion und Auswahl von Test-Items muss also bestimmten Regeln unterliegen. Mit diesen Regeln beschäftigen sich Testtheorien. Die große Mehrzahl der heute gebräuchlichen Tests basiert auf der historisch ältesten dieser Theorien, die heute als *Klassische Testtheorie* bezeichnet wird (eine Einführung in die Klassische Testtheorie findet man z.B. bei Bühner, 2010). Im Kontext des Themas „Messen" ist die klassische Testtheorie für uns vor allem deswegen interessant, weil sie mithilfe bestimmter Kriterien beurteilt, ob und wie gut ein Test geeignet ist, ein bestimmtes Merkmal zu erfassen. Diesen Gütekriterien sollte aber auch jede beliebige andere Messung genügen. Sie können also ganz generell als ein „Standard" aufgefasst werden, den „gute" sozialwissenschaftliche Messungen erfüllen sollten.

3.5 Gütekriterien beim Testen und Messen

Man unterscheidet drei sogenannte Hauptgütekriterien: Objektivität, Reliabilität und Validität. Diese Gütekriterien bauen in bestimmter Hinsicht aufeinander auf: Hohe Reliabilität kann nicht erreicht werden, wenn der Test nicht objektiv ist. Reliabilität ist wiederum eine Voraussetzung für Validität. Ein Test kann also durchaus reliabel

sein und dennoch das Kriterium der Validität nicht oder nur schlecht erfüllen. Der umgekehrte Fall ist hingegen ausgeschlossen: Ein hoch valider Test ist stets auch reliabel (und objektiv).

3.5.1 Objektivität

Selbstverständlich sollte das Ergebnis einer Messung nicht durch die Person beeinflusst werden, die das jeweilige Messinstrument anwendet. Genügt ein Messinstrument dieser Anforderung, so ist es objektiv. Ein Test ist also völlig objektiv, wenn verschiedene Testleiter bei demselben Probanden das gleiche Ergebnis erzielen. Wird das Ergebnis eines Tests jedoch durch das Verhalten des Testleiters bei der Durchführung oder durch seine individuellen Deutungen der Antworten des Probanden beeinflusst, so ist der Test nicht objektiv. Man kann drei Aspekte der Objektivität eines Tests differenzieren: Durchführungs-, Auswertungs- und Interpretationsobjektivität.

Durchführungsobjektivität

Die Durchführungsobjektivität betrifft die Frage, inwieweit die Testergebnisse von Verhaltensvariationen des Untersuchers während der Testdurchführung unabhängig sind. Eine Beeinträchtigung der Durchführungsobjektivität bestünde etwa dann, wenn verschiedene Testleiter den Probanden unterschiedlich verständliche Erläuterungen zu den Testaufgaben geben. Um eine hohe Durchführungsobjektivität zu gewährleisten, enthalten die meisten psychometrischen Tests präzise Anweisungen für den Testleiter, die festlegen, wie er sich während der Durchführung verhalten soll. Zum Beispiel sind die Instruktionen für die Probanden in der Regel wörtlich vorgegeben und müssen vom Testleiter lediglich vorgelesen werden. Darüber hinaus wird oftmals große Mühe darauf verwandt, die Instruktionen so verständlich und eindeutig zu formulieren, dass Rückfragen an den Testleiter nicht notwendig sind. Um die Durchführungsobjektivität nicht zu gefährden, werden soziale Interaktionen zwischen Testleiter und Probanden in solchen Tests also auf ein Minimum reduziert.

Auswertungsobjektivität

Ein Test erfüllt die Forderung nach Auswertungsobjektivität dann, wenn verschiedene Anwender aufgrund der Antworten eines Probanden zu demselben Testergebnis gelangen. Bei den meisten psychometrischen Tests stellt die Auswertungsobjektivität kein Problem dar. Da die Probanden lediglich zwischen verschiedenen vorgegebenen Antwortoptionen wählen und zudem in der Testanweisung eindeutig festgelegt wird, wie eine Antwort zu bewerten ist, hat der Anwender bei der Auswertung der Antworten keinerlei Spielraum. Allerdings gibt es auch Tests, die offene Fragen enthalten, bei denen die Probanden ihre Antwort frei formulieren können. Hier muss die Bewertung der Antworten dann vom Testanwender vorgenommen werden. Dies gefährdet die Auswertungsobjektivität. In diesem Fall sollte ein Test möglichst umfassende und klare Anweisungen enthalten, in denen definiert wird, welche freien Antworten als richtig zu bewerten sind bzw. welche Antworten eine größere Ausprägung des Merkmals, das gemessen werden soll, anzeigen.

Interpretationsobjektivität

Die Interpretationsobjektivität betrifft die Frage, ob verschiedene Anwender aus demselben Testergebnis die gleichen Schlüsse ziehen. Offensichtlich wäre z.B. ein Intelligenztest nutzlos, bei dem dasselbe Testergebnis von einem Psychologen als Ausdruck einer besonders hohen Intelligenz und von einem anderen Psychologen als Anzeichen eines problematischen Intelligenzdefizits gedeutet wird. Derart divergierende Interpretationen werden bei psychometrischen Tests durch die Angabe von Normen vermieden. Diese Normen werden anhand repräsentativer Stichproben erhoben und dienen als Vergleichsmaßstab. Setzt man den Testwert eines Probanden in Bezug zu einer solchen Norm, so wird eindeutig erkennbar, ob der Proband eine unterdurchschnittliche, überdurchschnittliche oder auch stark überdurchschnittliche Merkmalsausprägung aufweist. Oftmals enthalten Tests nicht nur eine Norm für die Gesamtpopulation, sondern auch für verschiedene Subgruppen. Typisch wären etwa getrennte Normen für verschiedene Bildungsniveaus, Altersgruppen oder Männer und Frauen. Durch diese zusätzlichen Normen wird eine detailliertere Beurteilung der Merkmalsausprägung eines Probanden möglich.

Repräsentative und differenzierte Normen reichen allerdings nicht zwingend aus, um eine hohe Interpretationsobjektivität zu gewährleisten. In der psychologischen Praxis werden psychometrische Tests häufig eingesetzt, um konkrete diagnostische Fragestellungen zu beantworten. Ist die Intelligenz eines Rehabilitanden ausreichend, um ihm eine Umschulung zum Industriekaufmann zu empfehlen? Hier genügt es nicht zu wissen, dass der Rehabilitand ein durchschnittliches Testergebnis erzielt hat. Offensichtlich müsste das Testergebnis zu den Anforderungen, die mit dem Beruf des Industriekaufmanns verbunden sind, in Beziehung gesetzt werden. Standardisierte Interpretationen von Testergebnissen für solche konkreten Fragestellungen können in Testanweisungen zumeist höchstens beispielhaft angegeben werden. Ein Grund dafür besteht darin, dass der Inhaltsbereich, in dem ein Test sinnvoll eingesetzt werden kann, oftmals zu groß ist, um für alle denkbaren Fragestellungen standardisierte Interpretationen zur Verfügung zu stellen.

3.5.2 Reliabilität

Mit dem Begriff Reliabilität wird die Zuverlässigkeit oder Messgenauigkeit eines Messinstruments bezeichnet. Generell sollten wiederholte Messungen eines Objekts, das sich nicht verändert, selbstverständlich stets zu demselben Messergebnis führen. Bestimmen wir etwa mit einer Briefwaage zwei Mal das Gewicht dieses Buches, so werden wir erwarten, dass wir zwei Mal zu demselben Messwert gelangen. Sofern wir eine moderne, einigermaßen brauchbare Waage benutzen, wird dies kein Problem darstellen. Die Messwerte werden allenfalls geringfügig schwanken. Es tritt also nur ein geringer Messfehler auf, die Waage ist reliabel.

Verglichen mit der Reliabilität einer Waage oder anderer physikalischer Messinstrumente werden psychologische Messinstrumente häufig nur eine geringe Reliabilität aufweisen. So können die Messwerte psychometrischer Tests aufgrund einer Reihe unsystematischer und unkontrollierter Einflüsse schwanken. Möglicherweise werden die Ergebnisse eines Intelligenztests durch die Motivation, Müdigkeit oder Testängstlichkeit eines Probanden beeinflusst. Andere Einflüsse könnten durch Veränderungen der Untersuchungssituation verursacht werden. Vielleicht variieren die Testergeb-

nisse mit der Tageszeit oder der Raumtemperatur bei der Testdurchführung. Schließlich können Messfehler auch auf Eigenschaften des Tests zurückgehen. Denkbar wäre etwa, dass manche Items eines Tests von einem Probanden bei wiederholten Messungen nicht stets in derselben Weise aufgefasst werden. Diese Items würden dann unterschiedliche Antwortprozesse und somit auch unterschiedliche Antworten auslösen. Vielleicht weist der Test auch keine perfekte Auswertungsobjektivität auf. In diesem Fall könnte ein Proband selbst dann unterschiedliche Testergebnisse erzielen, wenn er bei allen Testungen exakt dieselben Antworten gibt. Hier wird deutlich, dass Objektivität eine Voraussetzung für Reliabilität darstellt: Ein Test kann nicht zuverlässig und genau sein, wenn seine Ergebnisse bereits davon abhängen, *wer* den Test durchführt, auswertet und interpretiert.

Jeder Messwert kann also mit einem Messfehler behaftet sein. Eine Grundannahme der klassischen Testtheorie besagt nun, dass sich der beobachtete Messwert X einer Person in einem Test aus dem konstanten „wahren" Wert T (z.B. der tatsächlichen Intelligenz eines Probanden) und dem Messfehler *E* zusammensetzt:

$$X = T + E$$

Gemäß weiterer Grundannahmen der Testtheorie handelt es sich bei dem Messfehler E um einen Zufallsfehler – oder auch einen unsystematischen Fehler. Dies bedeutet zunächst, dass der Messfehler nicht dazu führen wird, dass wir die wahre Merkmalsausprägung von Probanden *systematisch* über- oder unterschätzen. Messen wir etwa die Intelligenz unendlich vieler Testteilnehmer, so werden sich die unterschiedlichen, positiven und negativen Messfehler, die bei den einzelnen Probanden auftreten, ausmitteln. Der Mittelwert des Messfehlers beträgt also 0. Dies heißt auch, dass der Mittelwert der beobachteten Messwerte der wahren mittleren Intelligenz der Probanden entspricht. Die gleiche Überlegung trifft ebenfalls für wiederholte Messungen an einem Probanden zu: Könnten wir unendlich häufig die Intelligenz einer Person messen, so würden sich die Messfehler, die bei den einzelnen Testungen auftreten, ausgleichen. Der Mittelwert der beobachteten Messwerte wäre also der wahre Intelligenzwert dieser (gequälten) Person.

Generell können wir also demnach bei einer größeren Anzahl von Messungen erwarten, dass sich der Messfehler nicht im Mittelwert der Messwerte niederschlagen wird. Der Messfehler wird sich allerdings auf die Unterschiedlichkeit der Messwerte auswirken. Nehmen wir an, wir messen die Intelligenz von 100 Personen. Selbstverständlich sollten sich die Messwerte dieser Personen unterscheiden – wir führen die Testung überhaupt nur deswegen durch, weil wir davon ausgehen, dass zwischen Personen Intelligenzunterschiede bestehen. Wie stark die Messwerte variieren, hängt nun aber nicht nur davon ab, wie groß die Unterschiede zwischen den wahren Intelligenzwerten der Personen in unserer Stichprobe sind. Die Unterschiedlichkeit der Messwerte wird zudem durch die Messfehler bei den einzelnen Messungen vergrößert. Mit einem Test, bei dem häufig große Messfehler auftreten, werden wir in unserer Stichprobe eine größere Unterschiedlichkeit der Messwerte finden als mit einem Test, bei dem lediglich kleine Messfehler auftreten.

Ein Maß für die Unterschiedlichkeit (bzw. die *Streuung*) von Werten ist die *Varianz*, die mit s^2 gekennzeichnet wird (genauere Erläuterungen zum Konzept der Varianz und zu ihrer Berechnung finden Sie in *Kapitel 6*). Die Varianz der Messwerte (s_X^2) von

Testteilnehmern kann nun aufgeteilt werden in die Varianz der wahren Werte (s_T^2) der Teilnehmer und die Varianz der Messfehler (s_E^2). Die Reliabilität (r_{tt}) eines Tests ist wie folgt definiert:

$$r_{tt} = \frac{s_T^2}{s_X^2} = \frac{s_T^2}{s_T^2 + s_E^2}$$

Die Reliabilität entspricht also dem Anteil der Varianz der wahren Werte an der Varianz der beobachteten Messwerte. Ist die Varianz der Messfehler gering – was nichts anderes heißt, als dass die einzelnen Messfehler kaum von ihrem Mittelwert von 0 abweichen, so ist die Reliabilität hoch. Tritt gar kein Messfehler auf, so beträgt auch die Varianz der Messfehler 0. In diesem Fall nimmt die Reliabilität den Wert 1 an. Mit steigender Varianz der Messfehler sinkt die Reliabilität. Den Wert 0 nimmt sie allerdings nur an, wenn die wahren Werte keine Varianz aufweisen. In diesem Fall misst der Test keine Unterschiede zwischen Personen (es bestehen keine!), sondern er erfasst ausschließlich Unterschiede zwischen den Messfehlern, die bei der Testung der einzelnen Personen aufgetreten sind.

Wie können wir nun die Reliabilität eines Tests bestimmen? Es gibt verschiedene Verfahren der Reliabilitätsermittlung, die jeweils mit spezifischen Stärken und Schwächen behaftet sind. Die grundlegende Idee besteht bei den meisten dieser Verfahren darin, für jeden Probanden in einer Stichprobe zwei Messwerte zu ermitteln. Erzielen die einzelnen Probanden bei beiden Messungen ähnliche Ergebnisse, so ist der Test hoch reliabel – offensichtlich werden die Messwerte der Probanden nur wenig durch Messfehler verfälscht. Besteht zwischen den Ergebnissen der Probanden bei beiden Messungen dagegen nur eine geringe Übereinstimmung, so haben die Messfehler offensichtlich einen starken Einfluss auf die Messergebnisse. Der Test verfügt über eine niedrige Reliabilität. Die Übereinstimmung (oder der *Zusammenhang*) zwischen den Messwerten der Probanden bei beiden Messungen wird dabei durch den *Korrelationskoeffizienten* ausgedrückt (*Kapitel 7*). Der Korrelationskoeffizient nimmt den Wert 1 an, wenn zwischen den beiden Messwertreihen eine perfekte Übereinstimmung gegeben ist. Besteht zwischen den Messwertreihen dagegen gar kein Zusammenhang, so beträgt der Korrelationskoeffizient 0. Im Folgenden erläutern wir kurz einige Verfahren der Reliabilitätsermittlung.

Die Retest-Methode

Die nächstliegende Vorgehensweise zur Bestimmung der Reliabilität besteht vermutlich darin, ein und denselben Test derselben Stichprobe von Probanden in einem gewissen Zeitabstand zwei Mal vorzulegen und den Korrelationskoeffizienten zwischen den Ergebnissen der beiden Messungen zu ermitteln. Genau dies geschieht bei der Retest-Methode. Allerdings ist diese Vorgehensweise mit einigen Problemen verbunden. Zunächst kann die wiederholte Durchführung von Tests zu Übungseffekten führen. Möglicherweise erlernen einige Probanden einen verbesserten Umgang mit Tests und schneiden daher bei der zweiten Messung besser ab. Dies wird die beobachtete Reliabilität vermindern. Ein noch schwerwiegenderes Problem haben wir, wenn Erinnerungseffekte auftreten. Da wir den Probanden zwei Mal denselben Test vorlegen, können sich die Probanden bei der zweiten Messung möglicherweise an ihre ersten Antworten erinnern. Sie müssten den eigentlich beabsichtigten Beantwortungs- oder Lösungsprozess dann gar kein zweites Mal durchlaufen, sondern könnten ein-

fach ihre ersten Antworten nochmals notieren. In diesem Fall bestimmen wir mit der Retest-Methode nicht die Reliabilität des Tests, sondern die Erinnerungsleistung der Probanden! Um Erinnerungseffekten vorzubeugen, wird daher oftmals empfohlen, zwischen der ersten und der zweiten Testdurchführung einen längeren Zeitraum verstreichen zu lassen (zumeist mehrere Wochen). Dies kann allerdings ein anderes Problem auslösen: Von vielen psychischen Merkmalen kann man nicht annehmen, dass sie über einen beliebig langen Zeitraum völlig konstant sind. Möglicherweise verbessert sich die Intelligenzleistung eines Kindes aufgrund einer Fördermaßnahme. Vielleicht verändert sich die Persönlichkeit eines Probanden aufgrund eines kritischen Lebensereignisses (z.B. einer schweren Erkrankung). Ermitteln wir also für einen Test eine geringe Retest-Reliabilität, so könnte dies nicht nur auf das Auftreten großer Messfehler, sondern auch auf Veränderungen der wahren Werte zurückzuführen sein. Tatsächlich findet man in der Regel, dass die Retest-Reliabilität umso geringer ausfällt, je größer der Zeitraum ist, der zwischen den beiden Testungen liegt. Umgekehrt heißt dies aber auch: Ermitteln wir trotz eines großen Zeitabstands zwischen den beiden Testungen eine hohe Reliabilität, so ist der Test mit geringen Messfehlern verbunden *und* das gemessene Merkmal ist stabil. Die Retest-Reliabilität wird daher gelegentlich auch als *Stabilität* bezeichnet.

Die Paralleltest-Methode

Auch bei der Paralleltest-Methode werden an derselben Stichprobe von Probanden zu unterschiedlichen Zeitpunkten zwei Messungen durchgeführt und die Ergebnisse dieser Messungen korreliert. Allerdings werden bei den beiden Messungen nicht genau dieselben Test-Items eingesetzt, sondern es werden *parallele* oder *äquivalente Formen* eines Tests verwendet. Diese parallelen Formen bestehen aus unterschiedlichen Items, die aber exakt dasselbe Merkmale in exakt derselben Weise messen müssen. Übungseffekte können auch bei der Paralleltest-Methode auftreten. Da den Teilnehmern zu den beiden Testzeitpunkten verschiedene Items vorgelegt werden, sind Erinnerungseinflüsse jedoch ausgeschlossen. Die Paralleltest-Methode ermöglicht es also, die Reliabilität auch bei kürzeren Zeitabständen zwischen den Testungen zu bestimmen. Dafür stehen wir nun vor dem Problem, zwei parallele Formen eines Tests entwickeln zu müssen. Diese Parallelformen sollten dieselbe Anzahl an Items enthalten und mit den denselben Instruktionen und Erläuterungen dargeboten werden. Damit wir annehmen können, dass beide Formen tatsächlich dasselbe Merkmal in derselben Weise messen, sollten sich die Items in den Testversionen zudem auf möglichst ähnliche Inhalte beziehen und auch formal möglichst ähnlich gestaltet sein. Dies schließt beispielsweise ein, dass jeweils die gleiche Anzahl an Antwortoptionen vorgegeben wird und dass die Items in beiden Formen die gleiche Schwierigkeit aufweisen. Diese Anforderungen sind oftmals schwierig zu erfüllen, und es ist vorab nicht zu klären, ob es tatsächlich gelungen ist, zwei äquivalente Testformen zu entwickeln. Eine niedrige Paralleltest-Reliabilität kann also sowohl darauf zurückgehen, dass bei den Messungen große Messfehler auftreten, als auch darauf, dass die Testformen nicht exakt dasselbe messen. Finden wir dagegen eine hohe Paralleltest-Reliabilität, so heißt dies, dass die Ergebnisse beider Testformen nur wenig durch Messfehler verfälscht werden *und* dass die Testformen äquivalent sind.

Die Testhalbierungsmethode

Bei der Testhalbierungsmethode bearbeiten die Probanden in der Untersuchungsstichprobe lediglich *einmalig* einen Test. Allerdings werden die Items dieses Tests in zwei Hälften aufgeteilt. Auf der Basis dieser Testhälften werden für jeden Teilnehmer zwei Messwerte bestimmt. Treten nur geringe Messfehler auf, so ist natürlich zu erwarten, dass diese Messwerte eines Probanden ähnlich ausfallen. Die Reliabilität des Tests kann also auch hier ermittelt werden, indem man den Korrelationskoeffizienten zwischen den Ergebnissen der Teilnehmer (in beiden Testhälften) berechnet.[4]

Da durch die Halbierung des Tests quasi zwei Parallelformen entstehen, ähnelt die Testhalbierungsmethode der Paralleltest-Methode. Auch hier stellt sich die Frage nach der Äquivalenz der Testhälften. Es wäre beispielsweise sinnlos, die Messergebnisse in der ersten Hälfte eines Tests mit den Messergebnissen in der zweiten Hälfte zu korrelieren, wenn die Items zu Beginn des Tests leichter sind als gegen Ende. Eine sinnvollere Vorgehensweise bei der Halbierung des Tests könnte darin bestehen, Items mit gerader Reihungsnummer in die eine Testhälfte und Aufgaben mit ungerader Reihungsnummer in die andere Testhälfte aufzunehmen (*odd-even-Methode*). Auf diese Weise wäre auch ausgeschlossen, dass in beiden Testhälften unterschiedlich große Übungseffekte auftreten.

3.5.3 Validität

Ein Test ist valide, wenn er das misst, was er zu messen vorgibt. Dieser Satz klingt vielleicht zunächst etwas merkwürdig. Wenn wir an physikalische Messinstrumente denken, ist zumeist völlig eindeutig, was diese Messinstrumente messen. Selbstverständlich misst eine Waage das Gewicht, ein Zollstock die Länge von Objekten. Anders als diese physikalischen Messinstrumente sollen psychometrische Tests jedoch nicht direkt beobachtbare Merkmale erfassen, sondern latente Variablen messen. Damit stellt sich die Frage, ob die Antworten zu den Items eines Tests tatsächlich Indikatoren desjenigen latenten Merkmals sind, das gemessen werden soll. Erfassen beispielsweise die Items eines Intelligenztests tatsächlich die intellektuelle Leistungsfähigkeit der Probanden oder messen sie eher deren Konzentrationsfähigkeit? Werden die Ergebnisse des Intelligenztests vielleicht systematisch durch andere Persönlichkeitsmerkmale der Probanden – z.B. ihre Leistungsmotivation – beeinflusst? In diesen Fällen wäre der Intelligenztest nicht valide.[5] Bei der Frage nach der Validität eines Tests geht es also um die Güte der Operationalisierung des interessierenden Merkmals. Es ist zu klären, ob ein Test eine gelungene Operationalisierung derjenigen latenten Variable darstellt, die gemessen werden soll.

Wie lässt sich die Validität eines Tests nun beurteilen? Auch bei der Validität können mehrere Aspekte unterschieden werden. Drei dieser Aspekte seien hier vorgestellt.

4 Genau genommen gibt der Korrelationskoeffizient bei der Testhalbierungsmethode die Reliabilität der Testhälften an. Da die Reliabilität eines Tests mit der Anzahl seiner Items zunimmt, muss der Korrelationskoeffizient nach oben korrigiert werden, um der Reliabilität des gesamten Tests zu entsprechen.

5 Auch ein Test mit niedriger Reliabilität kann nicht sonderlich valide sein: Die Ergebnisse eines Tests mit geringer Reliabilität bringen zu einem großen Teil Messfehler zum Ausdruck – also eben nicht die Ausprägung des Merkmals, das gemessen werden soll.

Inhaltsvalidität

Ein denkbares Vorgehen zur Entwicklung eines validen Tests bestünde darin, zunächst alle Items zu sammeln, in denen sich das interessierende Merkmal ausdrückt. Aus diesem „Itemuniversum" könnte dann eine repräsentative Teilmenge von Items ausgewählt und in den Test aufgenommen werden. Auf diese Weise wäre sichergestellt, dass der Test das zu messende Merkmal in seinen wesentlichen Aspekten erschöpfend erfasst. Der Test wäre inhaltsvalide.

Diese Vorgehensweise erfordert natürlich, dass das Universum aller Items, die ein Merkmal abbilden, eindeutig bestimmt werden kann. Dies ist insbesondere dann möglich, wenn der Test einfache, umgrenzte Fähigkeiten messen soll. Dies wäre etwa bei einem Test zur Messung der Kenntnisse in den Grundrechenarten der Fall. Zwar werden wir auch hier kaum alle relevanten Aufgaben sammeln können oder wollen, aber es ist klar, wie sich das Itemuniversum zusammensetzt: Alle möglichen Items enthalten Zahlen aus irgendeinem definierbaren Zahlenraum (bei einem Test für fortgeschrittene Grundschulkinder vielleicht aus dem Zahlenraum 1 bis 100) und eine der vier grundlegenden Rechenoperationen. Ein Test, der lediglich das kleine Einmaleins abprüft, nur aus Divisionsaufgaben besteht oder nur Zahlen enthält, die kleiner als 10 sind, würde also keine repräsentative Itemsammlung darstellen und wäre nicht inhaltsvalide. Die Inhaltsvalidität eines Tests sollte beispielsweise auch dann relativ leicht sicherzustellen sein, wenn Schulkenntnisse gemessen werden sollen. Ein Biologietest für das neunte Schuljahr wäre etwa dann inhaltsvalide, wenn seine Aufgaben den Unterrichtsstoff gut repräsentieren.

Bei breiteren und komplexeren Fähigkeiten ist es in der Regel nicht möglich, in irgendeiner Form ein Itemuniversum zu definieren. So dürfte es beinahe unendlich viele und sehr divergente Aufgaben geben, bei denen die Antworten unterschiedliche Ausprägungen des Merkmals Intelligenz anzeigen. Anders ausgedrückt: Uns fehlt eine hinreichend präzise Vorstellung von der Gesamtheit aller Aufgaben, die das Merkmal Intelligenz abbilden. Dennoch sollte prinzipiell natürlich auch ein Intelligenztest dem Kriterium der Inhaltsvalidität genügen. Die Test-Items sollten eine repräsentative Auswahl aller Items sein, die das Merkmal Intelligenz erfassen. Allerdings ist es in diesem Fall sehr schwierig zu beurteilen, ob ein Test tatsächlich eine repräsentative Itemmenge enthält. Eine formale Möglichkeit, die Höhe der Inhaltsvalidität zu bestimmen und in einer Zahl auszudrücken, besteht in diesem Fall gar nicht. Stattdessen wird einem solchen Test ausschließlich auf der Basis subjektiver (und hoffentlich übereinstimmender) Urteile von Experten Inhaltsvalidität bescheinigt oder abgesprochen.

Kriteriumsvalidität

Dieser Aspekt der Validität eines Tests wird geprüft, indem man die Übereinstimmung zwischen den Testwerten und sogenannten Kriterien bestimmt. Bei diesen Kriterien handelt es sich um Variablen, mit denen die Testwerte zusammenhängen sollten, sofern der Test tatsächlich das misst, was er zu messen vorgibt. Wir könnten z.B. ermitteln, wie die Ergebnisse von Schulkindern in einem Intelligenztest mit dem Urteil der Lehrer über die Intelligenz der Kinder übereinstimmen. Wie bei der Reliabilität wird die Übereinstimmung auch hier durch den Korrelationskoeffizienten ausgedrückt. Die Höhe der Kriteriumsvalidität kann also in einer Maßzahl angegeben werden.

Ein offensichtliches Problem bei dieser Vorgehensweise besteht darin, eine geeignete Kriteriumsvariable zu finden. Welches objektiv und reliabel feststellbare Kriterium würde uns fehlerfrei über die Intelligenz, den Neurotizismus, die Extraversion oder die Verträglichkeit von Probanden informieren? Im obigen Beispiel sind natürlich auch die Lehrerurteile kein perfekter Indikator der Intelligenz der Kinder. Wir müssen etwa damit rechnen, dass diese Urteile durch die Sympathie für die Kinder oder andere Urteilstendenzen systematisch verfälscht werden. Zudem sind die Lehrerurteile sicher auch nicht völlig reliabel: Eine wiederholte Befragung der Lehrer wird nicht zu identischen Ergebnissen führen. Selbst wenn unser Intelligenztest ein sehr valides Messinstrument ist, werden wir daher keine perfekte Übereinstimmung zwischen den Testergebnissen und den Lehrerurteilen erwarten können. Wir werden also bereits mit Korrelationskoeffizienten mittlerer Höhe zufrieden sein müssen. Fänden wir dagegen gar keine Übereinstimmung zwischen Testergebnissen und Lehrerurteilen, so wäre dies ein Grund, zumindest zu bezweifeln, dass unser Test tatsächlich Intelligenz misst. Da sich in der Regel kein ideales Kriterium für ein Merkmal finden lässt, ist es sinnvoll, einen Test an mehreren Kriterien zu validieren (bei einem Intelligenztest böten sich vielleicht auch der Studien- oder Berufserfolg an). Ein Test verfügt dann auch nicht nur über eine Kriteriumsvalidität, sondern über so viele Kriteriumsvaliditäten wie Variablen zu seiner Überprüfung herangezogen werden.

Oftmals werden zur Validierung eines Tests nicht nur *Außenkriterien* (wie die Lehrerurteile oder der Berufserfolg) verwendet, sondern auch andere Tests, die dasselbe Merkmal messen. Dieses Vorgehen wird als *innere Validierung* bezeichnet. In diesem Fall sollten wir hohe Korrelationskoeffizienten finden, da beispielsweise zwei unterschiedliche Intelligenztests zumindest etwas sehr Ähnliches messen sollten. Dieses Vorgehen ist natürlich nur dann sinnvoll, wenn der als Kriterium verwendete Test bereits als valides Messinstrument anerkannt ist. Andernfalls können wir zwar feststellen, dass die Tests tatsächlich dasselbe messen, es bleibt aber offen, was sie messen. An irgendeiner Stelle muss also zwangsläufig der Bezug der Tests zu Außenkriterien hergestellt werden.

Nach dem Zeitpunkt, zu dem das Kriterium erhoben wird, unterscheidet man die *Übereinstimmungsvalidität* und die *Vorhersagevalidität*. Bei der Übereinstimmungsvalidität werden Test- und Kriteriumswerte (fast) gleichzeitig ermittelt. Bei der Vorhersagevalidität werden die Kriteriumswerte nach der Testdurchführung erhoben. Dies ist insbesondere dann sinnvoll, wenn der Test in der diagnostischen Praxis zur Vorhersage künftigen Verhaltens eingesetzt werden soll. Ein Schulreifetest soll etwa den künftigen Schulerfolg prognostizieren. Ein Berufseignungstest wird verwendet, um künftigen Berufserfolg vorherzusagen. Zu demselben Zweck kann unter Umständen auch ein Intelligenztest bei der Personalauswahl eingesetzt werden. Die Kriterien (Schul- und Berufserfolg) sollten in diesen Fällen nach der Testdurchführung ermittelt werden. Die Höhe der Vorhersagevalidität der Tests informiert uns dann auch darüber, wie gut die Tests in der Praxis geeignet sind, ihren diagnostischen Zweck zu erfüllen.

Konstruktvalidität

Die beiden vorangegangenen Abschnitte zeigen, dass oftmals weder die Inhaltsvalidität noch die Kriteriumsvalidität zu einer eindeutigen, unzweifelhaften Aussage darüber führen, was ein Test tatsächlich misst. Dies gilt insbesondere dann, wenn ein Test Merkmale wie Intelligenz oder Neurotizismus messen soll, die nur schwer operational (also in eindeutigen beobachtbaren Indikatoren) zu fassen sind. Diesem Umstand trägt die Konstruktvalidität Rechnung. Die Konstruktvalidierung eines Tests ist ein längerer, fortdauernder Prozess, in dem theoretische Aussagen über das zu messende Merkmal mithilfe des Tests überprüft werden. Die Validität des Tests wird hier also nicht nur anhand einzelner Außenkriterien ermittelt, sondern indem geprüft wird, ob möglichst vielfältige Hypothesen über das Merkmal durch die Testwerte bestätigt werden. Dies setzt natürlich zunächst voraus, dass solche Hypothesen abgeleitet werden können. Messen wir etwa Aggressivität, so ist vielleicht zu erwarten, dass wir bei jüngeren Männern höhere Testwerte finden als bei älteren Männern. Messen wir Depressivität, so sollten wir bei einer Gruppe von Probanden, die wegen einer Depression in therapeutischer Behandlung sind, höhere Werte ermitteln als bei einer Gruppe von Probanden, die nicht über depressive Symptome klagen. Die Intelligenztestwerte von Sonderschülern sollten niedriger sein als die Intelligenztestwerte von Stipendiaten einer Eliteuniversität. Es ist zu erwarten, dass Intelligenztestwerte mit dem Berufserfolg zusammenhängen. Da Intelligenz als ein stabiles Merkmal aufgefasst wird, sollten wir innerhalb kurzer Zeit keine großen systematischen Schwankungen in den Testwerten eines Probanden finden. Die Konzentrationsfähigkeit eines Probanden sollte unter starkem Alkoholeinfluss geringer sein als in nicht alkoholisiertem Zustand. Im Zuge der Konstruktvalidierung eines Tests werden derartige Aussagen über ein Merkmal mithilfe des Tests überprüft. Da diese Überprüfung zum Teil darauf hinauslaufen kann, Kriteriumswerte mit Testwerten zu korrelieren, schließt die Konstruktvalidität unter Umständen die Kriteriumsvalidität ein. Allerdings werden hier auch Kriterien verwendet, die *nicht* mit den Testwerten übereinstimmen sollten. Ein Intelligenztest sollte natürlich Intelligenz und nicht Konzentrationsfähigkeit messen. Es wäre also zu prüfen, ob die Intelligenztestwerte tatsächlich nicht mit den Ergebnissen aus einem Test zur Messung der Konzentrationsfähigkeit zusammenhängen.

Können mit einem Test möglichst viele Hypothesen über ein Merkmal bestätigt werden, so spricht dies für die Konstruktvalidität des Tests. Da potenziell stets weitere und auch neue Hypothesen über ein Merkmal überprüft werden können, führt die Konstruktvalidierung nicht zu einer endgültigen, numerischen Aussage über die Validität eines Tests. Stattdessen kann aufgrund der Konstruktvalidierung angegeben werden, wie gut sich ein Test bisher bewährt hat. Je mehr Hypothesenüberprüfungen erfolgreich verlaufen sind, desto überzeugender ist die Annahme, der Test sei valide. Die einzelnen Hypothesenüberprüfungen sind dabei allerdings nur dann eindeutig interpretierbar, wenn die jeweilige Hypothese bereits vor der Validierung des Tests als gültig betrachtet werden kann. Müssen wir etwa bezweifeln, dass jüngere Männer aggressiver sind als ältere, so bleibt unklar, was gleiche Testwerte für jüngere und ältere Männer bedeuten: Misst der Test nicht Aggressivität oder gibt es tatsächlich keine altersabhängigen Unterschiede in der Aggressivität?

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

In der sozialwissenschaftlichen Forschung ist es notwendig, Merkmale von Personen zu messen. Dies heißt nichts anderes, als dass die jeweiligen Merkmalsausprägungen in Zahlen ausgedrückt werden müssen. Beim Messen werden also Personen (oder Objekten) hinsichtlich eines bestimmten Merkmals Zahlen zugeordnet. Diese Zuordnung kann natürlich nicht willkürlich vorgenommen werden. Von einer Messung kann erst dann gesprochen werden, wenn die Zuordnung so erfolgt, dass bestimmte empirisch feststellbare Relationen zwischen den Personen (z.B. „hat das gleiche Geschlecht" oder „ist intelligenter") auch durch entsprechende Relationen zwischen den zugeordneten Zahlen zum Ausdruck kommen. Genügt eine Zuordnung dieser Anforderung, so wird sie auch als *homomorphe Abbildung* eines *empirischen Relativs* in ein *numerisches Relativ* bezeichnet. Bei der Erarbeitung von homomorphen Abbildungen ergeben sich für die Messtheorie nun bestimmte Probleme. Beim *Repräsentationsproblem* geht es um die Frage, ob ein bestimmtes Merkmal überhaupt messbar ist. Dies ist dann der Fall, wenn die empirisch feststellbaren Relationen zwischen Personen (oder allgemeiner: Messobjekten) bestimmte Eigenschaften aufweisen, die dazu führen, dass diese Relationen auch durch Zahlen wiedergegeben werden können. Das *Eindeutigkeitsproblem* betrifft die Frage, wie die zugeordneten Zahlen (die Messwerte) transformiert werden können, ohne dass Information über die Messobjekte verloren geht. Schließlich ist zur Lösung des *Bedeutsamkeitsproblems* zu klären, welche mathematischen Operationen mit Messwerten sinnvoll sind. Eine bestimmte Verrechnung von Messwerten kann dabei immer dann als sinnvoll betrachtet werden, wenn sie zu Aussagen führt, die auch empirisch zutreffend sind.

Jede Messung erfolgt auf einem bestimmten *Skalenniveau*. Verschiedene Skalenniveaus unterscheiden sich aufgrund der Relationen, die zwischen den Messobjekten empirisch bestimmbar sind und die daher auch durch die Messwerte wiedergegeben werden. Messwerte auf unterschiedlichen Skalenniveaus haben somit einen unterschiedlich großen Informationsgehalt. Auf *Nominalskalenniveau* wird empirisch ausschließlich festgestellt, ob die Messobjekte gleich oder ungleich sind. Auch die Messwerte informieren hier daher ausschließlich über die Gleichheit oder Ungleichheit der Messobjekte.

Messungen auf einer *Ordinalskala* enthalten zusätzlich Informationen darüber, bei welchem von zwei Messobjekten ein Merkmal stärker ausgeprägt ist. Auf dem *Intervallskalenniveau* sind auch Aussagen über die Größe des Unterschieds zwischen Messobjekten möglich. Schließlich erlauben Messungen auf dem *Verhältnisskalenniveau* Aussagen über Verhältnisse zwischen Messobjekten.

Auf dem Niveau einer *Absolutskala* existiert zusätzlich eine natürliche Maßeinheit. Das Skalenniveau einer Messung muss bei der weiteren Analyse der Messwerte beachtet werden, da jedes statistische Verfahren ein bestimmtes (minimales) Skalenniveau voraussetzt.

Psychometrische Tests sind Messinstrumente, die spezifisch in der Psychologie verwendet werden. Derartige Tests dienen zur Messung latenter Variablen – also solcher Merkmale, die nicht direkt beobachtet werden können. Die *klassische Testtheorie* beurteilt anhand bestimmter Kriterien, ob und wie gut ein Test geeignet ist, ein bestimmtes Merkmal zu erfassen. Diese *Gütekriterien* können aber auch generell als ein Standard aufgefasst werden, dem „gute" Messungen genügen sollten.

▶

Die Hauptgütekriterien sind *Objektivität*, *Reliabilität* und *Validität*. Alle diese Begriffe können weiter ausdifferenziert werden. Objektiv ist eine Messung dann, wenn das Messergebnis unabhängig von der Person ist, die das jeweilige Messinstrument anwendet. Bei psychometrischen Tests werden die Aspekte *Durchführungs-*, *Auswertungs-* und *Interpretationsobjektivität* unterschieden. Mit dem Begriff Reliabilität wird die Messgenauigkeit eines Tests bezeichnet. Jede Messung kann durch zufällige, unsystematische Messfehler beeinflusst werden. Ein Test ist dann reliabel, wenn seine Messergebnisse nicht oder nur wenig durch solche Messfehler verfälscht werden. Aufgrund der Verfahren die zur Ermittlung der Reliabilität eines Tests verwendet werden können, unterscheidet man die *Retest-Reliabilität*, die *Paralleltest-Reliabilität* und die *Testhalbierungs-Reliabilität*.

Die Validität betrifft schließlich die Frage, ob ein Test tatsächlich diejenige latente Variable misst, die er zu messen vorgibt. Auch hier werden verschiedene Aspekte unterschieden: *Inhaltsvalide* ist ein Test dann, wenn seine Items eine repräsentative Auswahl aller Items darstellen, die geeignet sind, das fragliche Merkmal zu erfassen. *Kriteriumsvalidität* ist gegeben, wenn die Testergebnisse mit anderen Indikatoren des zu messenden Merkmals übereinstimmen. *Konstruktvalidität* besteht schließlich, wenn zahlreiche (gesicherte) theoretische Aussagen über die latente Variable, die gemessen werden soll, mithilfe des Tests bestätigt werden können.

ZUSAMMENFASSUNG

Weiterführende Literatur

Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson Studium.

Gut lesbares Lehrbuch zum Thema Testtheorie und Testentwicklung.

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.

Eine umfassende, einführende Darstellung der Klassischen Testtheorie.

Orth, B. (1974). *Einführung in die Theorie des Messens*. Stuttgart: Kohlhammer.

Eine „leserfreundliche" Behandlung der Grundlagen der Messtheorie.

Übungsaufgaben mit Lösungen sowie weitere Informationen zu diesem Buchkapitel finden Sie auf der Companion Website zum Buch unter *http://www.pearson-studium.de*

Copyright