Forschungsmethoden und Statistik

für Psychologen und Sozialwissenschaftler

3., aktualisierte und erweiterte Auflage

Peter Sedlmeier
Frank Renkewitz

Pearson

Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler

Forschungsmethoden und Statistik

für Psychologen und Sozialwissenschaftler

3., aktualisierte und erweiterte Auflage

Peter Sedlmeier
Frank Renkewitz

Pearson

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

21 20 19 18

ISBN 978-3-86894-321-4 (Buch)
ISBN 978-3-86326-808-4 (E-Book)

Programmleitung: Kathrin Mönch, kmoench@pearson.de
Lektorat: Elisabeth Prümm, epruemm@pearson.de
Korrektorat: Christian Schneider
Herstellung: Claudia Bäurle, cbaeurle@pearson.de
Satz: Gerhard Alfes, mediaService, Siegen (www.mediaservice.tv)
Coverillustration: 123rf.com
Druck und Verarbeitung: DZS-Grafik d.o.o., Ljubljana

Printed in Slovenia

Inhaltsübersicht

Inhaltsverzeichnis

Vorwort zur 3. Auflage

Wir freuen uns, dass die zweite Auflage des Buchs noch mehr Anklang gefunden hat als die erste und hoffen natürlich, dass dieser Trend auch für diese dritte, nun wiederum deutlich modifizierte und erweiterte Auflage anhält. Und wieder haben wir es hauptsächlich den Rückmeldungen unserer Studierenden zu verdanken, dass diese Neuauflage weniger Ungereimtheiten enthält und (hoffentlich) noch leichter lesbar ist. Ein besonderer Dank gebührt den fleißigen und für uns erfreulicherweise – weil für ein Methodenbuch sehr wichtig – sehr genau lesenden Hagener Studierenden, deren Rückmeldungen uns freundlicherweise Lena Schützler hat zukommen lassen. Auch von weiteren kritischen Leserinnen und Lesern haben wir immer wieder Mails bekommen, die uns auf Unzulänglichkeiten hingewiesen haben. Wir sind allen Mails nachgegangen und haben (fast) immer entsprechende Modifikationen vorgenommen – herzlichen Dank! Wir hoffen auf ihr Verständnis dafür, dass wir nicht alle an dieser Stelle namentlich nennen können.

Wir haben kleinere Veränderungen in allen Kapiteln vorgenommen und dabei neuere Entwicklungen mit aufgenommen. Das hat sich unter anderem in einem merkbar erweiterten Literaturverzeichnis niedergeschlagen. Einige Kapitel des Buchs haben wir ausführlicher überarbeitet. So enthält **Kapitel 2** *Wissenschaftstheorie, Theorien und Hypothesen* nun erheblich mehr Informationen über den derzeit wohl wichtigsten wissenschaftstheoretischen Ansatz, den Kritischen Rationalismus. **Kapitel 16** *Kontrastanalyse* ist nun etwas kürzer – wir haben einige Varianten herausgenommen, die in der Praxis wenig benutzt werden. In **Kapitel 25** *Indirekte Effekte, latente Variablen und multiple Analyseebenen* geben wir, einen Trend in der Forschung aufgreifend, der Mehrebenenanalyse jetzt deutlich mehr Platz.

Wie die 2. Auflage mit dem neuen Buchteil zum *Allgemeinen Linearen Modell* wurde auch diese 3. Auflage um einige zusätzliche Kapitel erweitert. Drei Kapitel sind ganz neu und zum Teil im Zusammenhang mit unseren eigenen Forschungsinteressen entstanden. Frank Renkewitz (FR) hat sich in den letzten Jahren ausführlich mit der Replikationskrise in der Psychologie auseinandergesetzt, woraus **Kapitel 20** *Probleme der klassischen Inferenzstatistik in der Forschungspraxis* und **Kapitel 21** *Replikation, Präregistrierung, Open Science* entstanden sind. Peter Sedlmeier (PS) fand sich in seiner Forschungspraxis vermehrt mit dem Problem der Heterogenität von Studienteilnehmern konfrontiert und sieht als eine Antwort darauf die *Experimentelle Einzelfallanalyse*, die nun in **Kapitel 30** behandelt wird. Das frühere Kapitel *Inferenzstatistik: Erweiterungen und Ergänzungen* haben wir in zwei neue Kapitel aufgesplittet. Eines davon, **Kapitel 19** *Resampling Verfahren* (PS) enthält nun umfassendere und detailliertere Informationen zu Bootstrap-Verfahren und Randomisierungstests und das zweite, **Kapitel 22** *Bayesianische Statistik* (FR) behandelt den Bayesianischen Ansatz deutlich ausführlicher. Auch das **Kapitel 28** *Metaanalyse* (PS) ist grundlegend überarbeitet und auf den neuesten Stand gebracht. Über die Einbeziehung und Gewichtung spezifischer Inhalte in diesen neuen oder grundlegend überarbeiteten Kapiteln waren wir uns – keine Seltenheit in der Wissenschaft – nicht immer völlig einig und im Zweifelsfall lag es dann an der Entscheidung des federführenden Autors, was Sie nun in diesen Kapiteln lesen.

Die neue Auflage hat enorm davon profitiert, dass viele Kollegen und Mitarbeiter große Teil davon gelesen und uns Rückmeldungen dazu gegeben haben. Herzlichen Dank an Friederike Brockhaus, Markus Burkhardt, Vivien Röder, Thomas Schäfer, Johannes Titz, Isabell Winkler und insbesondere an Johannes Hönekopp, der das gesamte Buch durchgesehen und viele wertvolle Verbesserungsvorschläge gemacht hat. Besonderen Dank verdient auch Guada Paralta Ramos, die die Illustrationen in Kapitel 30 in professioneller Weise erstellt hat. Melanie Keiner danken wir für die Abbildungen in den Kapiteln 20 bis 22 und für viele hilfreiche Hinweise zu diesen Kapiteln. Last not least möchten wir Kathrin Mönch vom Pearson Verlag danken, die uns zu dieser Revision ermuntert und uns dann über den gesamten Prozess der nicht immer einfachen Revisionsarbeit kompetent und verständnisvoll begleitet hat.

Chemnitz und Erfurt

Peter Sedlmeier und Frank Renkewitz

Vorwort zur 2. Auflage

Die Rückmeldungen auf die erste Auflage dieses Buchs waren sehr positiv (worüber wir uns natürlich gefreut haben) und das nicht nur von Seiten der ursprünglich angesprochenen Leserschaft, den Psychologiestudierenden und Psychologen. Auch Studierende und Lehrende aus den Sozialwissenschaften haben das Buch offensichtlich mit Gewinn in ihrer Methodenausbildung benutzt, und das hat uns dazu bewegt, diese Personengruppe in der nun vorliegenden zweiten Auflage noch etwas stärker anzusprechen und den Titel des Buchs zu ändern. Zudem haben wir in dieser Auflage versucht, kleinere Ungereimtheiten und Unvollkommenheiten im gesamten Buch auszumerzen und auch die Literaturangaben auf den neuesten Stand gebracht. Das Buch enthält nun aber auch einige weitergehende inhaltliche Neuerungen. Die wichtigste und umfangreichste betrifft die Einführung des *Allgemeinen Linearen Modells* (*ALM*), das es erlaubt, eine Vielzahl konventioneller Verfahren wie etwa Mittelwertsvergleiche oder Varianz- und Kovarianzanalysen aus einer einheitlicheren Perspektive zu betrachten. Darüber hinaus ist das ALM die Grundlage für einige, teilweise ziemlich komplexe Verfahren, die in der psychologischen und sozialwissenschaftlichen Forschung immer mehr an Bedeutung gewinnen. Diesem neuen Thema haben wir einen neuen Buchteil mit drei Kapiteln gewidmet. Nach einer Einführung in das Allgemeine Lineare Modell (**Kapitel 20**) geben wir einen Einblick in die vielfältigen Möglichkeiten der Regressionsrechnung (**Kapitel 21**) und erörtern schließlich (**Kapitel 22**) einige Verfahren, die es erlauben, auch indirekte Effekte, latente Variablen und mehrere Analyseebenen in die Datenanalyse mit einzubeziehen. Die beiden letzten Kapitel sind etwas schwerer zu lesen als der Rest des Buchs, aber manche Dinge sind eben etwas komplizierter.

Darüber hinaus behandelt dieses Buch einige neue Effektgrößen (**Kapitel 9**) und beschreibt Ergänzungen zur Berechnung von Konfidenzintervallen für Effektgrößen (inklusive der neu eingeführten – **Kapitel 24**). Auch für die Bewertung von Teilergebnissen in Metaanalysen werden in dieser Auflage einige zusätzliche Verfahren (wie

z.B. Konfidenzintervalle für aggregierte Effektgrößen) vorgeschlagen (**Kapitel 25**) und schließlich wird das Thema „Fehlende Daten“ nun ausführlich erörtert (**Kapitel 26**).

Unser Dank gilt zunächst unseren Studierenden, die uns auf kleine und etwas größere Probleme in der ersten Auflage aufmerksam gemacht haben und von deren Rückmeldungen wir auch bei der Gestaltung der neuen Kapitel sehr profitiert haben. Zu diesen Kapiteln haben wir auch von einigen Kollegen und Kolleginnen sehr hilfreiche Rückmeldungen erhalten: ein herzlicher Dank an Friederike Brockhaus, Juliane Eberth, Joachim Engel, Thomas Schäfer, Hermann Singer, Martin Spieß, Isabell Winkler, Marcus Schwarz und Juliane Kämpfe!

Last not least möchten wir uns auch wieder bei unseren Ansprechpartnern vom Pearson Verlag für die zahlreichen Ermunterungen, Hinweise und vor allem ihre Geduld bedanken. Es waren dies Christian Schneider, Andra Riemhofer, und vor allem Alice Kachnij und Kathrin Mönch.

Peter Sedlmeier & *Frank Renkewitz*

Vorwort zur 1. Auflage

Anders als in vielen anderen Wissenschaftsbereichen können alle Menschen über psychologische Themen ohne spezielle Ausbildung mitreden: Jeder erinnert sich, löst Probleme, trifft Entscheidungen, hat Gefühle, ist mehr oder weniger motiviert, lernt, handelt, versucht anderen zu helfen und denkt über all dies nach. Letzteres ist im Grunde auch, was ausgebildete Psychologen tun, wenn sie menschliches Erleben und Verhalten erforschen. Der Unterschied zwischen dem, was jeder tut und was man als „Alltagspsychologie“ bezeichnen könnte, und dem, was akademische Psychologinnen und Psychologen tun, ist der Gegenstand dieses Buches: *Forschungsmethoden und Statistik*. Das Buch gliedert sich in fünf Teile, deren Inhalte wir nun kurz vorstellen.

Der erste Teil, *GRUNDLAGEN EMPIRISCHER FORSCHUNG*, befasst sich mit grundlegenden Annahmen und Forschungsmethoden, die in jeder empirischen Untersuchung eine wichtige Rolle spielen. Zunächst wird in **Kapitel 1** der Unterschied zwischen Alltagspsychologie und wissenschaftlicher Psychologie thematisiert und die wissenschaftliche Vorgehensweise kurz vorgestellt. **Kapitel 2** führt erst einmal weg von der Psychologie in die Philosophie, genauer gesagt, in die Wissenschaftstheorie, in der es um die Begründung wissenschaftlichen Arbeitens geht. In diesem Kapitel wird auch die wissenschaftliche Methode vorgestellt und erläutert, wie man Theorien erhält und vor allem, wie man diese überprüfen kann. Zur Überprüfung von Theorien benötigt man in der Regel Daten. Wie man für psychologische Fragestellungen relevante Daten erhält und wann und wie man sie sinnvoll interpretieren kann, ist Gegenstand von **Kapitel 3**. Die zwei generellen Methoden, empirische Daten zu erhalten, sind Befragung und Beobachtung. In **Kapitel 4** erläutern wir die mannigfaltigen Möglichkeiten, die sich hierbei ergeben, aber auch Grenzen und Fehlermöglichkeiten. In der wissenschaftlichen Psychologie werden Studien auf sehr systematische Weise durchgeführt. Wichtige und bewährte Vorgehensweisen, die es unter anderem auch erlauben, Ursache-Wirkungs-Schlüsse zu ziehen, werden in **Kapitel 5** ausführlich beschrieben.

Die *DESKRIPTIVE UND EXPLORATIVE DATENANALYSE* ist Gegenstand des zweiten Teils des Buches. In der psychologischen Forschung sieht man sich meist die Daten für eine Stichprobe (eine Auswahl) von Personen zusammen an. In **Kapitel 6** wird erläutert, wie man die Werte einer Stichprobe mithilfe von sogenannten Lage- und Streuungsmaßen zusammenfassen kann. **Kapitel 7** befasst sich damit, wie man den Zusammenhang zwischen zwei Merkmalen (z.B. der Größe und des Gewichts von Personen) grafisch und zahlenmäßig darstellt. Dieser Zusammenhang zwischen zwei Merkmalen kann auch „gerichtet" sein, man könnte also beispielsweise versuchen, aufgrund der Kenntnis der Körpergröße von Personen ihr Gewicht vorherzusagen. Die dazugehörige Methode wird in **Kapitel 8** vorgestellt. Schließlich wird in **Kapitel 9** beschrieben, wie man die Größe von Effekten, also beispielsweise den Unterschied zwischen Männern und Frauen hinsichtlich eines Merkmals, unabhängig von den Besonderheiten einer Einzelstudie bestimmen kann, so dass die Ergebnisse aus unterschiedlichen Studien miteinander vergleichbar werden.

Während sich die deskriptive und explorative Datenanalyse auf die vorhandenen Daten beziehen, kann man mithilfe der *INFERENZSTATISTIK*, die im dritten Teil des Buches beschrieben wird, Schlussfolgerungen auf die zugrunde liegenden Grundgesamtheiten oder Populationen ziehen. In **Kapitel 10** erläutern wir die theoretischen Grundlagen für solche Schlüsse und in **Kapitel 11** und **Kapitel 12** führen wir die zwei hauptsächlichen Verfahren dazu ein, *Konfidenzintervalle* und *Signifikanztests*. Welche Art von inferenzstatistischen Verfahren man in einem gegebenen Fall benutzen sollte, hängt von der Fragestellung und von der Art der zur Verfügung stehenden Daten ab. Die **Kapitel 13 bis 18** decken die wichtigsten Arten von Signifkanztests, die in der psychologischen Forschung verwendet werden, ab. Das abschließende **Kapitel 19** befasst sich mit alternativen Verfahren, die gegenwärtig in der Psychologie noch keine besonders große Rolle spielen, aber unseres Erachtens in manchen Fällen mit großem Gewinn eingesetzt werden können.

Der fünfte Teil des Buchs, *WEITERE VERFAHREN DER DATENERHEBUNG UND DATENANALYSE*, geht teilweise deutlich über die Inhalte existierender Methodenbücher hinaus. **Kapitel 23** befasst sich mit Methoden, die dazu dienen, Daten durch die Anwendung grafischer Verfahren genauer zu explorieren und besser zu verstehen. In **Kapitel 24** wird noch einmal zusammenfassend diskutiert, wie man die Größe von Effekten interpretiert. Wenn man die Effekte aus vielen vergleichbaren Studien zur Verfügung hat, kann man den diesen Studien zugrunde liegenden Effekt in der Population sehr genau schätzen. Wie das geht, wird in **Kapitel 25** erläutert. **Kapitel 26** geht auf verschiedene Fehlermöglichkeiten bei der Erhebung von Daten ein und beschreibt u.a. Verfahren der Datenerhebung bei „sensiblen" Daten, etwa wenn durch eine Tendenz zu sozial erwünschten Antworten die Gefahr von verfälschten Angaben besteht. Die Themen der beiden letzten Kapitel dieses vierten Teils sind normalerweise in einführenden Methodenbüchern nicht zu finden, können jedoch das psychologische Methodenarsenal sehr bereichern. **Kapitel 27** beschreibt, wie man mithilfe von Computermodellierung deutlich präzisere Theorien erhalten kann und **Kapitel 28** gibt einen Überblick über sogenannte qualitative Methoden, bei denen oft Texte (und nicht Zahlen) sowohl als Daten fungieren als auch das Ergebnis der Datenanalyse sind.

Der letzte Teil des Buchs, die *REFLEXION* besteht nur aus einem einzigen Kapitel, in dem die Inhalte des Buchs rekapituliert werden und wir unsere Vorstellungen über

die Rolle von Forschungsmethodik und Statistik in der Psychologie noch einmal zusammenfassend thematisieren.

Einiges in diesem Buch ist etwas anders als in vergleichbaren Büchern. Wir haben versucht, Konzepte und Verfahren möglichst intuitiv zu erklären, dabei aber möglichst wenig an Präzision einzubüßen. Naturgemäß kamen wir dabei manchmal an unsere Grenzen: Manche Dinge sind eben komplex. Auch die starke Gewichtung von grafischen Verfahren und Effektgrößen ist konsistent mit dem Versuch, die Datenanalyse ohne Verlust an Präzision so einfach und verständlich wie möglich zu gestalten und folgt Vorschlägen einer prominent besetzten Kommission des weltweit größten Psychologenverbands, der *American Psychological Association* (Wilkinson et al., 1993). Als häufig bessere Alternative zu dem wohl am weitesten verbreiteten inferenzstatistischen Verfahren, der Varianzanalyse, schlagen wir die sogenannte Kontrastanalyse vor und haben ihr auch ein eigenes Kapitel gewidmet. Wie schon erwähnt, wird auch ein beträchtlicher Anteil der nützlichen Methoden, die wir im vierten Teil des Buchs beschreiben, bisher eher selten in einführenden Methodenbüchern behandelt.

Beim Schreiben des Buches haben wir von den hilfreichen Rückmeldungen von Probelesern enorm profitiert. Dabei haben wir versucht, Eindrücke von verschiedenen Personengruppen zu sammeln. Zunächst möchten wir den engagierten Studierenden danken, die die Mühe auf sich genommen haben, jeweils mehrere Kapitel zu lesen, und uns ihre Eindrücke und Empfehlungen mitgeteilt haben. Es waren dies Judith Bernauer, Doreen Drechsler, Frederik Haarig, Sebastian Hänsel, David Käthner, Janet Kleber, Sonja Kunze, Cynthia Pönicke, Marcus Schenkel, Katrin Sedlmeier und Corina Ulshöfer. Nicht weniger wertvoll waren die Rückmeldungen unserer Kolleginnen und Kollegen. Wir danken herzlich Martin Baumann, Joachim Engel, Madlen Glauer, Oswald Huber, Georg Jahn, Juliane Kämpfe, Thomas Schäfer, Manfred Wettler und Isabell Winkler. Wir haben ihre Ratschläge nicht immer befolgt, aber über alle (manchmal sehr lange) nachgedacht und manche Rückmeldungen haben uns auf Probleme aufmerksam gemacht, an die wir vorher nicht gedacht hatten. Ein besonderes Dankeschön verdienen Anita Hewer als kritische „fachfremde“ Leserin und Sonja Kunze sowie Corina Ulshöfer, die viele der Abbildungen auf professionelle Weise erstellten oder schon vorhandene deutlich verbesserten. Last not least möchten wir uns auch bei unseren Ansprechpartnern beim Pearson-Verlag, Christian Schneider, Martin Keidel, Mailin Bremer und Stephan Dietrich herzlich für die stets angenehme Arbeitsatmosphäre, ihr Eingehen auf unsere Wünsche und ihr Verständnis für unsere nicht immer optimale Zeitplanung bedanken.

Chemnitz und Erfurt

Peter Sedlmeier & *Frank Renkewitz*

TEIL I

Grundlagen und Konzepte

Alltagswissen versus Wissenschaft: Beispiel Psychologie

1

ÜBERBLICK

» Warum hat Karin ihren Freund verlassen? Wieso hat Tobias während der Vorlesung gelacht? Warum weint das Kind? Wie kommt es, dass ich mich so ärgere? Fragt man jemanden, der in der jeweiligen Situation dabei war, wird diese Person in der Regel eine plausible Antwort parat haben. Wir können das Verhalten von anderen Menschen und auch das von uns selbst – falls wir überhaupt darüber nachdenken – oft erklären, ohne lange nachzudenken. Erklären von Erleben und Verhalten wird häufig als zentrale Aufgabe der Psychologie und der Sozialwissenschaften betrachtet: Jeder scheint somit eine Psychologin oder ein Psychologe zu sein. Wozu brauchen wir also ein langwieriges Studium?

Vielleicht um Fragen zu beantworten, die nicht nur auf eine Person, sondern auf eine Gruppe von Personen oder vielleicht sogar alle Menschen bezogen sind? Wie wirkt sich Ängstlichkeit auf das Ergebnis von mündlichen Prüfungen aus? Warum empfinden wir Gefühle? Ist Frontalunterricht schlechter als Gruppenunterricht? Wie entstehen Vorurteile? Warum entwickeln wir eine Vorliebe für eine bestimmte Art von Musik? Hier kann es schon sein, dass einige Befragte etwas mit ihrer Antwort zögern, aber die meisten werden auch auf solche Fragen schnell Antworten parat haben. In der Tat ist es so, dass sich die wissenschaftliche Psychologie hauptsächlich mit Fragen beschäftigt, die sich nicht nur auf einen Einzelfall beziehen, sondern auf „durchschnittliches“ Verhalten und Erleben. Entsprechende Theorien kann man jedoch auch in Alltagsgesprächen finden. Wenn man aber über durchschnittliches Verhalten eine gute Theorie hat, dann lassen sich auch Aussagen über den Einzelfall machen. Der Unterschied zwischen Alltagspsychologie und wissenschaftlicher Psychologie besteht also offensichtlich nicht darin, ob man eine Aussage über einen Einzelfall oder eine allgemeinere Aussage machen möchte. Was ist es dann, was der wissenschaftlichen Psychologie einen besonderen Stellenwert verleiht?

Die Unterschiede zwischen Alltagspsychologie und wissenschaftlicher Psychologie sind manchmal qualitativer Art. So stimmen die Alltagsvorstellungen darüber, wie das Gedächtnis funktioniert oder was Intelligenz ist, oft nicht mit wissenschaftlichen Erkenntnissen überein. Oft sind die Unterschiede aber eher quantitativ oder graduell. Auch die „Methoden“ der Alltagspsychologie – häufig „Bauchgefühl“ oder „intuitives Urteil“ – führen zwar nicht selten zu richtigen Ergebnissen, aber sie sind in einem sehr viel höheren Maß fehleranfällig als die der wissenschaftlichen Psychologie. Außerdem ist die Sprache, die man in der Wissenschaft gebraucht, deutlich präziser als die im Alltag verwendete. Um diese Unterschiede deutlich zu machen, sehen wir uns zunächst einige besonders auffällige Beispiele für die Fehleranfälligkeit der Alltagspsychologie etwas genauer an: Sie sollen zeigen, was beim Wahrnehmen, Erinnern, logischen Denken und beim Umgang mit Wahrscheinlichkeiten alles schief gehen kann. Dann werden wir uns mit dem Unterschied zwischen Alltags- und Wissenschaftssprache beschäftigen und beschließen das Kapitel mit einer Beschreibung der *wissenschaftlichen Methode*. «

1.1 Die Fallstricke der Alltagspsychologie

Im Alltag hinterfragen wir selten die Art und Weise, wie wir die Welt wahrnehmen und Informationen darüber verarbeiten. Das ist auch meist vernünftig, weil uns ein dauerndes Hinterfragen in unseren Entscheidungen und Handlungsmöglichkeiten drastisch behindern würde. Der Nachteil einer solchen „spontanen" Vorgehensweise ist jedoch, dass wir manchmal Fehler begehen, die wir nicht bemerken. Solche Fehler können zu unzutreffenden Erklärungen menschlichen Verhaltens und Erlebens führen. Hier sind einige Beispiele für solche Fehler, die selbst wieder Gegenstand psychologischer Forschung geworden sind.[1]

1.1.1 Fehler beim Wahrnehmen

Nehmen wir die Welt so wahr wie sie ist? Das scheint zumindest manchmal nicht der Fall zu sein. Was sehen Sie in der linken Zeichnung in ▶ Abbildung 1.1, wenn Sie die beiden waagrechten Striche vergleichen? Wenn es Ihnen so geht wie den meisten Menschen, dann haben Sie den Eindruck, der obere Strich sei länger als der untere. Die beiden sind jedoch gleich lang (prüfen Sie es nach – z.B. mit einem Lineal!). Was ist passiert? Eine gängige Erklärung ist, dass wir bei Abbildungen dieser Art automatisch die Perspektiven-Information berücksichtigen, die in den Pfeilen an den Enden der waagrechten Striche gegeben wird. Die Perspektiven-Information deutet an, dass es sich beim oberen Strich um den hinteren Rand und beim unteren Strich um den vorderen Rand eines Objekts handeln könnte – der obere Strich müsste also weiter von uns entfernt sein als der untere. Unser Wahrnehmungssystem scheint diese angedeutete Entfernungs-Information automatisch mit zu verrechnen. Wenn die Abbilder zweier Objekte in der Netzhaut gleich groß oder lang sind (wie die waagrechten Striche in diesem Fall), aber eines davon weiter entfernt ist als ein anderes, dann muss es größer oder länger sein als das andere. Das ist eigentlich eine großartige (und automatische) Leistung unseres Wahrnehmungssystems, aber in diesem Fall verleitet es uns zu einer fehlerhaften Wahrnehmung, bei der es auch nichts hilft, dass wir die richtige Lösung kennen.

Wie ist es mit der rechten Zeichnung in ▶ Abbildung 1.1? Die meisten Menschen sehen hier zwei Dreiecke: ein weißes Dreieck ohne Rand, das über einem weißen Dreieck mit einem schwarzen Rand liegt. Keines der beiden Dreiecke ist aber tatsächlich vorhanden. Alles was die Zeichnung enthält, sind drei Winkel und drei schwarze „Törtchen", aus denen jeweils ein Stück herausgeschnitten ist. Was geschieht hier? Wir ergänzen Informationen. Auch das ist im Alltag sehr hilfreich, weil Objekte oft durch andere verdeckt sind und wir sie durch dieses Ergänzen unseres Wahrnehmungssystems doch erkennen können. Aber in unserem Beispiel verleitet uns diese Ergänzungsfunktion eben dazu, etwas zu sehen, was gar nicht da ist.

1 Die psychologische Forschung hat allerdings auch herausgefunden, wie zumindest einige dieser Fehler durch kleine Veränderungen in der Art wie Informationen dargeboten werden oder durch Instruktion vermieden werden können (z.B. Gigerenzer et al. 2008; Sedlmeier, 2007; Sedlmeier & Hilton, 2012).

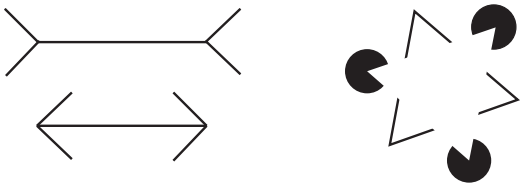


Abbildung 1.1: Zwei Beispiele für optische Täuschungen, links die Müller-Lyer Täuschung und rechts ein Kanisza-Dreieck (beide benannt nach ihren „Entdeckern").

Im Alltagsleben sind unsere Wahrnehmungen nicht isoliert voneinander: Was wir aktuell wahrnehmen, beeinflusst häufig unsere Erwartung dessen, was wir im nächsten Moment wahrnehmen werden. Wenn wir aus einiger Entfernung einen Bratwurststand sehen, erwarten wir, dass es nach Bratwurst riechen wird, wenn wir etwas näher kommen. Auch wenn wir uneindeutige Eindrücke interpretieren müssen, benutzen wir häufig frühere Wahrnehmungen und darauf aufbauende Erwartungen. ▶ Abbildung 1.2 zeigt ein Beispiel hierfür. Was stellt die Figur in ▶ Abbildung 1.2a dar? Wenn Versuchsteilnehmer zunächst die obere Reihe der Zeichnungen in ▶ Abbildung 1.2b sehen, dann sehen die meisten von ihnen das Gesicht eines Mannes, wenn sie aber als erstes die untere Reihe von Zeichnungen in ▶ Abbildung 1.2b gezeigt bekommen, dann sehen sie eher eine kniende Frau.

(a)

(b)

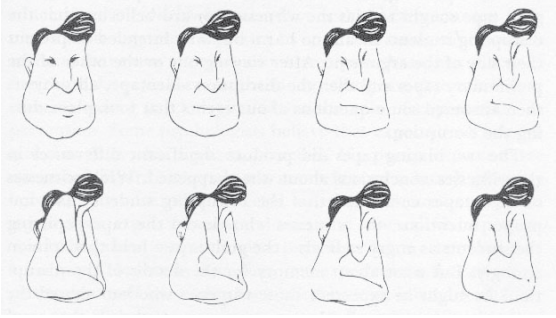


Abbildung 1.2: Ambige Figur a sowie die graduellen Übergänge zwischen zwei eindeutigen Figuren (a) Gesicht eines Mannes und kniende Frau – b) und dieser ambigen Figur (modifiziert nach Loftus, 1979, 45).

Eine fehlerhafte Wahrnehmung kann unter Umständen katastrophale Folgen haben. Das im Kasten „Fehlwahrnehmung mit Todesfolge“ beschriebene Beispiel ist sicher ein Extremfall, aber die Schlussfolgerung, die man daraus ziehen kann, ist dieselbe wie bei den Zeichnungen in ▶ Abbildung 1.1 und ▶ Abbildung 1.2: Unsere Wahrnehmungen sind beeinflusst von unserem Hintergrundwissen, unseren Erfahrungen und Erwartungen. Das ist in vielen Fällen äußerst hilfreich, weil es uns die Wahrnehmung der Welt sehr vereinfacht, aber in manchen Fällen – wenn wir uns dieses Einflusses nicht bewusst sind – führt uns dieser Mechanismus in die Irre und kann unsere alltagspsychologischen Theorien, z.B. dazu, was gerade in unserer Umwelt passiert, beeinflussen.

Fehlwahrnehmung mit Todesfolge

Ein drastisches Beispiel für die Auswirkung von Erwartungen auf die Wahrnehmung, das sich gegen Ende der fünfziger Jahre in Kanada ereignet hat, wird von Sommer (1959) berichtet.

An einem Winternachmittag gingen fünf Männer auf die Hirschjagd. Als sie über ein morastiges Feld fuhren, blieben sie im Schnee stecken und nach einiger Zeit fiel das Getriebe aus. Zwei der Männer erboten sich, zu einem nahe gelegenen Farmhaus zu gehen und Hilfe zu holen. Von den anderen drei Männern blieb einer im Auto und zwei hielten sich vor dem Auto auf. Unterdessen entschied einer von den beiden Männern, die unterwegs zur Farm waren, dass es keinen Grund dafür gebe, dass sie beide Hilfe holten; stattdessen wolle er inzwischen versuchen, einen Hirsch aufzuspüren. Die Männer, die beim Auto geblieben waren, wussten nicht, dass ihr Freund nun an einem Hügel vor ihnen herumstreifte. Einer der beiden Männer, die beim Auto standen, sah, dass sich etwas bewegte und sagte zu seinem Freund: „Das ist ein Hirsch, oder?“ Der Freund antwortete, dass er das auch denke, und der erste Mann schoss auf den Hirsch. Der Hirsch sprang daraufhin vorwärts und schrie – ein Laut, der wie der Schrei eines verwundeten Hirsches klang. Als der Hirsch nun versuchte wegzulaufen, rief der Freund: „Lass ihn nicht entkommen: bitte erleg ihn für mich.“ Der erste Mann feuerte noch einen Schuss ab. Da sich der Hirsch immer noch bewegte, wurde auch noch ein dritter Schuss abgefeuert, der den Hirsch endgültig niederstreckte. Die Männer rannten zu ihm und sahen erst jetzt, dass es überhaupt kein Hirsch war, sondern ihr Freund. Und der Freund war tot.

So merkwürdig diese Geschichte einem Außenstehenden auch vorkommen mag: Die Jäger, die voller Jagdfieber die Gegend nach einem Hirsch durchsuchten, nahmen das sich bewegende Objekt (ihren Freund) als Hirsch wahr. „In meinen Gedanken und in meinen Augen war es ein Hirsch“, sagte einer der Männer in der anschließenden gerichtlichen Befragung.

Dieser Unfall war kein Einzelfall: Fehlwahrnehmungen bei Jägern scheinen – auch angesichts der nicht immer eindeutigen visuellen Signale (Personen sind häufig durch Bäume oder Zweige teilweise verdeckt, Lichtverhältnisse sind nicht immer optimal) immer wieder vorzukommen. So erschoss beispielsweise am 11. Oktober 2004 ein 60-jähriger Jäger im westmecklenburgischen Landkreis Hagenow während der nächtlichen Pirsch versehentlich seinen ein Jahr jüngeren Bekannten. Beide Jäger hatten sich am Rand eines abgeernteten Maisfeldes postiert, aber der Jüngere hatte seinen Posten verlassen und sich seinem Kollegen genähert, der ihn offenbar für Wild hielt (Quelle: *www.n24.de*).

1.1.2 Fehler beim Erinnern

Eine im Alltag verbreitete Vorstellung ist, dass unser Gedächtnis so ähnlich funktioniert wie ein Tonband, ein Fotoapparat oder eine Filmkamera: Wenn wir uns erinnern, dann rufen wir aus einer Art Speicher das ab, was wir gesehen, gehört oder gedacht haben. Das ist, wie man mittlerweile weiß, jedoch eher selten der Fall. Meist rufen wir unsere Vergangenheit nicht ab, sondern *rekonstruieren* sie. Dabei kann unsere Erinnerung in vielfältiger Weise beeinflusst werden. Gut untersucht sind der Einfluss von Information, die wir vor dem Erinnern erhalten, und der Einfluss unserer „impliziten Theorien", das heißt, Vorstellungen und Erwartungen, die wir zu bestimmten Gegenstandsbereichen haben, die uns in der Regel aber nicht bewusst sind.[2]

Ein sehr bekannter Effekt in der Psychologie ist der „Rückschaufehler", der in vielen Studien systematisch untersucht wurde. ▶ Abbildung 1.3 illustriert, was in Studien zum Rückschaufehler normalerweise gemacht wird. Zunächst geben die Untersuchungsteilnehmer ein Urteil oder eine Einschätzung ab, z.B. darüber, welche Note sie in der anstehenden Klausur schreiben werden. Danach bekommen sie die „wahre" Information, z.B. das Ergebnis ihrer Klausur. Wenn man die Teilnehmer nun bittet, sich an ihr ursprüngliches Urteil zu erinnern, erhält man oft Werte, die zwischen dem Original-Urteil und der „wahren" Information liegen (natürlich nur, wenn sich Urteil und „wahre" Information unterscheiden). Wenn jemand ursprünglich gesagt hat, dass er eine 2,7 in der Klausur erwartet und das tatsächliche Ergebnis 2,0 war, dann wird er sich z.B. tendenziell erinnern, damals eine 2,3 erwartet zu haben (Sedlmeier & Jaeger, 2007).

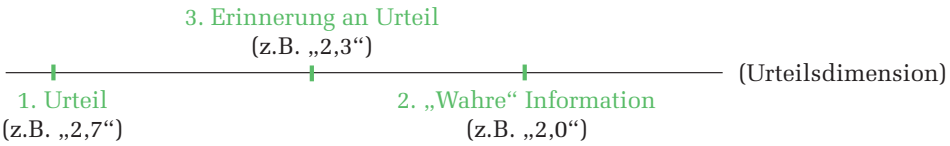


Abbildung 1.3: Typisches Ergebnis in einer Studie zum Rückschaufehler – die Erinnerung an ein Urteil ist ein Kompromiss zwischen dem ursprünglichen Urteil und einer später gegebenen Information.

Die Informationen, die wir bei der Rekonstruktion von Gedächtnisinhalten benutzen, müssen nicht immer so klar sein wie etwa die Information, dass das tatsächliche Ergebnis in der Klausur eine 2,0 war. Manchmal sind Zusatzinformationen auch in der Art und Weise versteckt, wie jemand eine Befragung durchführt. Vor allem wenn sich die Befragten nicht sicher sind, können subtile Informationen in der Frage den Abruf (und auch die Neukonstruktion) von Gedächtnisinhalten beeinflussen. Von einem besonders drastischen Beispiel wird im Kasten „Falsche Erinnerung mit Konsequenzen" berichtet.

2 In der Psychologie wird häufig zwischen impliziten und expliziten Denk- und Urteilsprozessen unterschieden. Explizit sind solche Prozesse, derer wir uns bewusst sind, wenn wir etwa verschiedene Vor- und Nachteile beim Kauf eines Autos gegeneinander abwägen. Impliziten Urteilsprozessen beim Autokauf würden wir folgen, wenn wir uns einfach „nach Gefühl" entscheiden würden.

Falsche Erinnerungen mit Konsequenzen

In den frühen 80er Jahren hatten nach und nach 360 Kinder, die einen Kindergarten – die *McMartin Preschool* – in Kalifornien besuchten oder besucht hatten, berichtet, dass sie in diesem Kindergarten sexuell missbraucht worden seien (siehe z.B. *http://en.wikipedia.org/wiki/McMartin_preschool*). Gegen die Betreuer liefen lange Gerichtsverfahren – der Hauptbeschuldigte saß fünf Jahre in Untersuchungshaft – bis sich 1990 herausstellte, dass sie alle unschuldig waren. Wie war es zu diesem eklatanten Fehlurteil gekommen? Ausgelöst wurde der Fall durch die Mutter eines Kindes, die zur Polizei ging und angab, ihr Sohn sei von einem Betreuer des Kindergartens sexuell missbraucht worden. Sie kam zu dieser Schlussfolgerung, weil ihr Sohn schmerzhafte Blähungen hatte. Der Sohn seinerseits bestritt jedoch, von dem Betreuer belästigt worden zu sein. Die Polizei befragte den Verdächtigen, erhob aber keine Anklage, weil die Beweislage zu gering war. Allerdings schickten die Polizisten einen offenen Brief an etwa 200 Eltern von Kindergartenkindern, in dem sie schrieben, dass ihre Kinder möglicherweise sexuell missbraucht worden seien und sie sie darüber befragen sollten. Dies führte dazu, dass einige hundert Kinder in einer Klinik für Opfer von Missbrauch in Los Angeles (*Children' s Institute International*) zu den Vorgängen im Kindergarten befragt wurden. Es gab Berichte über sexuellen Missbrauch in Autowaschanlagen und in Flughäfen sowie über Nacktfotografien. Aber selbst für Letzteres wurde keinerlei Beweis gefunden. Anscheinend waren viele dieser Erinnerungen hauptsächlich durch die Befragung in der Klinik „entstanden". Die Befragenden gingen offensichtlich meist davon aus, dass der Missbrauch tatsächlich stattgefunden hatte und formulierten ihre Fragen entsprechend. Wie leicht durch den Einfluss geeigneter Fragetechniken falsche Erinnerungen bei Kindern entstehen können, haben Garven et al. (1998) in einer Studie demonstriert, die die Techniken der Klinikmitarbeiter anwandten und damit bei über der Hälfte der befragten Kinder nachweisbar falsche Erinnerungen erzeugten.

Wie können unsere Gedächtnisinhalte durch unsere impliziten Theorien beeinflusst werden? ▶ Abbildung 1.4 zeigt zunächst drei schematische Grafiken darüber, wie wir uns Veränderungen von Einstellungen, Eigenschaften und Fähigkeiten über die Zeit hinweg vorstellen könnten. Das erste Schema stellt eine verbreitete Ansicht über politische Einstellungen dar: Sie ändern sich nicht über die Zeit. Das zweite Schema beschreibt, was man von einem effektiven Trainingsprogramm erwartet: Kenntnisse und Fertigkeiten verbessern sich über die Zeit hinweg. Das dritte Schema schließlich zeichnet nach, wie man sich die Entwicklung vieler Fähigkeiten und Fertigkeiten im Laufe des Leben vorstellt: So steigt nach Meinung der meisten Menschen die intellektuelle Leistungsfähigkeit im Laufe der Entwicklung erst an, erreicht dann im mittleren Erwachsenenalter ihren Höhepunkt und sinkt dann wieder mit zunehmendem Alter.

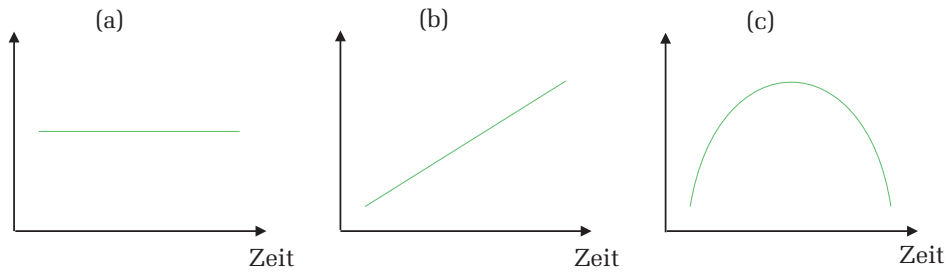


Abbildung 1.4: Darstellung dreier impliziter Theorien zu Veränderungen über die Zeit hinweg.

Wie können diese impliziten Theorien nun zu Gedächtnisfehlern führen? Diese Frage wurde in zahlreichen Studien untersucht (siehe Ross, 1989). Man hat beispielsweise in den USA Wahlberechtigte im Abstand von 20 Jahren nach ihren politischen Präferenzen (Republikaner oder Demokraten) gefragt. Dabei stellte sich heraus, dass viele der Befragten beim zweiten Befragungstermin (die Befragten waren inzwischen 40 bis 50 Jahre alt) fälschlicherweise meinten, sie hätten vor 20 Jahren schon dieselbe politische Einstellung vertreten. Was passiert, wenn man der Ansicht ist, ein Trainingsprogramm zur Verbesserung der Studienleistungen sei effektiv, obwohl es keinen Effekt hat? Die Befragten können ihre Leistungen nach dem Training realistisch einschätzen. Wenn sie sich zu diesem Zeitpunkt erinnern sollen, wie gut ihre entsprechenden Studienleistungen vor dem Training waren, tendieren sie dazu, diese Leistungen *vor* dem Training systematisch zu unterschätzen. Ähnliche Ergebnisse fand man auch zu weiteren impliziten Theorien, wie der in Kurve 3 in ▶ Abbildung 1.4. Generell scheinen implizite Theorien zu Veränderungen über die Zeit so zu funktionieren, dass man vom Status Quo ausgeht und seine Gedächtnisinhalte in Übereinstimmung mit der verwendeten impliziten Theorie angleicht.

Beide hier beschriebenen Einflussmöglichkeiten auf unser Gedächtnis, nachträgliche Informationen über Ereignisse und unsere impliziten Theorien, spielen im Alltag eine große Rolle und haben damit auch das Potenzial, unsere alltagspsychologischen Theorien stark zu beeinflussen.

1.1.3 Fehler beim logischen Denken

Ist logisches Denken schwierig? Einige Befunde aus der psychologischen Urteilsforschung legen diese Schlussfolgerung nahe. Hier ist ein Beispiel – eine Variante einer bekannten Aufgabe (bekannt unter der Bezeichnung „Wason-Aufgabe“) zum logischen Denken. Nehmen Sie an, Sie haben vier Karten vor sich wie die in ▶ Abbildung 1.5. Alle Karten sind auf der einen Seite mit einem Buchstaben und auf der anderen Seite mit einer Zahl beschriftet. Sie sollen jetzt überprüfen, ob folgende Regel für die Karten in ▶ Abbildung 1.5 stimmt: *Wenn auf der einen Seite der Karte ein Konsonant ist, ist auf der anderen eine ungerade Zahl.* Die Frage ist nun, welche Karte(n) Sie unbedingt umdrehen müssen, um zu überprüfen, ob diese Regel verletzt wurde. Versuchen Sie erst, die richtige Lösung zu finden, bevor Sie weiterlesen!

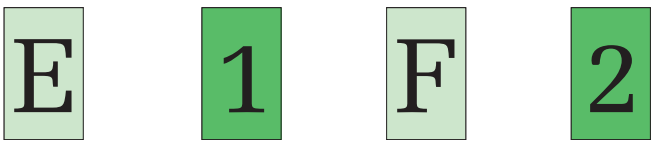


Abbildung 1.5: Eine Variante einer bekannten Logik-Aufgabe.

Die Lösungsraten bei Aufgaben dieser Art sind nicht sehr hoch – nur etwa 20% der Befragten kommen auf die richtige Lösung. Richtig ist: Man muss die Karte mit dem „F“ und die mit der „2“ umdrehen. Warum? Wenn auf der anderen Seite der „F“-Karte eine gerade Zahl ist, dann ist die Regel verletzt, genauso wenn auf der anderen Seite der „2“-Karte ein Konsonant steht. Auf der anderen Seite der „1“-Karte kann stehen,

was will: Bei einem Konsonanten passt die Regel und bei einem Vokal ist sie nicht anwendbar; deswegen macht es auch nichts, ob auf der Rückseite der „E"-Karte eine gerade oder ungerade Zahl steht.

Logik wird oft als Bindeglied zwischen verschiedenen zusammengehörenden Aussagen benutzt, die dann insgesamt eine Theorie ergeben. Wenn die Anwendung logischer Schlussfolgerungen fehlerhaft ist – und das scheint im Alltag manchmal der Fall zu sein –, dann hat das natürlich auch Auswirkungen auf eine entsprechende alltagspsychologische Theorie.

1.1.4 Fehler beim Umgang mit Wahrscheinlichkeiten

In *Abschnitt 1.1.2* haben wir gesehen, dass wir nicht immer sicher sein können, dass sich das, woran wir uns erinnern, auch genauso ereignet hat. Noch schlimmer steht es mit Vorhersagen: Bei genauerer Betrachtung müssen wir feststellen, dass wir kaum etwas mit absoluter Sicherheit vorhersagen können. Oft hilft es aber, zumindest über die Wahrscheinlichkeit, mit der ein bestimmtes Ereignis eintritt, gut Bescheid zu wissen. Wie gut sind wir im Umgang mit solchen Wahrscheinlichkeiten? Unter bestimmten Bedingungen scheinen wir auch dabei recht fehleranfällig zu sein. Sehen wir uns folgendes Beispiel an:

> *Ein Schüler hat im Abschlusszeugnis in Mathematik die Note Vier erzielt. Welche der folgenden Aussagen trifft für ihn mit größerer Wahrscheinlichkeit zu?*
>
> *a) Er hatte eine Sechs in Mathe im Halbjahreszeugnis.*
>
> *b) Er hatte eine Sechs in Mathe im Halbjahreszeugnis, hat aber im zweiten Halbjahr Nachhilfe in Mathe erhalten.*

Bei dieser Aufgabe entscheiden sich die meisten Befragten regelmäßig (und fälschlicherweise) für die Alternative b. Auch Aufgaben der folgenden Art scheinen nicht so einfach zu sein:

> *Welches der folgenden Ereignisse ist wahrscheinlicher?*
>
> *a) Eine Hausfrau hat promoviert.*
>
> *b) Eine promovierte Frau ist Hausfrau.*
>
> *c) Beides ist gleich wahrscheinlich.*

Die häufigste Antwort ist, dass beide Ereignisse gleich wahrscheinlich sind (Antwort c), obwohl tatsächlich die zweite Aussage (eine promovierte Frau ist Hausfrau) die wahrscheinlichere ist. Wenn Ihnen nach wie vor unklar ist, warum im ersten Beispiel die Antwort a und im zweiten die Antwort b richtig ist, dann können Sie die Lösungen in *Kapitel 10* finden, in dem wir uns ausgiebig mit Wahrscheinlichkeiten befassen. Festzuhalten bleibt, dass es im Alltag nicht immer einfach ist, Wahrscheinlichkeiten richtig einzuschätzen: ein weiteres Argument dagegen, auf alltagspsychologische Theorien zu sehr zu vertrauen.

1.2 Sprachgebrauch in Alltag und Wissenschaft

Abgesehen von den in *Abschnitt 1.1* angedeuteten Fehlermöglichkeiten gibt es im Alltag eine weitere Quelle für Missverständnisse und fehlerhafte Interpretationen: die Umgangssprache. Unsere Äußerungen im Alltag sind oft unvollständig und nicht grammatisch. Trotzdem haben wir in der Regel keine Probleme, unsere Gesprächspartner zu verstehen – der Kontext hilft uns dabei: Wir besitzen fast immer Vorwissen über den Gegenstand einer Unterhaltung und auch über unsere Gesprächspartner und benutzen außerdem viele unterschiedliche Hinweise, wie z.B. Gestik und Mimik, um fehlerhafte und unvollständige Aussagen zu verstehen. In der Wissenschaft geht es jedoch oft gerade darum, Gesetzmäßigkeiten zu finden, also verallgemeinernde Aussagen zu machen, bei denen der Kontext keine Rolle spielen soll. Dieses Ziel wird durch die Präzisierung der Wissenschaftssprache verfolgt.

1.2.1 Missverständnisse beim Verstehen von Sprache im Alltag

Unsere Alltagssprache ist oft nicht ganz eindeutig. Was bedeutet es beispielsweise, wenn jemand sagt: „Das war ein teures Abendessen" – wie viel wird das Essen gekostet haben? Offensichtlich ist es zur Beantwortung dieser Frage notwendig, etwas mehr über den Sprecher zu wissen. Wenn es ein Schüler gesagt hat, dann wird „teuer" etwas anderes bedeuten, als wenn dieser Satz vom Vorstandsvorsitzenden einer Bank stammt. Ein anderes Beispiel: Ein Freund fragt Sie „Hast du ein Taschentuch?" Wenn Sie ein überzähliges (Papier-)Taschentuch haben, werden Sie sich in der Regel nicht darauf beschränken „Ja" zu sagen, sondern ihm eines geben. Unklarheiten dieser Art können zu Missverständnissen führen, wenn man nichts über Sprecher und Kontext weiß. Viele Ausdrücke in der Umgangssprache sind mehrdeutig – Beispiele sind Adjektive wie „groß", „viel", „stark" usw. Die Mehrdeutigkeit von Wörtern wird auch deutlich, wenn man Texte von einer Sprache in eine andere übersetzen möchte. Ein Beispiel, das in vielen Büchern zur Kritik der Künstlichen-Intelligenz-Forschung zitiert wird, illustriert diese Schwierigkeit. Danach soll von einem automatischen Übersetzungsprogramm der Satz „Der Geist ist willig, aber das Fleisch ist schwach" erst ins Russische und dann wieder zurück übersetzt worden sein mit folgendem Ergebnis: „Der Wodka ist gut, aber das Steak ist lausig".

Die Mehrdeutigkeit der Umgangssprache eröffnet Möglichkeiten für systematische Missverständnisse, seien sie gewollt oder ungewollt. Auch dazu gibt es umfangreiche Forschungsergebnisse (siehe z.B. Harris & Monaco, 1978). So hat man herausgefunden, dass kleine Unterschiede in der Formulierung von Fragen die Ergebnisse stark verändern können. Wenn Patienten etwa gefragt werden: „Haben Sie *häufig* Kopfschmerzen, und wenn ja, wie oft?", dann fallen die Angaben wesentlich höher aus, als wenn die Frage lautet: „Haben Sie *gelegentlich* Kopfschmerzen, und wenn ja, wie oft?". Ähnliche Unterschiede erhält man, wenn man Augenzeugen bei einem Unfall Fragen darüber stellt, wie schnell die Autos fuhren. „Wie schnell waren die Autos ungefähr, als sie *aufeinander krachten*?", erzeugt in der Regel deutlich höhere Geschwindigkeitsschätzungen als „Wie schnell waren die Autos ungefähr, als sie *kollidierten*?". Gewollte Missverständnisse kann es etwa bei Aussagen vor Gericht geben. Wenn beispielsweise

Mitglieder einer Jury einschätzen sollen, ob der Angeklagte Geld gestohlen hat oder nicht, macht es anscheinend keinen Unterschied, ob dieser antwortet: „Ich habe das Geld nicht gestohlen“ oder „Ich war nicht gezwungen, das Geld zu stehlen“. In der von Harris und Monaco (1978) berichteten Studie hat offenbar stattgefunden, was die Autoren „pragmatische Implikation“ nennen: Uneindeutige oder fehlende Information in Alltagskonversationen wird vom Hörer in passender Weise „ergänzt“.

Die Umgangssprache bildet zudem logische Strukturen nicht genau so ab, wie es in der Sprache der Logik geschieht. Wenn jemand sagt: „Bei dem Unfall gab es 22 Tote und Verletzte“, dann ist das „und“ logisch gesehen ein „oder“: Jede der 22 Personen war entweder tot oder verletzt. Wenn in der Aussagenlogik der Ausdruck „und“ gebraucht wird, heißt das, dass beides (z.B. „tot“ und „verletzt“) zutreffen muss, damit die Aussage wahr ist. Selbst ein „oder“ ist in der Umgangssprache oft nicht eindeutig. Bei der Frage „Willst du etwas trinken oder essen?“, kann das „oder“ die Bedeutung von „entweder oder“ (nur trinken oder nur essen) haben oder die von „sowohl als auch“ (nur trinken, nur essen oder beides). Auch das logische „und“ muss in der Umgangssprache nicht als solches benutzt werden: In „Er ist intelligent, aber er arbeitet sehr langsam“, steht das „aber“ für ein logisches „und“.

Was sollen diese Beispiele zeigen? Sie sollen illustrieren, dass Aussagen in der Umgangssprache oft mehrdeutig sind und somit Manipulationen des Sprachverständnisses ermöglichen. Das ist natürlich keine gute Grundlage für eine wissenschaftliche Beschreibung von Erleben und Verhalten. Um Aussagen zu präzisieren, haben die Wissenschaften deswegen eigene Fachwörter entwickelt, und benutzen, wenn es möglich ist, die Sprache der Logik und mathematische Elemente.

1.2.2 Präzisierung der Sprache in der Wissenschaft

In den Anfangssemestern haben Studierende aller Fachrichtungen mit vielen neuen Ausdrücken zu kämpfen und bezweifeln manchmal, ob diese auch alle notwendig sind. Auch wenn es hin und wieder unnötiges „Fachchinesisch“ geben mag: Die Verwendung von Fachausdrücken stellt sicher, dass Verwechslungen und Mehrdeutigkeiten, die beim Benutzen der Umgangssprache entstehen können, weniger wahrscheinlich werden. In der wissenschaftlichen Psychologie gibt es eine Vielzahl von Fachausdrücken. Viele davon betreffen die Thematik dieses Buches: Forschungsmethodik und Statistik. Manche der Fachausdrücke kommen auch sporadisch in der Umgangssprache vor, sind aber in der Wissenschaft eindeutig (und anders) definiert, wie etwa „Signifikanz“, „Neurotizismus“, „Faktor“ usw. Andere Ausdrücke sind spezifisch für bestimmte Fächer. In der wissenschaftlichen Psychologie geht man teilweise so weit, dass man Persönlichkeitseigenschaften nur mit Abkürzungen wie „PF1“ oder „PF2“ bezeichnet, um Mehrdeutigkeiten zu verhindern.

In *Kapitel 2* werden wir sehen, dass Mehrdeutigkeiten auch dadurch minimiert werden, dass man die Operationen angibt, mithilfe derer bestimmte Eigenschaften oder Merkmale gemessen werden können. Das ist deswegen nötig, weil Ausdrücke, die wir wie selbstverständlich im Alltag benutzen – wie „Gedächtnis“, „Angst“ oder „Intelligenz“ – nichts bezeichnen, was wir direkt sehen könnten. Wir nehmen im Alltag ein-

fach an, dass alle dasselbe darunter verstehen (was nicht immer der Fall ist). In der Wissenschaft muss man sicherstellen, dass alle dasselbe Verständnis von Begriffen haben. Deswegen ist es notwendig, genau anzugeben, wie man beispielsweise das Ausmaß von „Angst" feststellen kann.

Wann immer es angebracht ist, versucht man in der Wissenschaft, Aussagen, Hypothesen oder sogar ganze Theorien dadurch zu präzisieren, dass man sie logisch eindeutig formuliert. Dies ist in der Psychologie nicht immer einfach. Aber je besser es möglich ist, eine Hypothese zu formalisieren, also entweder in der Sprache der Logik oder als mathematische Gleichung auszudrücken, desto besser kann man sie überprüfen. Logik dient manchmal auch dazu, die Struktur einer Aufgabe klar herauszuarbeiten. Ein Beispiel: Wie kann man die richtige Lösung des Karten-Problems im *Abschnitt 1.1.3* mithilfe einer logischen Formulierung darstellen? ▶ Abbildung 1.6 zeigt eine sogenannte „Wahrheitstafel", wie sie häufig in der Aussagenlogik benutzt wird. Die Aufgabe in *Abschnitt 1.1.3* lautete: *Wenn auf der einen Seite der Karte ein Konsonant ist, ist auf der anderen eine ungerade Zahl.* Nun lassen wir *A* für den „Wenn-Teil" der Aussage (Konsonant) und *B* für den „Dann-Teil" (ungerade Zahl) stehen. In der Wahrheitstafel stehen jetzt unter *A* verschiedene Möglichkeiten: „Ja" bedeutet Konsonant und „Nein" bedeutet Vokal (kein Konsonant). Unter *B* steht „Ja" für „ungerade Zahl" und „Nein" für „gerade Zahl" (keine ungerade Zahl). In der letzten Spalte steht, wann die gesamte Wenn-Dann-Aussage, die sogenannte logische Implikation richtig ist. Diese Implikation ist nur dann nicht wahr (*A→B: nein*), wenn auf der einen Seite der Karte ein Konsonant steht (*A: ja*), aber auf der anderen eine gerade Zahl (*B: nein*). Man muss die Karte also immer dann umdrehen, wenn man entweder einen Konsonanten oder eine gerade Zahl sieht (also die Karten „F" und „2" in ▶ Abbildung 1.5), denn nur dann kann die Implikation falsch sein; und nur wenn die Implikation falsch ist, kann die Regel verletzt sein.

A	B	A → B
ja	ja	ja
ja	*nein*	*nein*
nein	ja	ja
nein	nein	ja

Abbildung 1.6: Wahrheitstafel für die logische Implikation. Die Implikation ist nur falsch, wenn der Wenn-Teil (A) wahr ist und der Dann-Teil (B) falsch (siehe zweite Zeile).

Noch häufiger als logische Ausdrücke benutzt man in der wissenschaftlichen Psychologie mathematische Gleichungen. Der Kasten über das Weber'sche Gesetz zeigt ein einfaches Beispiel dafür, wie man psychologische Gesetzmäßigkeiten in Form von Gleichungen ausdrücken kann.

Weber'sches Gesetz

Ab welchem Unterschied im Zuckergehalt merkt man, dass eine Sorte Cola süßer ist als die andere? Ab welchem Unterschied im Gewicht bemerkt man ohne Waage, ob eine Münze schwerer ist als eine zweite (eine Fälschung beispielsweise), oder ab welchem Unterschied im Schalldruck hört sich ein Ton lauter an als ein anderer? Solche Fragen nach subjektiven Unterschiedsschwellen bei physikalischen und chemischen Reizen kann man, zumindest in einem gewissen Intensitätsbereich, gut mit dem sogenannten Weber'schen Gesetz erklären. Dieses Gesetz lautet, dass die Größe der Unterschiedsschwelle sich proportional zur Intensität eines Vergleichsreizes verhält. Wenn Sie von diesem Gesetz noch nichts gehört haben, müssen Sie wahrscheinlich etwas über die Bedeutung des letzten Satzes nachgrübeln. Für die meisten Leser wird es jedoch einfacher und auf alle Fälle präziser, wenn dieses Gesetz als Gleichung dargestellt wird:

$$\frac{\Delta I}{I} = k$$

Der Ausdruck k ist eine Konstante – die Weber'sche Konstante – die angibt, um welchen Anteil man die Reizintensität I vergrößern muss, um den gerade merklichen Unterschied ΔI zu erhalten. Die Konstante k unterscheidet sich für verschiedene Reizdimensionen. Für die Geschmackskonzentration gilt $k = 0{,}20$. Wenn ich etwa eine Cola mit 10%iger Zuckerkonzentration als Vergleichsreiz hätte, bräuchte ich, um festzustellen, dass eine andere süßer ist, eine Erhöhung der Zuckerkonzentration um $k \cdot I = 0{,}2 \cdot 10\% = 2\%$. Nach dem Weber'schen Gesetz könnte ich also eine Cola mit einem Zuckergehalt von 10% von einer mit 12% unterscheiden. Wenn jedoch die Zuckerkonzentration in der ersten Cola schon 50% wäre, würde ich den Unterschied bei einer zweiten erst bemerken, wenn die Konzentration um zehn Prozentpunkte angestiegen wäre.

In der Umgangssprache würde man das Weber'sche Gesetz wohl mithilfe von Beispielen beschreiben oder versuchen, es anschaulich zu formulieren. Als Gleichung – ein fester Bestandteil der Wissenschaftssprache – ist die Aussage aber eindeutiger, präziser und, wenn man sich mit Gleichungen etwas angefreundet hat, auch leichter zu verstehen.

Eine weitere Möglichkeit, die Wissenschaftssprache zu präzisieren, von der auch in der Psychologie immer mehr Gebrauch gemacht wird, ist, Theorien und Vorhersagen als Computerprogramme abzubilden (siehe *Kapitel 31*). Die dabei benötigten Programmcodes zwingen zur Präzision: sie *müssen* präzise sein, weil sonst das Programm nicht funktioniert.

1.3 Die wissenschaftliche Methode

Das Bestreben der Wissenschaftler, zu fundierten Aussagen über die Welt zu gelangen, die außerdem präzise und möglichst wenig fehlerbehaftet sind, hat zur Entwicklung einer Vielzahl von Vorgehensweisen, Verfahren und Techniken geführt, die in ihrer Gesamtheit als *wissenschaftliche Methode* bezeichnet werden. Die wissenschaftliche Methode wurde zuerst in den Naturwissenschaften angewandt, ist aber mittlerweile die anerkannte methodische Grundlage in nahezu allen Bereichen der psychologischen Forschung (auf einen alternativen Ansatz werden wir in *Kapitel 2* und in *Kapitel 32* noch zu sprechen kommen). Nicht alle Wissenschaftler verstehen genau dasselbe unter „wissenschaftlicher Methode" – so benutzen etwa Physiker andere Verfahren als Biolo-

gen oder Psychologen – aber die meisten würden wohl der Zusammenfassung in ▶ Abbildung 1.7 zustimmen. Die in der Abbildung aufgeführten Bestandteile finden sich auch in diesem Buch wieder und werden in späteren Kapiteln behandelt. Die Breite des jeweiligen Bestandteils repräsentiert dabei den Allgemeinheitsgrad des entsprechenden Schrittes im Forschungsprozess, nicht unbedingt jedoch den damit verbundenen zeitlichen Aufwand. Sehen wir uns zunächst kurz an, was sich hinter den Begriffen verbirgt.

Abbildung 1.7: Die wissenschaftliche Methode.

1.3.1 Theorien, Hypothesen und ihre Präzisierung

In *Kapitel 2* werden wir uns genauer damit befassen, was eine Theorie ist und welche Arten von Theorien es gibt. Wir werden auch diskutieren, welche Voraussetzungen und Bestandteile eine Theorie enthalten und wie man bei ihrer Überprüfung vorgehen sollte. Eine Theorie ist ein systematisches Gefüge von Ideen und Annahmen über einen definierten Gegenstandsbereich. Die Bestandteile der Theorie sollten dabei so präzise wie möglich sein. In der Regel kann man aber eine Theorie nicht direkt prüfen, weil sie zu komplex ist. Deshalb werden aus Theorien Hypothesen abgeleitet, die in einem weiteren Schritt so präzisiert werden, dass man sie direkt in empirischen Studien überprüfen kann. Hypothesen sind Annahmen oder Behauptungen, die man auch in Form einer Frage formulieren kann, und „empirisch" bedeutet, dass man Daten sammelt, die Aufschluss über die untersuchten Fragestellungen oder Hypothesen geben können. Die Hypothesen beziehen sich oft nicht nur auf das Verhalten oder Erleben der Teilnehmer einer bestimmten Studie, sondern auf die Grundgesamtheit (Population), aus der diese Teilnehmer stammen. So möchte man beispielsweise aufgrund der Ergebnisse einer Studie mit zufällig ausgewählten Schulkindern in einem Bundesland (Stichprobe) Schlüsse über alle Schulkinder in diesem Bundesland (Population) ziehen. Wenn sich solche Hypothesen auf „Statistiken", das heißt, auf zusammengefasste Werte wie beispielsweise Mittelwerte, beziehen, dann spricht man häufig von statistischen Hypothesen.

1.3.2 Design

„Design“ oder „experimentelles Design“ hat in der psychologischen Forschung eine andere Bedeutung als in der Alltagssprache: es bedeutet „Versuchsplanung“. Das Design einer Studie gibt an, wie die Studie durchgeführt werden sollte, und ist im Idealfall ausschließlich durch die Hypothese bestimmt. Die meisten Designs in der psychologischen Forschung befassen sich mit der Untersuchung von Gruppen. Das hängt damit zusammen, dass menschliches Verhalten und Erleben durch sehr viele Faktoren beeinflusst wird, die man oft gar nicht untersuchen möchte, und dass sich die Einflüsse dieser Faktoren „ausmitteln“, wenn man nicht Einzelwerte, sondern Mittelwerte betrachtet. Wenn man etwa im Rahmen von Ländervergleichen die mathematischen Fähigkeiten einer bestimmten Schülergruppe erheben möchte, dann können die Ergebnisse der einzelnen Schüler in einem entsprechenden Rechentest auch dadurch beeinflusst sein, wie ängstlich, intelligent, ausgeschlafen, hungrig, gut gelaunt usw. diese Schüler sind. Wenn beispielsweise die gute Laune eines Schülers das Ergebnis im Rechentest positiv und die schlechte Laune eines anderen Schülers es negativ beeinflusst, dann bekommt man durch das Mitteln einen Effekt, in dem die Auswirkung der Laune „kontrolliert“ ist. Das setzt natürlich voraus, dass man keine systematische Auswahl von Schülern mit besonders guter oder schlechter Laune getroffen hat. Das wiederum wird durch besondere Verfahren bei der Auswahl von Untersuchungsteilnehmern gewährleistet. „Design“ ist also eine zusammenfassende Bezeichnung für die Auswahl von Teilnehmern, die Wahl von Gruppen, die Methoden zur Datenerhebung, die zeitliche Struktur sowie weitere Aspekte bei der Durchführung von empirischen Studien (siehe *Kapitel 5*).

1.3.3 Durchführung von Studien

Bei der Durchführung einer Studie tritt der Untersucher zum ersten Mal direkt in Kontakt mit den Untersuchungsteilnehmern. Im Idealfall folgt die Durchführung in allen Punkten dem geplanten Design. In der Praxis kommt es allerdings häufig zu unvorhergesehenen Problemen. So können Versuchsteilnehmer ihren Termin versäumen, weil sie krank geworden sind oder es sich anders überlegt haben, Geräte können plötzlich nicht funktionieren oder anders, als es eigentlich gedacht war. Darüber hinaus könnten die Ergebnisse durch das Verhalten der Versuchsleiter oder durch die Zuweisung der Teilnehmer zu den Untersuchungsbedingungen systematisch beeinflusst werden. Gar nicht so selten entwickeln Versuchsteilnehmer auch selbst „Theorien“ darüber, was in der Studie untersucht werden soll, und solche Theorien können sich auf ihr Verhalten auswirken. Außerdem können ethische Probleme auftreten (z.B.: Wann ist es erlaubt, Versuchsteilnehmern zunächst nicht die ganze Wahrheit über die Studie zu sagen?). Eine ausführliche Diskussion der Problematik findet man in Sarris und Reiß (2012).

1.3.4 Datenanalyse und -interpretation

Der größte Teil dieses Buches wird sich damit befassen, wie man die Ergebnisse einer Studie analysieren und interpretieren kann. In *Kapitel 3* werden wir sehen, dass Zahlen nicht gleich Zahlen sind, sondern dass ihre Bedeutung erst durch Zuordnungsregeln zu beobachtbaren oder angenommenen Ereignissen entsteht. Abhängig von der

Bedeutung einer Zahl können dann unterschiedliche Verfahren zur weiteren Bearbeitung verwendet werden. In den meisten Fällen wird man versuchen, Zahlen zusammenzufassen, um ein erstes Bild davon zu erhalten, was in einer Studie herausgekommen ist (*Kapitel 6*). Die weitere Datenanalyse hängt dann noch stärker von der Ausgangshypothese ab. Dabei wird man zunächst die Ergebnisse in der Stichprobe beschreiben, aber nicht selten möchte man auch Schlüsse über die Grundgesamtheit oder Population ziehen. Die erste Art der Analyse wird häufig als *Deskriptive* oder *Explorative Statistik* bezeichnet (*Kapitel 6* bis *9* und *Kapitel 26*) und die zweite als *Inferenzstatistik* (siehe insbesondere die *Teile III* und *IV* dieses Buchs). Bei der Behandlung von Analysemethoden werden wir uns nicht darauf beschränken, „Kochbuchrezepte" zu geben, sondern wir werden versuchen, die Leser zum kritischen Anwenden der entsprechenden Methoden zu motivieren. Dies versuchen wir u.a. auch dadurch, dass wir Verfahren erläutern, die bisher kaum in einführenden Methodenlehrbüchern zu finden sind, die aber unseres Erachtens eine weitere Verbreitung verdienen (*Kapitel 26* bis *32*).

Die Interpretation der Ergebnisse hängt immer von der Hypothese und damit letztlich von der Theorie ab, die der Ausgangspunkt für die Forschung war. Aber auch die Art und Güte der Daten sowie die verwendeten Verfahren haben Einfluss auf die Interpretierbarkeit der Ergebnisse. Es kann durchaus sein, dass man Ergebnisse zunächst einmal nicht interpretieren kann, weil sie den vorhandenen Hypothesen widersprechen. In diesem Fall – und das ist möglicherweise die interessanteste Art von Ergebnissen – wird man versuchen, die vorhandene(n) Theorie(n) zu erweitern oder eine neue zu kreieren, die dann natürlich wieder der Überprüfung ausgesetzt werden muss (symbolisiert durch den Rückkopplungs-Pfeil in ▶ Abbildung 1.7).

1.4 Was gewinnen wir durch die wissenschaftliche Vorgehensweise?

Dieses Kapitel dreht sich im Grunde um die Frage, warum wir eine wissenschaftliche Psychologie brauchen, obwohl wir in gewisser Weise doch alle schon Psychologen sind: Wir können Verhalten und Erleben oft spontan erklären und vorhersagen. Unsere Antwort – illustriert anhand einiger Beispiele – ist: Die Methoden, die wir in der Alltagspsychologie anwenden – wie wir die Welt und uns selbst wahrnehmen, uns erinnern, Schlussfolgerungen ziehen, Wahrscheinlichkeiten schätzen und unsere Sprache benutzen –, sind potenziell ungenau und mit Fehlern behaftet. Das kann zu falschen Alltagstheorien, etwa über die Funktionsweise des Gedächtnisses führen. Alltagstheorien müssen allerdings nicht falsch sein. Aber auch dann gewährleistet nur die wissenschaftliche Methode, dass aus solchen Theorien plausible Hypothesen abgeleitet und sorgfältig überprüft werden können.

Die Überprüfung von alltagspsychologischen Theorien wird neben den oben angeführten methodischen Einschränkungen auch noch durch einige verbreitete aber wenig effektive „Strategien" erschwert. Wie überprüfen wir Theorien im Alltag? Wenn wir sie überhaupt überprüfen, dann verlassen wir uns manchmal ausschließlich auf unser „Gefühl": „Mein Gefühl sagt mir ganz eindeutig, dass der Mensch nicht vom Affen abstammt". Eine andere Art von suboptimaler Überprüfung, die man eher Rechtfertigung nennen könnte, ist die Berufung auf Autoritäten. Diese Autoritäten

sind häufig Personen, die in der Medienlandschaft präsent, aber nicht unbedingt auch in entsprechenden Fachkreisen anerkannt sind. „Fernsehpsychologen", „psychologische Ratgeber" in Illustrierten und viele prominente Erziehungsratgeber – etwa Nachrichtensprecherinnen oder Ehepartner von Politikern – fallen in diese Kategorie. Auch Nobelpreisträger der Naturwissenschaften scheinen sich manchmal für Universalexperten im Bereich der Psychologie und der Sozialwissenschaften zu halten. Eine weitere Vorgehensweise, um Alltagstheorien auf vermeintlich feste Füße zu stellen, ist das Anführen einiger möglichst lebendig dargestellter Beispiele. Ein oder zwei Beispiele für Sozialhilfeempfänger, die in sonnigen Ländern leben, können Alltagstheorien über das Verhalten von Sozialhilfeempfängern stärker beeinflussen als Statistiken, die Tausende von Beziehern der Sozialhilfe repräsentieren.

Eine verbreitete Methode, um unsere Alltagstheorien zu stützen, besteht darin, nur nach bestätigender Evidenz zu suchen. Diese konfirmatorische (bestätigende) Suche ist im Alltag durchaus sinnvoll und vereinfacht unser Leben, führt aber auch dazu, dass falsche alltagspsychologische Theorien nicht auffallen. Wenn man beispielsweise nach dem Essen von Kirschen nie Wasser trinkt, kann man auch nie herausfinden, dass das Trinken von Wasser *kein* Bauchweh erzeugt. Ein weiterer Grund, weswegen falsche Theorien im Alltag nicht auffallen, ist, dass auch aus falschen Theorien manchmal richtige Vorhersagen abgeleitet werden können (die man dann konfirmatorisch prüft). Auch wenn man der Ansicht ist, die Erde sei eine Scheibe, kann man mit einigen Zusatzannahmen die Stellung der Sonne im Jahresverlauf gut vorhersagen.

Was gewinnen wir also durch die Anwendung der wissenschaftlichen Methode? In einigen Fällen möglicherweise inhaltlich nicht viel. Manchmal treffen alltagspsychologische Theorien ja auch zu. Aber selbst in diesem Fall können wir nach einer methodisch fundierten Prüfung Aussagen über die entsprechende (präzisierte) Theorie mit größerer Sicherheit machen. Wenn die entsprechende Alltagstheorie aber nicht oder nur bedingt stimmt oder wenn sie nur teilweise vorhanden ist, dann hat die wissenschaftliche Vorgehensweise den unschätzbaren Vorteil, dass sie eine systematische und unvoreingenommene Prüfung der Gründe, Ursachen und Ziele menschlichen Verhaltens und Erlebens und eine systematische Präzisierung und Weiterentwicklung guter Theorien erlaubt. Fundierte und präzise Methoden sind kein Selbstzweck; sie sind unabdingbare Voraussetzung für wissenschaftlichen Erkenntnisfortschritt, der auch inhaltlich weit über die Themen der Alltagspsychologie hinausgeht.

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

Anders als in vielen naturwissenschaftlichen Bereichen, wie etwa der Atomphysik oder der Biochemie, wo das nur sehr eingeschränkt möglich ist, können Menschen über nahezu alle psychologischen und sozialwissenschaftlichen Themen ohne spezielle Ausbildung mitreden. Alltagspsychologische Erkenntnisse werden jedoch häufig mit Methoden wie „Intuition", „Gefühl" oder „Hörensagen" erreicht. Eine solche Vorgehensweise kann zu richtigen Theorien führen, birgt jedoch auch die Gefahr von Fehlinterpretationen. Die Art und Weise, wie wir die Welt wahrnehmen, uns erinnern, logische Schlüsse ziehen und mit unsicherer Information umgehen, ist potenziell fehlerbehaftet und kann zu falschen Theorien führen sowie theoretische Weiterentwicklungen blockieren. Ein weiteres Problem der Alltagspsychologie besteht darin, dass unsere Alltagssprache oft mehrdeutig und offen für Missverständnisse ist.

Die Wissenschaft bemüht sich dagegen, ihre Sprache zu präzisieren und wendet beim Ableiten, Prüfen und Weiterentwickeln von Theorien eine bewährte und stetig wachsende Sammlung von Verfahren an, zusammengefasst unter der Bezeichnung *wissenschaftliche Methode*. Die wissenschaftliche Methode minimiert Fehlermöglichkeiten und nur durch ihre Anwendung kann unser Wissen über menschliches Erleben und Verhalten auf ein solides Fundament gestellt und systematisch erweitert werden.

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

Weiterführende Literatur

Bunge, M. & Ardila, R. (1990). *Philosophie der Psychologie.* Tübingen: Mohr.

In diesem Buch wird die wissenschaftliche Methode erläutert.

Gilovich, T., Griffin, D. & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment.* Cambridge: Cambridge University Press.

Überblick über die Forschung zu Urteils- und Denkfehlern.

Wissenschaftstheorie, Theorien und Hypothesen

2

ÜBERBLICK

» In diesem Kapitel verlassen wir zunächst Psychologie und Sozialwissenschaften im engeren Sinn. Das mag die Lektüre für viele Leser erschweren, nicht zuletzt wegen der zahlreichen neuen Fachbegriffe, die dabei notwendigerweise eingeführt werden. Trotzdem lohnt sich die Mühe, denn wir berühren zwei zentrale Aspekte aller Arten von Wissenschaft: Vorannahmen, die (noch) nicht direkt überprüfbar sind, und generelle Zugangsmöglichkeiten zu Wissen. Dabei begeben wir uns in die Philosophie, genauer gesagt, in ein Spezialgebiet der Philosophie – die Wissenschaftstheorie. Die zwei zentralen Fragen der Wissenschaftstheorie befassen sich damit, was die Wirklichkeit ist und wie wir Erkenntnisse darüber gewinnen können. Wie der Begriff Wissenschafts*theorie* schon andeutet, sind die Antworten auf diese Fragen letztlich auch nur Theorien (keine Wahrheiten), und man sollte sich der jeweiligen Annahmen bewusst sein, wenn man Forschung betreiben oder auch nur Forschungsergebnisse verstehen und verwenden möchte. Denn die Antworten auf die zwei zentralen Fragen der Wissenschaftstheorie sind ausschlaggebend dafür, welche Art von Forschungsergebnissen man erwarten und wie man sie gewinnen kann.

Wir widmen uns zunächst den Fragen nach dem Wesen der Wirklichkeit und den Möglichkeiten, sie zu erkennen. Sodann geben wir einen Überblick über verbreitete wissenschaftstheoretische Positionen und alternative Ansätze. Nach dem Ausflug in die Wissenschaftstheorie werden wir wieder zu Psychologie und Sozialwissenschaften zurückkehren und zwei spezielle Probleme erörtern, die dort eine besondere Rolle spielen: das Erfassen von Eigenschaften, auf die man keinen direkten Zugriff hat, und das besondere Verhältnis zwischen Forschern und „Forschungsobjekten“, die im Prinzip selbst auch Forscher sein könnten. Im Anschluss daran beschäftigen wir uns damit, was man über die Entstehung von Theorien weiß, und werden sehen, dass das relativ wenig ist. Weit mehr Erkenntnisse liegen darüber vor, wie man eine Theorie fundiert überprüfen kann. Wir beschreiben die grundlegende Vorgehensweise und stoßen dabei wieder auf die in *Kapitel 1* beschriebene wissenschaftliche Methode. Insbesondere werden wir illustrieren, wie man aus einer Theorie empirisch überprüfbare Hypothesen ableitet. «

2.1 Was ist die Wirklichkeit und wie können wir sie erkennen?

Gibt es eine von uns unabhängig existierende Welt, also Pflanzen, Tiere und Menschen, über die jeder im Prinzip dasselbe herausfinden könnte, wenn er oder sie wollte? Diese Frage mag auf den ersten Blick trivial klingen und die Antwort „Ja“ ist auch die Grundlage für viele Ansätze in der Wissenschaftstheorie. Bei genauerem Hinsehen ist die Frage aber nicht trivial. Wie können wir beispielsweise die Ansicht widerlegen, dass wir in einer Welt nur für uns leben und dass alle Sinneseindrücke selbst erzeugt sind und nur in der eigenen Vorstellung existieren? Könnte es sein, wir stellten uns vor, wir seien mit anderen Menschen zusammen und unterhielten uns mit ihnen, aber tatsächlich gäbe es niemanden, der mit uns redete? Oder dieser Jemand sähe tatsächlich ganz anders aus oder verhielte sich ganz anders, als wir das erleben? Selbst wenn wir uns in den Arm kneifen und Schmerz empfinden würden – ein beliebter Realitätstest –, könnte auch das ausschließlich in unserer Vorstellung stattfinden. Nicht so weit entfernt von einer solchen Position ist der Wahrnehmungspsychologe Donald Hoffman, der argumentiert, dass wir wohl nie wissen werden, wie die Welt tatsächlich aussieht; was wir wahrnehmen, ist nur die „Benutzerschnittstelle“ zwischen uns und der Welt (Kasten „Der virtuelle Tennisplatz“). Die Benutzerschnittstelle ist festgelegt durch die Möglichkeiten und Grenzen unseres Wahrnehmungssystems. Hoffman lässt allerdings die Frage offen, ob es eine unabhängig von uns existierende Welt gibt oder nicht.

Der virtuelle Tennisplatz

Donald Hoffman argumentiert, dass unsere Wahrnehmung der Welt so ähnlich funktioniert wie unsere naive Wahrnehmung der Funktionsweise eines Computers. Wenn wir an einem Computer arbeiten, manipulieren wir häufig Icons auf dem Bildschirm. Wir klicken beispielsweise mit der Maus auf ein Icon, das eine Datei bezeichnet und ziehen es über ein anderes Icon, das einem Papierkorb ähnlich ist. Wir verstehen diesen Vorgang als Löschen einer Datei, aber was tatsächlich im Computer abläuft, sind ganz andere Prozesse (u.a. wird die Datei nicht wirklich physikalisch gelöscht, sondern zunächst nur ein Verweis auf diese Datei). Hoffman (2008) illustriert seine These, dass wir nicht die Realität (sollte sie unabhängig von uns existieren) wahrnehmen, sondern nur eine „Benutzerschnittstelle“, mit Beispielen. Eines davon befasst sich mit einem virtuellen Tennisplatz.

Nehmen Sie an, Sie und ein Freund von Ihnen spielen in absehbarer Zukunft (wenn die entsprechende Technologie zur Verfügung steht) virtuelles Tennis in einer entsprechend ausgerüsteten Spielhalle. Sie tragen Datenhelme und benutzen Anzüge, die mit elektronischen Sensoren versehen sind. Sie finden sich wieder im Roland-Garros-Stadion, dem Schauplatz der French Open. Nachdem Sie den Platz und das Stadion bewundert haben, servieren Sie den ersten Aufschlag und sind bald in das Spiel vertieft. Sie nehmen das Stadion, das Spielfeld, das Netz, den Ball und den Schläger wahr, aber all diese Dinge sind nur Bestandteile einer ausgeklügelten Benutzerschnittstelle. Wenn Sie den Ball schlagen, mit dem Sie den ersten Satz gewinnen, dann scheint es Ihnen, wie wenn die Art und Weise Ihres Schlags den Ball dazu gebracht hat, gerade noch übers Netz zu gehen. Tatsächlich haben Sie gar keinen Ball geschlagen: Ball und Schläger sind nur Pixel in der Benutzerschnittstelle in Ihrem Datenhelm und senden auch keine Signale an den Supercomputer, der das ganze Spiel steuert. Der Schläger und der Ball dienen nur dazu, Ihre Aktionen zu steuern. Ihre Körperbewegungen wiederum werden über die Sensoren Ihres Anzugs an den Supercomputer weitergeleitet und lösen dort komplexe, aber Ihnen verborgene Prozesse aus. ▶

Das Computerprogramm bringt dann jeweils die Speicherinhalte, die mit den Positionen von Schlägern und Ball korrespondieren, auf den neuesten Stand. In diesem Beispiel besteht die Realität aus dem Supercomputer und den darin ablaufenden Prozessen. Wahrgenommen wird jedoch nicht diese Realität, sondern eine „Übersetzung" davon (eigentlich nicht existierende Dinge wie ein Tennisplatz, der Gegner, Schläger, Bälle usw.) und diese Übersetzung hängt vom jeweiligen Wahrnehmungsapparat, der „Benutzerschnittstelle" ab. Laut Hoffman, der sich ausgiebig mit Wahrnehmungstäuschungen beschäftigt hat (Hoffman, 1998), spricht vieles dafür, dass wir nie in der Lage sein werden, die Welt so wahrzunehmen, wie sie tatsächlich ist. Was wir seiner Meinung nach wahrnehmen, ist durch eine Benutzerschnittstelle (spezifisch für die Spezies „Mensch") festgelegt. Bislang hat Hoffmans These allerdings nur den Status eines interessanten Gedankenexperiments, das jedoch formal gut untermauert ist (Hoffman et al., 2015).

Der Bereich der Philosophie, der sich mit der Frage danach, was die Wirklichkeit ist, befasst, wird als *Ontologie* (Seinslehre) bezeichnet. Die *Epistemologie* (Erkenntnistheorie) beschäftigt sich hingegen mit der Frage, wie wir diese Wirklichkeit erkennen können. Zu beiden Bereichen sind unzählige Bücher geschrieben worden und wir werden auf die zwei Fragen *Was ist die Wirklichkeit?* und *Wie können wir sie erkennen?* in mehr oder weniger versteckter Form noch einige Male in diesem Kapitel stoßen. Zur Sensibilisierung der Leser für die Thematik sehen wir uns zunächst jeweils einen wichtigen Aspekt dieser zwei Fragen etwas genauer an, bevor wir uns konkreten wissenschaftstheoretischen Ansätzen widmen. Besonders relevant für die Psychologie ist die ontologische Frage nach dem Verhältnis von Leib und Seele und für alle Wissenschaften relevant ist die epistemologische Frage nach dem Verhältnis der zwei generellen Vorgehensweisen bei der Erkenntnis der Wirklichkeit: Induktion und Deduktion.

2.1.1 Das Leib-Seele-Problem

Wie ist das Verhältnis zwischen Leib und Seele, also zwischen physischen und mentalen Zuständen? Das ist eine Frage, die die Philosophen und Sozialwissenschaftler seit Anbeginn ihrer Wissenschaft beschäftigt und die natürlich auch für Psychologen äußerst relevant ist. Um zu optimistische Erwartungen gleich zu dämpfen: Bis jetzt gibt es keine allgemein akzeptierte Antwort. Sehen wir uns zunächst die zwei extremen Antworten an, bei denen das Leib-Seele-Problem streng genommen gar nicht existiert, weil jeweils nur die Existenz von Leib (Materie) oder Seele (Geist) angenommen wird:

1. Alles ist Materie: Menschliches Verhalten und Erleben ist also in seiner Gänze zurückführbar auf Gehirnzustände und Gehirnprozesse.
2. Alles ist Seele oder Geist: Was wir wahrnehmen und mental verarbeiten, ist nicht die letzte Wirklichkeit – wir schaffen uns die Welt in unserer Vorstellung.

Für die zweite Position gibt es in ihrer Reinform im Westen kaum Anhänger, sie spielt aber eine zentrale Rolle in einigen asiatischen, z.B. indischen, Ansätzen (Sedlmeier, 2006a). Etwas abgeschwächt taucht sie allerdings in sogenannten konstruktivistischen Ansätzen auf (*Abschnitt 2.2.2*). Die erste Antwort, die bedeutet, dass mentale Zustände ausschließlich auf physischen Zuständen beruhen, scheint auch unter Psy-

chologen relativ verbreitet zu sein. Diese Position legt nahe, dass sich die Psychologie noch mehr als bisher auf das Studium neurophysiologischer Prozesse konzentrieren müsste. Manche Psychologen gehen einfach pragmatisch so vor, dass sie – soweit das möglich ist – Bezüge zwischen gehirnphysiologischen Prozessen auf der einen Seite und Erleben und Verhalten auf der anderen herstellen. Solche dualistischen Leib-Seele-Positionen (dualistisch deswegen, weil man die Unterteilung in zwei Bestandteile vornimmt, eben Gehirn und Seele oder Geist) postulieren entweder, dass Materie und Geist miteinander *interagieren,* oder dass mentale und physische Zustände *parallel* – koordiniert oder unabhängig voneinander – existieren. Die interaktionistische Sichtweise ist im Alltag weit verbreitet und auch in der Wissenschaft anzutreffen. Wir nehmen beispielsweise an, dass wir Körperbewegungen durch unseren Geist steuern können, dass aber Körperbewegungen – etwa wenn wir aus Versehen jemanden anrempeln – wieder Auswirkungen auf mentale Zustände haben können (es könnte uns z.B. peinlich sein). Die Sichtweise, dass mentale und physische Prozesse parallel ablaufen, ist demgegenüber weniger verbreitet. Bei dieser Sichtweise liegt auch die Frage nahe, warum das jeweils andere – Geist oder Gehirn – überhaupt für wissenschaftliche Erklärungen notwendig ist, wenn es keine kausale Wirkung entfaltet. Da die Frage bislang nicht zufriedenstellend beantwortet werden kann, ist der Weg von der Position des Parallelismus hin zu den anfangs erörterten monistischen Positionen nicht weit (monistisch deswegen, weil man nur *einen* Wirkmechanismus, beruhend entweder auf Gehirn oder Geist, annimmt).

2.1.2 Induktion vs. Deduktion

Nehmen wir einmal an, die Wirklichkeit existiert tatsächlich unabhängig von uns nach bestimmten Gesetzmäßigkeiten. Wie können wir diese Gesetzmäßigkeiten erkennen? Wir können im Prinzip auf zwei Weisen vorgehen: Wir können beobachten, was in bestimmten Situationen passiert und dann versuchen, dafür eine Erklärung zu finden; oder wir könnten schon eine (vorläufige) Erklärung im Kopf haben und nachprüfen, ob unsere Beobachtungen zu dieser Erklärung passen. Im ersten Fall würden wir induktiv vorgehen und im zweiten deduktiv. Die wissenschaftliche Methode, die wir in *Kapitel 1* vorgestellt haben, ist hauptsächlich eine deduktive Methode: Man beginnt mit einer Theorie, leitet daraus Vorhersagen ab und überprüft, ob diese Vorhersagen zutreffen. In der Praxis ist eine rein deduktive Vorgehensweise allerdings äußerst selten. Schon im Verlauf einer Studie kann sich herausstellen, dass man bei der Ableitung der Fragestellung bedeutende Einzelheiten nicht in Betracht gezogen hat und dass sie deswegen modifiziert werden muss. Noch öfter wird allerdings nach der Durchführung der Studie deutlich, dass die zuvor aufgestellte Hypothese so nicht stimmen kann. Wie kommt man zu einer Modifikation oder Neufassung der Hypothese? Das ist ein induktiver Akt, der sich z.B. auf Aussagen der Versuchsteilnehmer stützt oder auf Ideen, die der Forscher beim Analysieren der Daten hat. Im normalen Forschungsprozess spielt also immer beides eine Rolle: Induktion *und* Deduktion.

In der Wissenschaftstheorie hat die deduktive Vorgehensweise allerdings einen deutlich höheren Stellenwert. Das liegt zum Teil daran, dass sie – mithilfe verschiedener Verfahren der Logik – sehr gut formalisiert werden kann. Es liegt aber auch daran, dass Ursache-Wirkungs- und Wenn-Dann-Beziehungen – die zentralen Komponenten beim Erklären von Erleben und Verhalten – deduktiv sehr viel besser untersucht werden können. Die Logiker benutzen zur Illustration der Schwäche der Induktion in Bezug

auf solche Beziehungen häufig das „Schwäne-Beispiel“: Selbst wenn ich tausend weiße Schwäne gesehen habe, kann ich nicht schlussfolgern, dass alle Schwäne weiß sind. Es kann ja immer sein, dass der nächste Schwan, den ich beobachten werde, schwarz sein wird. Dann stimmt die Beziehung „wenn Schwan, dann weiß“ nicht mehr.

Der relative Anteil von Induktion und Deduktion verändert sich charakteristischerweise im Laufe des Forschungsprozesses: Am Anfang einer Forschungsrichtung steht immer die Induktion (*Abschnitt 2.4*): Erst muss eine Idee oder Fragestellung vorhanden sein, bevor sie (deduktiv) überprüft werden kann. Induktion fängt jedoch meist nicht bei null an. In der Psychologie gibt es viele Theorien und Hypothesen, bei denen der induktive Prozess darin bestand, Analogien zwischen mathematischen oder technischen Modellen, die schon vorhanden waren, und Aspekten menschlichen Verhaltens herzustellen. Je besser ausgearbeitet eine Theorie ist und je genauer die daraus abgeleiteten Vorhersagen sind, desto gewichtiger wird der deduktive Anteil im Forschungsprozess sein.

2.2 Wissenschaftstheoretische Ansätze im Überblick

Bislang gibt es keine allgemein anerkannte wissenschaftstheoretische Position für Psychologie und Sozialwissenschaften und vermutlich wird es eine solche auch in absehbarer Zeit nicht geben. Trotzdem werden Forscher in diesen Disziplinen, wenn man sie zu ihrer Meinung über Wissenschaftstheorie befragt, einige Begriffe und Namen häufiger nennen als andere. In den folgenden Abschnitten werden wir zunächst auf solche „konventionellen Ansätze“ eingehen. Alternative Ansätze, manchmal als „konstruktivistische Ansätze“ bezeichnet, führen bislang eher ein Schattendasein. Wir geben nur einen kurzen Einblick in solche Alternativen im *Abschnitt 2.2.5*, kommen jedoch in *Kapitel 32* noch einmal darauf zu sprechen.

Die hier diskutierten konventionellen Ansätze gehen alle davon aus, dass es eine unabhängig von unseren Vorstellungen existierende Welt gibt. Sie unterscheiden sich jedoch in ihren Schwerpunkten. Manche Ansätze stellen Regeln dafür auf, wie Theorien aussehen sollen oder wie man sie überprüfen sollte, andere legen den Schwerpunkt auf die Analyse der tatsächlich in den Wissenschaften ablaufenden Prozesse und Mechanismen (▶ Abbildung 2.1). Am prominentesten unter Psychologen und Sozialwissenschaftlern ist wohl der von Karl Popper entwickelte *Kritische Rationalismus*, auf den wir daher auch besonderes Gewicht legen wollen. Die *Methodologie wissenschaftlicher Forschungsprogramme*, die auf Imre Lakatos zurückgeht, lässt sich als eine Weiterentwicklung dieses Ansatzes verstehen. Aber auch der u.a. von Rudolf Carnap entwickelte *Logische Empirismus* und die *historisch-soziologische Analyse* von Thomas Kuhn spielen eine gewisse Rolle in der psychologischen und sozialwissenschaftlichen Forschung. Die folgenden Ausführungen zu den Ansätzen in ▶ Abbildung 2.1 sind vereinfacht und sollen nur einen ersten Überblick geben (siehe Westermann, 1987, für eine ausführlichere Darstellung der in diesem Absatz diskutierten Ansätze).

Wir werden uns zunächst den beiden *aprioristischen* Positionen, dem Logischen Empirismus und dem Kritischen Rationalismus, zuwenden. Etwas vereinfacht gesagt ist der Logische Empirismus eine Theorie darüber, wie Theorien aussehen sollen, und der Kritische Rationalismus eine Theorie darüber, wie man Theorien überprüfen sollte. Beide machen Vorschriften und Empfehlungen, deren Gültigkeit von vornherein *(a priori)* und unabhängig von den Einzelwissenschaften angenommen wird (oben in ▶ Abbildung 2.1). Die beiden anderen in diesem Abschnitt diskutierten Ansätze, Kuhns historisch-soziologische Analyse und Lakatos' Methodologie wissenschaftlicher Forschungsprogramme, beziehen demgegenüber Erkenntnisse und methodische Ansätze aus den Einzelwissenschaften durchaus mit ein, sie arbeiten also *quasi-empirisch* (unten in ▶ Abbildung 2.1).

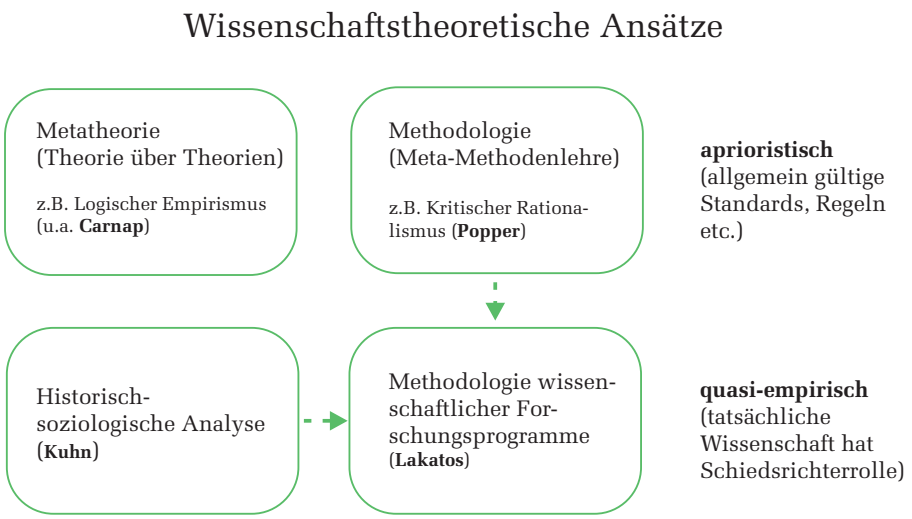


Abbildung 2.1: Überblick über die in diesem Kapitel diskutierten konventionellen wissenschaftstheoretischen Ansätze.

2.2.1 Logischer Empirismus

Der Logische Empirismus, auch *Logischer Positivismus* genannt, entstand aus einem 1923 gegründeten Diskussionszirkel in Wien heraus („Wiener Kreis"), unter anderem als Gegenbewegung zur Psychoanalyse. Seine Vertreter verlangten, dass alle bedeutungsvollen Aussagen der Wissenschaft auf Beobachtungen (daher der „Empirismus") zurückführbar sein müssen, was etwa bei der Psychoanalyse nicht der Fall ist. Beobachtung oder Erfahrung ist also für den logischen Empirismus die Grundlage von Wissenschaft. Zu seinen Begründern zählen Rudolf Carnap, Hans Reichenbach und Herbert Feigl. Dem Logischen Empirismus zufolge sollen alle Theorien in einer formalen Sprache wie der Aussagen- oder Prädikatenlogik (daher „logischer" Empirismus) ausgedrückt oder „axiomatisiert" werden können. Durch die „Axiomatisierung" sollen die Aussagen einer Theorie die Ambiguität der Alltagssprache vermeiden. Die Sprache des logischen Empirismus enthält neben logischen Termen (wie z.B. „und", „oder" und „nicht") zwei Arten von Begriffen: theoretische Begriffe und Beobachtungsbegriffe. Dabei sollen die theoretischen Begriffe eindeutig und vollständig auf Beobachtungsbegriffe zurückführbar sein. Diese drei Anforderungen an eine Theorie

sind auch als „Standardkonzeption wissenschaftlicher Theorien“ bekannt (▶ Abbildung 2.2). Alle Theorien sollen also logisch auf Erfahrung zurückführbar oder, mit anderen Worten, empirisch verifizierbar (vollständig bestätigbar) sein. Die Erfahrungen oder Beobachtungen werden als *Protokollsätze* (B_1, B_2, … B_n in ▶ Abbildung 2.2) bezeichnet. Wenn sich nun eine theoretische Aussage (H in ▶ Abbildung 2.2) aus mehreren Protokollsätzen vollständig ableiten lässt, gilt sie als empirisch verifiziert.

Standardkonzeption wissenschaftlicher Theorien

1. Formale Axiomatisierung
2. Beobachtungs- und theoretische Begriffe
3. Theoretische Begriffe (H) auf Beobachtungsbegriffe (B) zurückführbar

$$(B_1 \wedge B_2 \ldots \wedge B_n) \rightarrow H \qquad (\wedge: \text{und}, \rightarrow: \text{impliziert})$$

Abbildung 2.2: Standardkonzeption wissenschaftlicher Theorien entsprechend dem Logischen Empirismus.

Der Logische Empirismus geht also im Prinzip induktiv vor, von den Beobachtungen zur Theorie. Das Verifizieren einer Theorie mithilfe der Induktion ist aber, wie wir schon am Beispiel mit den weißen und schwarzen Schwänen gesehen haben, problematisch. Induktive Schlüsse über eine Population (alle Mitglieder einer definierten Gruppe) sind nur dann ohne Probleme zu ziehen, wenn wir die ganze Population überprüfen können oder wenn unsere Hypothese besagt, dass eine Gesetzmäßigkeit nur bei mindestens einem Objekt oder einer Person zutreffen muss (z.B. mindestens ein Schwan ist weiß). Beides ist in der Wissenschaft äußerst selten. Dieses logische Problem, dass man mithilfe der Induktion keine Theorie verifizieren oder „beweisen“ kann, wurde früh als Kritik am Logischen Empirismus vorgebracht. Allerdings hat auch Carnap in späteren Schriften (z.B. Carnap, 1946) anerkannt, dass eine rein induktive Vorgehensweise logisch nicht haltbar ist. Weiterhin wird am logischen Empirismus kritisiert, dass eine durchgehende Axiomatisierung von Theorien praktisch nicht möglich ist. Ein dritter Kritikpunkt bezieht sich schließlich darauf, dass eine theoriefreie Beobachtung sehr schwierig bis unmöglich ist (siehe *Kapitel 1*). Letzteres führt dazu, dass die Beobachtungsbegriffe nicht mehr eindeutig anwendbar sind, weil sie möglicherweise durch die subjektiven Theorien des Beobachters verfälscht sind. Der logische Empirismus spielt in seiner Reinform heutzutage keine bedeutsame Rolle mehr. Er hatte allerdings großen Einfluss auf spätere Ansätze. Außerdem ist die Idee, Theorien in einer möglichst eindeutigen (logischen) Sprache auszudrücken, auch heute noch aktuell und wurde in neueren wissenschaftstheoretischen Ansätzen wieder aufgegriffen.

2.2.2 Kritischer Rationalismus

Wenn Psychologen oder Sozialwissenschaftlern nur ein einziger Name zu Wissenschaftstheoretikern einfällt, dann ist das mit hoher Wahrscheinlichkeit der von (Sir) Karl Popper, dem Begründer des Kritischen Rationalismus. Popper (z.B. Popper, 1989; Erstveröffentlichung 1934) lehnte die zentralen Ideen des Logischen Empirismus vollständig ab. Insbesondere kritisierte er die zwei Annahmen, die notwendig sind, um argumentieren zu können, dass Theorien verifizierbar sind: a) Es gibt sichere theorieunabhängige Beobachtungen und b) Induktionsschlüsse sind gerechtfertigt. Aus seiner Sicht boten weder Beobachtung noch Induktion eine Grundlage für verifizierbares, sicheres Wissen.

Fallibilismus

Betrachten wir zunächst noch einmal kurz das Problem der Induktion. Es ist leicht einzusehen, dass auch 1000 weiße Schwäne nicht beweisen, dass der 1001. Schwan ebenfalls weiß sein wird. Dennoch teilen Sie vermutlich die Intuition, dass die vielen positiven Beispiele (die weißen Schwäne) die Plausibilität der Verallgemeinerung („alle Schwäne sind weiß“) stärken. Wäre es also zumindest gerechtfertigt anzunehmen, dass ein weiterer weißer Schwan die *Wahrscheinlichkeit* dafür erhöht, dass alle Schwäne weiß sind? Auch diese Auffassung war schon lange vor Popper unter massive Kritik geraten. Der schottische Philosoph David Hume[1] hatte bereits im 18. Jahrhundert argumentiert, dass es grundsätzlich nicht rational begründbar ist, allein aus wiederholten Erfahrungen irgendeine Schlussfolgerung über weitere Beispiele (den nächsten Schwan) oder künftige Ereignisse abzuleiten. Auch sehr viele Beobachtungen, die eine bestimmte Verallgemeinerung bestätigen, schließen nicht aus, dass wir künftig teilweise oder auch ausnahmslos Beobachtungen machen werden, die dieser Verallgemeinerung entgegenstehen. Dass weitere bestätigende Beobachtungen auch nicht zu einer erhöhten Wahrscheinlichkeit der Verallgemeinerung führen, lässt sich leicht an einigen Beispielen zeigen, mit denen wir tatsächlich sehr viel Erfahrung gesammelt haben: Wir (und sicher auch ein großer Teil der Leser dieses Buchs) waren an Tausenden von Tagen in ununterbrochener Folge in der Lage, uns morgens aus unseren Betten zu erheben. Aus diesem Umstand werden Sie kaum folgern wollen, dass wir auch künftig an allen Tagen in der Lage sein werden, aufzustehen. Auch erhöhen zusätzliche „erfolgreiche“ Tage nicht die Wahrscheinlichkeit, dass noch ein weiterer erfolgreicher Tag hinzukommen wird – mit wachsendem Alter wird die Wahrscheinlichkeit für morgendliches Aufstehen eher sinken. Die letzte Zuflucht der Induktion könnte in dem Argument liegen, dass sie sich dennoch bewährt hat: Sie wurde und wird genutzt, um Schlüsse zu ziehen, die sich dann als nützlich und erfolgreich erweisen. Auch dieses Argument hat bereits Hume entkräftet. Das Problem ist hier, dass das Argument selbst bereits die Gültigkeit der Induktion voraussetzt. Auch wenn Induktion in vielen Beispielen erfolgreich sein mag, folgt daraus nicht, dass sie das auch in weiteren Beispielen sein wird – es sei denn, man akzeptiert die Gültigkeit der Induktion. Hume schließt, dass Menschen aus reiner Gewohnheit aus

1 Hume ist ein Vertreter des *Empirismus*, eines philosophischen Ansatzes, der in (Sinnes-)Erfahrungen die wesentliche oder alleinige Grundlage für Wissen und Erkenntnis sieht. Er gilt als der erste Philosoph, der sich mit dem *Induktionsproblem* befasst hat, das daher auch als „*Hume-Problem*“ bezeichnet wird.

ihren Erfahrungen lernen und Schlussfolgerungen ableiten (wie etwa, dass auch der nächste Schwan weiß sein wird oder – in einem anderen bekannten Beispiel – dass Brot sie weiterhin sättigen wird) und dass dies praktisch auch durchaus nützlich ist. Aber diesen Schlussfolgerungen fehlt laut Hume jede rationale Begründung. Induktion ist generell nicht gerechtfertigt.

Popper (1989) macht sich nicht nur diese Kritik an der Induktion zu eigen, er greift darüber hinaus auch die Idee des logischen Empirismus an, dass es sichere theoriefreie Beobachtungen geben könne. Um seine Position zu illustrieren, nutzt er den vermeintlich sehr einfachen Beobachtungssatz „Hier steht ein Glas Wasser" (S. 61). Dieser Satz enthält die Allgemeinbegriffe „Glas" und „Wasser" (sogenannte *Universalien*). Diese Begriffe bezeichnen offensichtlich keine unmittelbaren Erfahrungen. Wir benötigen zusätzliches Wissen, Erwartungen, also eigene Theorien, um diese Begriffe anwenden zu können. Popper argumentiert nun, dass jeder Beobachtungssatz Begriffe enthält, die über die reine Wahrnehmung hinausgehen (er ersetzt den Ausdruck Beobachtungssatz daher auch durch den Ausdruck *Basissatz*). Damit kann aber auch die einfache Aussage „Hier ist ein Glas Wasser" nicht durch unmittelbare Erfahrungen verifiziert werden. Auch Basissätze sind theorieabhängig und fehlbar.

Poppers Sicht auf die Grundlagen des logischen Empirismus verdeutlicht ein erstes wesentliches Merkmal seiner wissenschaftstheoretischen Position: Der Kritische Rationalismus ist *fallibilistisch*. Sicheres Wissen ist demnach generell unmöglich. Weder bietet Beobachtung eine Basis für sicheres Wissen noch ist Induktion geeignet Theorien zu beweisen. Alle unsere Theorien bleiben also stets Vermutungen, die möglicherweise falsch sind. Dass dies nicht nur eine philosophische Spitzfindigkeit ist, zeigt ein Blick in die Wissenschaftsgeschichte: Zu Poppers Lebzeiten wurde die Newtonsche Physik – vermutlich die bis dahin erfolgreichste Theorie überhaupt – durch die Relativitätstheorie und die Quantentheorie abgelöst. Die Newtonschen Gesetze waren über mehr als zwei Jahrhunderte immer wieder durch zahllose Beobachtungen „bestätigt" worden. Dennoch betrachtet die moderne Physik sie als falsch.

Falsifikationismus

Wenn wir grundsätzlich kein sicheres Wissen erlangen können, wie unterscheidet sich Wissenschaft dann noch von „Nicht-Wissenschaft" oder Metaphysik? Wenn Beobachtung und Induktion keine Grundlage bieten, um Theorie zu verifizieren, welche Funktion hat Empirie dann überhaupt? Wie kann Empirie dann noch etwas über Theorien aussagen? Poppers (1989) Antworten auf diese Fragen beruhen auf der Idee der Falsifikation: Wir können zwar nicht induktiv beweisen, dass unsere Theorien wahr sind, aber auf dem Wege der Deduktion ergibt sich eine logisch korrekte Möglichkeit, Beobachtungen zu nutzen, um zu zeigen, dass unsere Theorien *falsch* sind. ▶ Abbildung 2.3 zeigt die logische Grundlage für die Falsifizierbarkeit von Aussagen. Wenn eine Hypothese (*H*) vorhersagt, dass ein bestimmtes Ereignis beobachtet werden kann (*B*), dann lässt sich aus dem Nicht-Eintreten des Ereignisses schlussfolgern, dass die Hypothese nicht stimmt – eine in der Logik als *Modus Tollens* bekannte Schlussform. Ein Beispiel: Die Hypothese (*H*) sei, dass der Siedepunkt von Wasser bei 50 Grad Celsius liegt (entspricht der ersten Zeile in ▶ Abbildung 2.3, *B* ist das Ereignis, dass das Wasser bei 50 Grad Celsius siedet). Das Wasser wird nun auf 50 Grad erhitzt und die Beobachtung ergibt, dass das Wasser nicht siedet (zweite Zeile in ▶ Abbildung 2.3). Somit gilt die Hypothese als falsifiziert (dritte Zeile in ▶ Abbildung 2.3).

$$\begin{array}{l} H \rightarrow B \\ \neg\, B \\ \hline \neg\, H \end{array}$$

(*H*: Hypothese, *B*: Beobachtung)

Abbildung 2.3: Die logische Grundlage der Falsifizierbarkeit von Aussagen: der *Modus Tollens*. Die Zeichen in der Abbildung haben folgende Bedeutungen: → = logische *Implikation* und ¬ = *Negation* (= trifft nicht zu).

Es gibt also eine grundsätzliche Asymmetrie in der Möglichkeit, Theorien durch Beobachtungen zu falsifizieren oder zu verifizieren. Auf dieser Asymmetrie beruht Poppers gesamter wissenschaftstheoretischer Ansatz und sie begründet das zweite wesentliche Merkmal des Kritischen Rationalismus: Er ist *falsifikationistisch*. Wissenschaft besteht demnach darin, prüfbare Theorien vorzuschlagen und sich anschließend darum zu bemühen, diese Theorien zu falsifizieren. Eine Theorie sollte also möglichst strengen Tests unterzogen werden, in denen sie sich als falsch erweisen kann. Tut sie dies tatsächlich, so ist das ein Anlass, die bisherige Theorie zu verwerfen und nach Möglichkeit durch eine bessere zu ersetzen. Diese Theorie sollte mit den falsifizierenden Beobachtungen der alten Theorie vereinbar sein – und sie muss wiederum neuen Falsifikationsversuchen ausgesetzt werden. Übersteht sie diese, so ist sie natürlich nach wie vor nicht bewiesen. In der Terminologie Poppers gilt sie damit als *„bewährt"*. Der Bewährungsgrad einer Theorie steigt mit der Zahl und der Strenge der Falsifikationsversuche, denen sie widerstanden hat (auf den Begriff der Strenge der Prüfung werden wir unten noch einmal zurückkommen).

Die Idee des Falsifikationismus mag auf den ersten Blick wenig intuitiv erscheinen. Wir sind sicher oftmals stärker motiviert, die Richtigkeit unserer Annahmen und Theorien aufzuzeigen, als uns darum zu bemühen, ihre Fehlerhaftigkeit nachzuweisen. Gerade darin liegt aber vielleicht eine wesentliche Nachricht des Falsifikationismus: Er ändert unsere „Suchrichtung". Eine in der Psychologiegeschichte sehr alte Hypothese besagt, dass Frustration Aggression erzeugt. Sicher lassen sich zahllose Beobachtungen finden, die mit dieser Hypothese in Einklang stehen – auf Frustration folgt oftmals Aggression. Ein Anhänger der Hypothese mag uns nun eine weitere entsprechende Beobachtung zeigen. Aber was nutzt uns dieses zusätzliche positive Beispiel? Weder beweist es die Richtigkeit der Hypothese noch erhöht es deren Wahrscheinlichkeit. Es ist erkenntnistheoretisch belanglos. Unsere Erfahrung liefert uns in genau dem Moment eine bedeutsame Rückmeldung aus der Umwelt, in dem wir eine Beobachtung machen, die *nicht* mit der Hypothese übereinstimmt. Nun ergibt sich die Gelegenheit etwas zu lernen – vielleicht über Umstände, unter denen Frustration nicht zu Aggression führt. Wir können neue, verbesserte Hypothesen bilden und diese wiederum testen. Wir sollten also nach falsifizierenden Beobachtungen suchen, um dazulernen zu können. (Tatsächlich hat sich die Hypothese „Frustration erzeugt Aggression" in dieser sehr einfachen Form als falsch erwiesen, Frustration erzeugt nicht immer Aggression. Dies hat wiederum zu einer komplexeren Theoriebildung geführt, die beispielsweise zu erklären sucht, unter welchen Bedingungen Frustration tatsächlich Aggression nach sich zieht; siehe z.B. Berkowitz, 1974).

Das Abgrenzungsproblem

Popper (1989) nutzt das Prinzip der Falsifikation auch, um zu bestimmen, wie sich Wissenschaft von Nicht-Wissenschaft unterscheidet. Wissenschaft kann nur falsifikationistisch vorgehen, wenn sie sich mit Aussagen befasst, die im Prinzip widerlegbar sind: Eine Theorie muss also bestimmte Beobachtungen ausschließen, um falsifizierbar zu sein und Gegenstand einer empirisch-wissenschaftlichen Betrachtung werden zu können. Falsifizierbarkeit ist daher Poppers *„Abgrenzungskriterium"* der (empirischen) Wissenschaft von der Nicht-Wissenschaft oder Metaphysik. Nicht-Wissenschaft behandelt Aussagen, die nicht falsifizierbar sind. Einfache Beispiele für nicht falsifizierbare Aussagen liefert vielleicht die Astrologie: Wenn Sie für geraume Zeit das Horoskop des Astrologen Ihres Vertrauens beziehen, werden Sie vermutlich feststellen, dass sich darin kaum Aussagen finden, für die sich eindeutig angeben ließe, welche Beobachtungen sie falsifizieren könnten. Astrologie ist daher keine Wissenschaft. Generell nicht falsifizierbar sind auch sogenannte Tautologien (also Aussagen, die unabhängig davon wahr sind, ob die in ihnen enthaltenen Teilaussagen zutreffen). Ein einfaches Beispiel wäre hier der Satz „Kräht der Gockel auf dem Mist, ändert sich's Wetter oder es bleibt wie es ist". Eine Falsifizierung ist auch nicht möglich bei Aussagen, die zu unpräzise sind. Der Satz „Meditation *kann* die Konzentrationsfähigkeit erhöhen" schließt keine denkbare Beobachtung aus – er ist also nicht falsifizierbar. Auch der Satz „Alle Menschen sind sterblich" ist nicht der Wissenschaft, sondern der Metaphysik zuzurechnen. Auch hier gibt es keine Beobachtung, die diesen Satz widerlegen könnte (auch bei einem 400-Jährigen schließt nichts aus, dass er noch sterben wird).

Die Bedeutung der Kritik

Wir können falsifizierende Beobachtungen als eine Möglichkeit verstehen, Theorien zu kritisieren und die Konstruktion besserer Theorien anzuregen. Beobachtungen sind aber nicht die einzige Möglichkeit zur Kritik an theoretischen Aussagen. Popper (1994, S. 213 f) gibt ein Beispiel aus der vorsokratischen Philosophie: Thales von Milet lehrte um 600 v. Chr. ein Weltbild, demzufolge „die Erde von Wasser getragen wird, auf dem sie reitet, ähnlich wie ein Schiff". Sein Schüler Anaximander kritisierte diese Theorie und kam zu einer völlig anderen Auffassung. Demnach wird „die Erde … von nichts gehalten, aber sie bleibt dadurch an ihrem Ort, dass sie von allen Dingen den gleichen Abstand hat. Ihre Form ist die einer Trommel … Auf einer ihrer flachen Seiten gehen wir, und die andere Seite liegt gegenüber." Dies erscheint in vielerlei Hinsicht immer noch recht naiv, dennoch war es Anaximander offensichtlich zumindest in einem Aspekt gelungen, sich modernen Vorstellungen anzunähern. Popper argumentiert nun, dass die Grundlage für diese Verbesserung keine Beobachtung gewesen sein kann. Mit den technischen Möglichkeiten der Antike führt Empirie bei diesem Problem eher in die Irre – die Erde erscheint flach und alles scheint von irgendetwas anderem gehalten zu werden. Anaximanders Kritik beruhte eher auf einem logischen Problem: Wenn die Erde auf Wasser schwimmt, muss irgendetwas dieses Wasser halten. Wenn dies eine riesige, halb gefüllte Kugel ist (eine Vorstellung, die ebenfalls auf Thales zurückgeführt wird), muss wiederum irgendetwas diese Kugel halten. Offensichtlich führt uns dies in einen „infiniten Regress". Wir benötigen eine Theorie, die uns aus diesem Problem befreit.

Theorien können und sollten also auch mit nicht-empirischen Mitteln kritisiert werden. Sie sollten beispielsweise im Hinblick auf ihre logischen Eigenschaften und ihre Widerspruchsfreiheit untersucht werden. Poppers Blick in die vorsokratische Philosophie illustriert zudem, dass er nicht-falsifizierbare, metaphysische Aussagen keineswegs per se für sinnlos hält. Auch metaphysische Aussagen können Gegenstand einer rationalen Kritik werden, die vielleicht dazu führt, dass sie aufgegeben und durch verbesserte Aussagen ersetzt werden.[2] Im Bereich der Metaphysik kann es also ebenfalls „Wissenszuwachs" geben (allerdings gilt natürlich auch hier das Prinzip des Fallibilismus: Auch dieses Wissen bleibt vorläufig und kann sich möglicherweise noch als falsch erweisen). Popper (1994) betrachtet metaphysische Überlegungen zudem als einen möglichen Ausgangspunkt für die Entstehung wissenschaftlicher Theorien. Metaphysik unterscheidet sich von Wissenschaft lediglich dadurch, dass ihr die Möglichkeit der Falsifikation durch Empirie fehlt. Wo immer dies möglich ist, sollten allerdings falsifizierbare Aussagen getroffen werden – eben, um sich dieser Möglichkeit der Kritik nicht zu entziehen.

Für Popper ist die Aufeinanderfolge der Theorien von Thales und Anaximander schließlich noch aus einem anderen Grund erwähnenswert: Offensichtlich hat Anaximander seinen Lehrer Thales überhaupt kritisiert und eine eigene Theorie entwickelt. Dies ist alles andere als selbstverständlich. Bis dahin (und natürlich auch in der weiteren Weltgeschichte immer wieder) haben „Schulen" vor allem Lehrmeinungen verbreitet. Diese Lehrmeinungen wurden weder kritisiert noch verändert. Im Gegenteil, die Funktion einer Schule bestand eher darin, die Lehrmeinung in möglichst „reiner" Form zu erhalten. Die Schule des Thales scheint nun die erste Schule gewesen zu sein, in der etwas grundsätzlich anderes passiert ist. Popper (1994, S. 234) beschreibt, dass er sich Thales als den ersten Lehrer vorstellt, der zu seinen Schülern gesagt hat „So sehe ich die Dinge – das ist meine Theorie. Versucht nun, meine Lehren zu verbessern". Thales hätte demnach also Kritik ermutigt und könnte als Begründer der „kritischen Tradition" angesehen werden. Kritische Diskussion ist laut Popper der einzige Weg, auf dem wir unser Wissen mehren, also zu verbesserten Vermutungen über die Welt kommen können. Sicheres Wissen ist generell unmöglich. Es gibt also keine „guten Gründe", etwas für wahr zu halten. Es ist allerdings durchaus möglich aufzuzeigen, dass eine Theorie falsch ist. Wir können neue Erkenntnisse gewinnen, indem wir mit einer Vermutung beginnen und anschließend versuchen, an dieser Vermutung Kritik zu üben und Fehler aufzuzeigen. Ist diese kritische Diskussion erfolgreich, so kann eine neue und verbesserte Vermutung aufgestellt werden, die der bisherigen Kritik standhält. Im Sinne der kritischen Tradition sollten wir nun Wege suchen, Mängel dieser neuen Vermutung aufzudecken. – Popper argumentiert, dass mit der „Begründung" der kritischen Tradition durch Thales in der vorsokratischen Philosophie ein immenser Wissenszuwachs in nur sehr kurzer Zeit einherging. Dieser Vorgang hat sich in der Geschichte wiederholt. Die kritische Tradition ging noch in der Antike wieder verloren. Sie wurde in der Renaissance wiederentdeckt. Wieder kommt es zu einem enormen Erkenntnisgewinn in sehr kurzer Zeit.

2 Thales und Anaximanders Theorien waren zu ihren Lebzeiten nicht falsifizierbar, sie sind es aber im Prinzip (wir können heute feststellen, dass die Erde nicht auf Wasser schwimmt wie ein Schiff). Sie sind damit nicht der Metaphysik zuzurechnen. Die Gemeinsamkeit besteht hier lediglich darin, dass die Mittel, mit denen diese Theorien in der Antike kritisiert wurden, auch auf metaphysische Aussagen anwendbar sein können.

Wissenschaftler sollten also Kritik üben. Sie sollten aber auch bereit sein, ihre eigenen Auffassungen kritisierbar zu machen, Kritik zu suchen und – nicht zuletzt – Kritik auszuhalten. Wissenschaftliche Institutionen (etwa Ihre Universität oder wissenschaftliche Zeitschriften) sollten Kritik fördern. Um Teil der kritischen Tradition zu werden, sollten Sie sich beim Lesen dieser Zeilen fragen, warum die darin enthaltenen Argumente falsch sein könnten, warum wir (die Autoren) möglicherweise einfach irren. Nur dies böte Ihnen und uns die Möglichkeit zu (noch) besseren Einsichten zu gelangen als denen, die wir Ihnen im Moment anbieten können.[3] Ein Mangel an Kritik kann Wissenschaft dagegen in Pseudowissenschaft verwandeln (s. den Kasten „Pseudowissenschaft").

Pseudowissenschaft

Popper (1994) hat neben den Begriffen Wissenschaft und Metaphysik auch den Begriff „Pseudowissenschaft" verwendet. Er kennzeichnet damit Praktiken, die den Anspruch erheben, wissenschaftlich zu sein, dies aber tatsächlich nicht sind. Seine zentralen Beispiele für solche Pseudowissenschaften sind Marxismus und Psychoanalyse. Der Psychoanalyse wirft er – ganz im Sinne des Abgrenzungskriteriums – vor, nicht falsifizierbar zu sein. Die Theorie könne jedes beliebige Verhalten erklären, etwa das eines Mannes, der ein Kind ins Wasser wirft, um es zu töten, ebenso gut wie das eines zweiten Mannes, der bei einem Sprung ins Wasser sein Leben riskiert, um das Kind zu retten (ersteres vielleicht als Ausdruck eines *verdrängten Ödipus-Komplexes*, letzteres als Folge einer *Sublimierung*). Interessanterweise billigt Popper dem Marxismus aber durchaus zu, zunächst falsifizierbare Vorhersagen getroffen zu haben. Beispielsweise die, dass die erste sozialistische Revolution in einem wirtschaftlich wie technisch hoch entwickelten Land stattfinden sollte. Tatsächlich ereignete sich die erste Revolution in Russland – einem Land, das im Jahr 1917 nicht sonderlich entwickelt war. Das Problem, das aus dem Marxismus eine Pseudowissenschaft macht, setzt nun dort ein, wo derartige Beobachtungen nicht genutzt werden, um die Theorie kritisch zu hinterfragen, sie zumindest teilweise zu modifizieren oder aufzugeben. Tatsächlich wurde der Marxismus laut Popper in einer Weise „angepasst", die ihn mit jedem beliebigen Ereignis vereinbar macht. Verifikation (im Sinne einer bestätigenden Beobachtung) ist nun sehr billig zu haben, wenn eine Theorie alles erklären kann. Die vermeintliche „Erklärungskraft" einer solchen Theorie ist eben keine besondere Stärke, sondern eine Schwäche. Entsprechend wirft Popper (1994, S. 48) den Vertretern und Anhängern von Psychoanalyse und Marxismus vor allem vor, dass sie imstande waren, jedes erdenkliche Ereignis als eine Verifikation ihrer Theorien zu interpretieren.

Das Beispiel des Marxismus zeigt aber auch, dass Wissenschaft nicht schon allein dadurch Wissenszuwachs erzeugt, dass sie falsifizierbare Theorien aufstellt. Die wissenschaftliche Gemeinschaft benötigt zudem eine „falsifikationistische Haltung". Sie muss aktiv nach Falsifikationen ihrer aktuellen Theorien suchen. Und sie muss bereit sein, diese Theorien sinnvoll weiterzuentwickeln oder aufzugeben, wenn falsifizierende Beobachtungen eingetreten sind. Andernfalls betreibt sie laut Popper lediglich Pseudowissenschaft.

3 Allerdings braucht Lernen zunächst einen gewissen „Vertrauensvorschuss" gegenüber den Lehrenden. Sie lernen bei der Lektüre dieses Buchs wahrscheinlich leichter, wenn Sie uns für den Anfang unterstellen, dass wir etwas Sinnvolles zu sagen haben, und wenn Sie zunächst versuchen, diesen Sinn nachzuvollziehen. Dennoch wird ein Satz natürlich auch nicht dadurch wahr, dass er in diesem Buch steht. Der Moment, in dem Sie sich fragen sollten, ob und warum unsere Argumente fehlerhaft sein könnten, wird also kommen.

Empirischer Gehalt und Theoriewahl

Wir haben schon angedeutet, dass sich Theorien hinsichtlich ihrer „Güte“ nicht nur dadurch unterscheiden, ob sie bereits falsifiziert wurden oder noch auf ihre Falsifikation warten. Ein weiteres wichtiges Kriterium, anhand dessen Theorien beurteilt werden können, ist ihr *empirischer Gehalt*. Der empirische Gehalt einer Theorie ist umso größer je mehr mögliche Beobachtungen sie verbietet. Wir können uns die aus einer Theorie abgeleiteten Vorhersagen als Wenn-dann-Sätze vorstellen. Den Wenn-Teil nennt Popper (1989) Allgemeinheit, den Dann-Teil Bestimmtheit. Die Allgemeinheit bezeichnet die Menge der Fälle, auf die sich die Vorhersage bezieht, die Bestimmtheit entspricht der Genauigkeit oder Präzision dieser Vorhersage. Popper vergleicht in einem (hier leicht abgewandelten) Beispiel den empirischen Gehalt der folgenden Hypothesen:

1 Wenn sich ein Himmelskörper auf einer geschlossenen Bahn bewegt, dann ist diese Bahn eine Ellipse.

2 Wenn sich ein Planet auf einer geschlossenen Bahn bewegt, dann ist diese Bahn eine Ellipse.

Die erste Hypothese verfügt über die größere Allgemeinheit und damit über den größeren empirischen Gehalt. Sie bezieht sich beispielsweise auch auf Monde und verbietet damit unter anderem, dass sich Monde in Quadraten bewegen. Die zweite Hypothese tut dies nicht. Den empirischen Gehalt der ersten Hypothese können wir nun weiter erhöhen, indem wir ihre Bestimmtheit steigern:

3 Wenn sich ein Himmelskörper auf einer geschlossenen Bahn bewegt, dann ist diese Bahn ein Kreis.

Jeder Kreis ist auch eine Ellipse, umgekehrt gilt dies aber nicht. Die Vorhersage der dritten Hypothese ist damit präziser. Nur sie verbietet, dass sich ein Himmelskörper anders bewegt als auf einer Kreisbahn. Das für unsere Zwecke Wesentliche ist nun, dass Hypothesen mit höherem empirischen Gehalt leichter zu falsifizieren sind. Beobachten wir etwa einen Mond, der sich auf einer nicht-kreisförmigen Ellipse bewegt, so widerlegt dies nur die dritte Hypothese. Umgekehrt gibt es keine Beobachtung, die die erste (oder zweite) Hypothese falsifizieren würde, nicht aber die dritte.

Theorien, die mehr mögliche Beobachtungen verbieten, sind also – positiv formuliert – informativer: Sie sagen uns für eine größere Menge von Fällen genauer, was wir erwarten sollten. Im Bereich der Sozialwissenschaften hätte eine Theorie etwa dann eine größere Allgemeinheit, wenn sie sich auf „alle Menschen“ bezieht und nicht nur auf „alle Kinder“ oder „alle Männer“. Eine Theorie über das Lernen wäre aber auch dann allgemeiner, wenn sich aus ihr Vorhersagen ableiten lassen, die sich sowohl auf das Lernen eines Tennisschlags beziehen als auch auf das Lernen von Vokabeln – und nicht nur auf einen dieser Bereiche. Theorien in den Sozialwissenschaften können sich natürlich auch hinsichtlich ihrer Bestimmtheit unterscheiden. Wir können beispielsweise leicht Theorien finden, die für irgendeine Lernaufgabe vorhersagen, dass die Menge des gelernten Stoffs mit der Lernzeit zunimmt. Es gibt aber auch Theorien, die diesen Zusammenhang genauer beschreiben, die also beispielsweise vorhersagen, ob in den ersten fünf Minuten genauso viel gelernt wird wie zwischen der 60. und 65. Minute und der 120. und 125. Minute (was einem linearen Zusammenhang zwischen

Lernzeit und gelernter Stoffmenge entspräche) oder ob vielleicht anfänglich schneller und später langsamer gelernt wird.

Aus dem Konzept des empirischen Gehalts ergeben sich nun eine Reihe wichtiger Konsequenzen: Nehmen wir an, eine Theorie erlaubt die Vorhersage, dass die Lebenszufriedenheit mit dem Einkommen wächst (mathematisch formuliert: zwischen Einkommen und Lebenszufriedenheit besteht eine monoton steigende Beziehung). Eine zweite Theorie hat einen höheren empirischen Gehalt, weil sie eine präzisere Vorhersage macht: Sie erwartet einen linearen Zusammenhang zwischen Einkommen und Zufriedenheit (eine Erhöhung des Einkommens um denselben Betrag geht stets mit dem gleichen Zuwachs in der Zufriedenheit einher). In einer Studie finden wir nun, dass die Zufriedenheit bis zu einem Jahreseinkommen von 12.000 Euro nur sehr langsam wächst, dann deutlich ansteigt, aber ab einem Einkommen von 80.000 Euro wieder nur langsam zunimmt. Diese Beobachtung falsifiziert die zweite Theorie, ist mit der ersten aber durchaus vereinbar. Wir sollten daher die erste Theorie bevorzugen. Wie verhält es sich aber, wenn wir in unserer Untersuchung tatsächlich einen linearen Anstieg der Zufriedenheit bei wachsendem Einkommen finden? In diesem Fall ist keine der beiden Theorien falsifiziert. Die zweite Theorie hat aber einer strengeren Prüfung standgehalten: Es gab für sie mehr Möglichkeiten, in unserer Studie falsifiziert zu werden. Theorien mit höherem empirischem Gehalt erlauben also strengere Prüfungen. Weil unsere Studie für die zweite Theorie eine strengere Prüfung war, hat sich zudem der *Grad der Bewährung* für diese Theorie stärker erhöht als für die erste Theorie – und dies, obwohl beide Theorien nur einem Test unterzogen wurden. Schließlich gibt uns der unterschiedliche Grad der Bewährung einen Grund, die zweite Theorie zu bevorzugen, obwohl keine der Theorien falsifiziert wurde: Die zweite Theorie erlaubt uns genauere Vorhersagen und konnte dennoch nicht widerlegt werden.

Wie steht's um den empirischen Gehalt psychologischer Theorien?

Sie werden in wenigen Seiten (im *Abschnitt 2.5.2*) in einem Beispiel die folgende Hypothese finden: Die Sachsen sind intelligenter als die anderen Deutschen. Sie werden dort auch lesen, dass diese Hypothese noch diverse Präzisierungen braucht, um prüfbar zu sein. Nehmen wir für den Moment an, dass eine der Präzisierungen lautet, dass die Sachsen *durchschnittlich* intelligenter sind als die übrigen Deutschen. Wie ist der empirische Gehalt dieser Hypothese zu bewerten? Wie wäre der empirische Gehalt der Hypothese zu bewerten, dass sich die durchschnittliche Intelligenz von Sachsen und anderen Deutschen unterscheidet? Für letztere Hypothese sollte deutlich sein, dass der empirische Gehalt extrem gering ist. Diese Hypothese schließt lediglich eine einzige mögliche Beobachtung aus, nämlich die, dass wir bei Sachsen und anderen Deutschen die gleiche durchschnittliche Intelligenz finden. Da wir aber nicht alle Sachsen und Deutschen untersuchen können, sondern nur eine Stichprobe, werden wir kaum zeigen können, dass die mittlere Intelligenz bei *allen* Sachsen und *allen* Deutschen exakt gleich ist. Man kann daher auch argumentieren, dass diese Hypothese so gut wie gar keinen empirischen Gehalt hat. Die ursprüngliche Hypothese (die Sachsen sind durchschnittlich intelligenter) ist offensichtlich präziser, aber auch ihr empirischer Gehalt ist noch gering. Sie verbietet nur die Hälfte der möglichen Beobachtungen. ▶

Dennoch werden Sie im weiteren Verlauf des Buchs viele Hypothesen finden, die eine ähnliche Struktur haben – also etwa besagen, dass die Gruppe A hinsichtlich irgendeines Merkmals durchschnittlich höhere Werte erzielt als die Gruppe B. Dies liegt zunächst vor allem daran, dass derartige Hypothesen in den statistischen Verfahren, mit denen wir Sie vertraut machen wollen, sehr gebräuchlich sind. Tatsächlich führen aber auch viele psychologische Theorien zu Hypothesen der Form „A ist größer als B". Sie sagen also beispielsweise vorher, dass irgendein Statistiktraining die Leistung verbessert, aber nicht, dass es die Leistung um zehn „Punkte" verbessert (was eine sehr präzise Hypothese wäre). Sie könnten vielleicht vorhersagen, dass höhere Intelligenz mit höherer Lebenszufriedenheit einhergeht, spezifizieren aber eher nicht, ob dieser Zusammenhang linear verläuft oder irgendeine andere Form hat. Dies ist durchaus ein Problem: Der niedrige empirische Gehalt der Vorhersage bedingt, dass auch eine erfolgreiche Prüfung der Theorie ihre Bewährung nur wenig erhöht. Das muss allerdings nicht bedeuten, dass auch der empirische Gehalt der gesamten Theorie gering ist. Die Theorie könnte durch den Umstand einen höheren empirischen Gehalt erzielen, dass sie weitere Vorhersagen erlaubt: Sie mag aus bestimmten Merkmalen von Statistiktrainings nicht nur folgern, dass Training A wirksamer ist als Training B, sondern auch, dass Training C beide anderen Trainings überbieten sollte und D am schlechtesten abschneidet. Eine Theorie, die Vorhersagen darüber macht, welche von zwei Optionen Sie bei einer Entscheidung wählen werden, mag auch Vorhersagen darüber machen, welche von zwei Entscheidungen Sie schneller treffen werden oder über welche Option Sie mehr Information suchen werden.

Es bleibt allerdings dabei, dass die erfolgreiche Prüfung der Hypothese „A ist größer als B" allein Sie nicht sonderlich von der dahinterliegenden Theorie überzeugen sollte. Ebenso sollten Sie nicht voraussetzen, dass eine Theorie schon weitere Vorhersagen erlauben wird (und dass Prüfungen dieser Vorhersagen erfolgreich verlaufen). Der geringe empirische Gehalt von „Theorien", die nur eine einzige Vorhersage der Form „A ist größer als B" erlauben, war schon vor Jahrzehnten Gegenstand heftiger Kritik (Meehl, 1967; siehe auch *Abschnitt 2.5.1*). Bei der Lektüre eines Forschungsartikels sollten Sie sich daher fragen, ob es um eine „brauchbare" Theorie geht, die weitere Vorhersagen erlaubt, und wie es um die Bestimmtheit dieser Vorhersagen steht. So die Theorie tatsächlich nur eine unpräzise Vorhersage macht, sollten Sie sich vielleicht einer anderen Theorie mit höherem empirischen Gehalt zuwenden – diese Theorie wäre informativer und damit einfach interessanter.

Der Grad der Bewährung kann selbst dann ein Grund sein, eine Theorie gegenüber einer anderen zu bevorzugen, wenn beide Theorien falsifiziert wurden. Die Falsifikation bedeutet natürlich, dass wir beide Theorien verwerfen und nach einer besseren Theorie suchen sollten. Dies muss eine falsifizierte Theorie aber nicht nutzlos machen. Nehmen wir an, dass zwei Theorien mit unterschiedlichem empirischen Gehalt diverse Prüfungen überstanden haben, nun aber beide durch dieselbe Beobachtung falsifiziert werden. Dann gilt immer noch, dass die Theorie mit höherem empirischen Gehalt an anderer Stelle mehr oder präzisere Vorhersagen zulässt. Denken Sie an die Newtonsche Physik: Diese Theorie ist durch bestimmte Beobachtungen falsifiziert, sie erlaubt aber offensichtlich dennoch zahlreiche präzise und erfolgreiche Vorhersagen. Auch eine falsche Theorie kann also *„wahrheitsnäher"* (ein Begriff, den Popper geprägt hat) sein als eine andere.

Falsifizierende Beobachtungen bedeuten nicht unbedingt, dass eine Theorie durch eine revolutionär andere Theorie ersetzt werden sollte oder wird. Die neue Theorie wird man oft als eine Weiterentwicklung verstehen können, die aus einer Modifikation der alten Theorie entsteht. Dabei besteht allerdings die Gefahr, die Theorie durch

sogenannte *Ad-hoc-Annahmen* lediglich so zu verändern, dass eine Falsifikation unmöglich wird.[4] Die Theorie wird „immunisiert". Ein Beispiel: Die Unconscious Thought Theory (Dijksterhuis & Nordgren, 2006) sagt vorher, dass Menschen bessere Entscheidungen treffen, wenn sie vor einer Entscheidung zwar etwas Zeit vergehen lassen, in dieser Zeit aber nicht über die Entscheidung nachdenken. Eine erste Überprüfung zeigt nun, dass diese Vorhersage zwar bei Männern zutrifft, bei Frauen aber falsch zu sein scheint (Dijksterhuis, 2004). Wir erhalten nun offensichtlich eine neue, nicht falsifizierte Theorie, wenn wir die ursprüngliche Theorie dahingehend modifizieren, dass sie nur für Männer gilt. Dieses Vorgehen wirkt aber ziemlich unbefriedigend. Sie können sich nun leicht erklären, warum: Der empirische Gehalt der Theorie ist durch die Ad-hoc-Annahme gesunken. Popper (1989) befasst sich daher mit der Frage, wann eine neue Theorie das Potenzial hat, „besser" zu sein als eine ältere Theorie, und daher einen ernsthaften Test lohnt. Er schlägt die Kriterien in ▶ Abbildung 2.4 vor. Die ersten beiden Postulate besagen zusammengenommen, dass die neue Theorie (T_2) mehr der bisher bereits bekannten Beobachtungen erklären können sollte als die alte Theorie (T_1). Der dritte Punkt entspricht der Forderung nach einem höheren empirischen Gehalt der neuen Theorie. Punkt 4 in ▶ Abbildung 2.4 meint schließlich, dass die neue Theorie in den zusätzlichen Prüfungen, die durch ihren höheren empirischen Gehalt möglich werden, auch erfolgreich sein muss. Erst wenn sich die neue Theorie in diesen Falsifizierungsversuchen bewährt hat, sollte sie die alte Theorie ersetzen.

Wann ist eine Theorie T_2 besser als eine Theorie T_1?

1. T_2 erklärt alles, was T_1 erklärt
2. T_2 erklärt einige der Beobachtungen, die T_1 nicht erklären kann
3. T_2 erlaubt weitergehende Prüfungen

4. T_2 bewährt sich in diesen Prüfungen

Abbildung 2.4: Kriterien des Kritischen Rationalismus für den Vergleich der Güte zweier Theorien.

Probleme der Falsifizierbarkeit

Falsifizierbarkeit scheint zunächst ein völlig eindeutiges Kriterium zu sein: Ein schwarzer Schwan beweist, dass die Vermutung „Alle Schwäne sind weiß" falsch ist. In der Praxis ergeben sich allerdings häufig Schwierigkeiten bei dem Versuch, darüber zu entscheiden, ob eine Theorie tatsächlich falsifiziert worden ist. Da der Kritische Rationalismus auf der Idee der Falsifizierbarkeit beruht, sind diese Schwierigkeiten auch ein Problem für den gesamten wissenschaftstheoretischen Ansatz.

Eine erste Schwierigkeit haben wir oben bereits kurz angesprochen: Empirisch fundierte Beobachtungsaussagen, die Basissätze, sind nicht theoriefrei und damit fallibel. Vielleicht war das vermeintliche Glas Wasser doch kein Glas Wasser. Wenn wir einen schwarzen Schwan gefunden haben, kann man infrage stellen, ob der fragliche Vogel tatsächlich ein Schwan war. Falsifikationen können also immer dadurch vermieden werden, dass man die falsifizierende Beobachtung bezweifelt oder zurückweist. Basis-

4 In der Literatur werden Sie auch die Bezeichnung Post-hoc-Annahme finden. Beide Begriffe werden synonym verwendet.

sätze beschreiben nicht einfach nur Fakten und sind folglich auch nicht eindeutig wahr (oder falsch). Sie haben eher den Charakter von einfachen Hypothesen, über deren Richtigkeit entschieden werden muss. Laut Popper (1989) kann (und sollte) diese Entscheidung in der Regel durch einen Konsens unter den beteiligten Forschern erfolgen. Man einigt sich also darauf, dass der Vogel tatsächlich ein Schwan war. Dies setzt natürlich zunächst voraus, dass die beteiligten Forscher gewillt sind, einen solchen Konsens zu suchen und anzugeben, unter welchen Bedingungen sie eine Beobachtung akzeptieren werden. Ein aus Poppers Sicht vorbildliches Beispiel gibt hier Einstein, der nicht nur angab, welche Beobachtung seine Relativitätstheorie falsifizieren würde, sondern auch wie diese Beobachtung durchzuführen wäre (die Vorhersage betrifft die Beugung des Lichts am Sonnenrand, die laut Einstein während einer Sonnenfinsternis zu beobachten sein müsste – und auch tatsächlich zu beobachten ist, wie sich herausstellte). Wenn wir uns stets die Möglichkeit offenhalten, Beobachtungen zurückzuweisen, ist Falsifikation nicht möglich und wir haben keine Chance mehr, aus Empirie zu lernen. Dennoch belegt auch eine Einigung unter Forschern nicht, dass ein Basissatz wahr ist. Basissätze bleiben selbst Gegenstand von Kritik und müssen möglicherweise aufgrund weiterer Beobachtungen später revidiert werden. Dies ändert aber nichts an der Asymmetrie zwischen Verifikation und Falsifikation: Ein akzeptierter Basissatz bietet eine logische Grundlage, um eine Theorie zu falsifizieren; er bietet keine Grundlage, die Theorie zu verifizieren.

Eine weitere Schwierigkeit betrifft allerdings einen logischen Aspekt der Falsifizierbarkeit. Betrachten wir noch einmal die Hypothese „Frustration erzeugt Aggression" und überlegen uns, wie diese Hypothese praktisch geprüft wird. Wie wir noch sehen werden, sind Experimente die beste Möglichkeit solche Hypothesen zu testen (*Kapitel 5*). Ein Experiment wiederum erfordert, dass wir zumindest zwei Gruppen miteinander vergleichen: In diesem Fall eine Gruppe von Personen, die frustriert wurden, mit einer Gruppe nicht-frustrierter Personen. Um dies zu realisieren, müssen wir etwas darüber wissen, wie wir Frustration herstellen können – und wir müssen letztlich annehmen, dass uns dies gelungen ist. Wir müssen zudem annehmen, dass sich die Gruppen nicht in anderer relevanter Weise unterscheiden, also etwa nicht aus Personen mit unterschiedlicher Aggressionsneigung bestehen, nicht unterschiedlich hungrig sind (auch dies könnte aggressiv machen) usw. Schließlich ist auch nicht offensichtlich, wie wir feststellen, wie aggressiv unsere Versuchsteilnehmer sind. Wir müssen also festlegen, wie wir Aggression messen – und wir müssen letztlich annehmen, dass uns diese Messung gelungen ist (siehe *Kapitel 3*). Die Prüfung unserer Hypothese ist also auf eine Reihe von Zusatzannahmen (oder „Hilfshypothesen") angewiesen. ▶ Abbildung 2.5 zeigt die veränderte Situation. Aus der Hypothese H und den mit logischem *und* verknüpften Zusatzannahmen A_1 bis A_n (alle Aussagen müssen zusammen zutreffen) lässt sich eine höhere Aggression in der Gruppe der Frustrierten (B) vorhersagen. Wenn diese Vorhersage nun nicht eintrifft ($\neg B$ in der zweiten Zeile in ▶ Abbildung 2.5), kann das verschiedene Ursachen haben: Zum einen kann die Hypothese falsch sein, aber es kann auch mindestens eine der Annahmen A_1 bis A_n nicht erfüllt sein (ausgedrückt durch die logischen *oder*-Verknüpfungen in der dritten Zeile in ▶ Abbildung 2.5). In der Wissenschaftstheorie wird dieses Problem als *Duhem-Quine-Problem* bezeichnet.

$$\frac{(H \wedge A_1 \wedge \ldots \wedge A_n) \rightarrow B \quad \neg B}{\neg H \vee \neg A_1 \vee \ldots \vee \neg A_n}$$

H: Hypothese
A_i: Zusatzannahme
B: Beobachtung

Abbildung 2.5: Die Problematik der Falsifizierbarkeit von Hypothesen in der Praxis, verursacht durch Zusatzannahmen (A_1 bis A_n), deren Erfülltsein vorausgesetzt werden muss. Die Zeichen in der Abbildung haben folgende Bedeutungen: $\wedge$ = logisches *und*, $\vee$ = logisches *oder*, $\rightarrow$ = logische *Implikation* und $\neg$ = *Negation*.

Poppers Antwort lautet hier, dass wir zum Zweck der Prüfung einer Theorie „Hintergrundwissen" nutzen und zumindest für den Moment als unproblematisch akzeptieren müssen. Wir müssen also beispielsweise voraussetzen, dass die von uns verwendeten Methoden zur Herstellung von Frustration und zur Messung von Aggression funktionieren. Andernfalls ist die eigentliche Theorie tatsächlich nicht falsifizierbar. Das Hintergrundwissen selbst ist aber natürlich nicht beliebig. Wir werden eine falsifizierende Beobachtung umso eher der Fehlerhaftigkeit der Theorie zuordnen können, je besser die von uns verwendeten Methoden bewährt sind. Wir sollten also etwa auf frühere Studien zurückgreifen können, in denen unsere Methoden zur Herstellung von Frustration und Messung von Aggression geprüft wurden. Der Test der Theorie sollte generell möglichst „verlässliches" Hintergrundwissen verwenden. Daraus ergibt sich, dass die Frage, ob eine Theorie falsifizierbar ist, nicht allein anhand der Formulierung der Theorie zu entscheiden ist. Falsifizierbarkeit hängt auch von der Güte der Methoden ab, die uns zur Prüfung der Theorie zur Verfügung stehen (Popper, 1989). Das Prinzip des Fallibilismus gilt aber natürlich auch für das am besten bewährte Hintergrundwissen. Auch dieses Wissen ist nie sicher, es bleibt kritisierbar und wird möglicherweise noch modifiziert werden. Eine Theorie wird daher kaum jemals in der Folge des falsifizierenden Ergebnisses einer einzelnen Studie aufgegeben werden. Die Studie selbst könnte in irgendeinem Sinn fehlerhaft sein. Studien sollten daher wiederholt werden (siehe *Kapitel 21*) und Theorien sollten in unterschiedlichen Experimenten überprüft werden, die unterschiedliches Hintergrundwissen nutzen (also etwa verschiedene Methoden zur Herstellung von Frustration).

Eine letzte Schwierigkeit der Falsifizierbarkeit betrifft die weitaus meisten Hypothesen in der Psychologe und den Sozialwissenschaften: Solche Hypothesen sind in aller Regel nicht deterministisch. Deterministische Hypothesen besagen, dass eine bestimmte Konsequenz immer und in allen Fällen eintreten wird, wenn die in der Hypothese spezifizierten Bedingungen gegeben sind. „Wenn Schwan, dann weiß" ist also ein Beispiel für eine deterministische Hypothese. „Auf Frustration folgt immer Aggression" wäre ein weiteres Beispiel. Allerdings würde die zweite Hypothese kaum in dieser Form formuliert werden. Das Auftreten von Aggression hängt recht offensichtlich auch von einer ganzen Reihe anderer Faktoren ab (etwa der Ängstlichkeit der frustrierten Person, den Machtverhältnissen zwischen den beteiligten Personen oder der Situation – im Fußballstadion ist Aggression vermutlich häufiger als in der Kirche). Frustration erzeugt also nicht *immer* Aggression. Hypothesen in den Sozialwissenschaften sind daher zumeist Wahrscheinlichkeitsaussagen (sogenannte *stochastische* oder *probabilistische* Hypothe-

sen). Die Hypothese „Frustration erzeugt Aggression“ meint eigentlich, dass Frustration die Wahrscheinlichkeit für Aggression erhöht. Wenn dies zutrifft, dann sollten wir im Mittel bei frustrierten Personen eine höhere Aggression finden als bei nicht-frustrierten. Allerdings bezieht sich diese Aussage auf Mittelwerte in der Population – auf alle Fälle von mehr oder weniger Frustration, die für die Hypothese relevant sind. Wir können aber natürlich niemals alle Fälle untersuchen. Uns liegt lediglich eine Stichprobe von frustrierten und nicht-frustrierten Personen vor. Die Mittelwerte in einer Stichprobe geben uns aber keine eindeutige Auskunft über die Mittelwerte in der Population. Nehmen Sie an, wir wollten die Hypothese prüfen, dass Männer im Mittel größer sind als Frauen. Es dürfte kaum Zweifel geben, dass diese Hypothese (in der Population) zutrifft. Das schließt aber nicht aus, dass wir auf Stichproben von (sagen wir) zehn Männern und Frauen stoßen, in denen die Frauen im Mittel größer sind. Diese Beobachtung würde die Hypothese nicht falsifizieren. Tatsächlich gibt es keine Beobachtung, die eine probabilistische Hypothese eindeutig falsifizieren würde. Wir können aber auch probabilistische Hypothesen einem strengen Test unterziehen: Ein Test ist hier dann streng, wenn sich die Wahrscheinlichkeit des Ergebnisses in Abhängigkeit davon, ob die geprüfte Hypothese wahr oder falsch ist, deutlich unterscheidet (diesen Gedanken werden Sie in der Grundidee des Signifikanztests wiederfinden, siehe *Kapitel 12*). Dies können wir zum Beispiel dadurch erreichen, dass wir die Stichprobengröße erhöhen. Wenn wir Stichproben von nur zwei Männern und Frauen betrachten, ist ein größerer Mittelwert für die beiden Frauen auch dann noch gut möglich, wenn Männer „eigentlich“ (also in der Population) größer sind. Mit einer Stichprobengröße von 2000 Männern und Frauen würde dieses Ergebnis sehr unwahrscheinlich. Wir können also Beobachtungen machen, die auch mit einer probabilistischen Hypothese nur schwer zu vereinbaren sind. Dadurch werden probabilistische Hypothesen zumindest „praktisch falsifizierbar“: Wiederholte, stark abweichende Beobachtungen liefern uns gute Gründe solche Hypothesen zurückzuweisen.

Die diskutierten Probleme der Falsifizierbarkeit sollten deutlich machen, dass es ein mühsamer und arbeitsreicher Weg sein kann, Theorien zu falsifizieren. In der Praxis wird oftmals nicht unmittelbar zu entscheiden sein, ob und inwieweit eine bestimmte Beobachtung eine Theorie tatsächlich falsifiziert. Die Frage, ob eine Theorie als falsifiziert zu betrachten ist, wird in der Regel Gegenstand langwieriger Diskussionen bleiben. Poppers Antwort bleibt bei all diesen Schwierigkeiten letztlich aber immer dieselbe: Wir sollten uns diesen Mühen nicht entziehen. Wir sollten unsere Theorien kritisch prüfen und möglichst strengen empirischen Tests unterziehen. Nur dies bietet uns die Chance, dazuzulernen.

2.2.3 Historisch-soziologische Analyse (Kuhn)

Während der Logische Empirismus gewissermaßen Regeln dafür vorgibt, wie Theorien auszusehen haben, und der Kritische Rationalismus dafür, wie man Theorien sinnvoll überprüfen kann, ging Thomas Kuhn, ein amerikanischer Physiker, Wissenschaftsphilosoph und Wissenschaftshistoriker, einen anderen Weg. Er untersuchte, wie Wissenschaft tatsächlich funktioniert. Als zentrale Begriffe in seinem Ansatz benutzt er die *wissenschaftliche Gemeinschaft* und das *Paradigma*. Kuhn kommt in seinen Arbeiten zu dem Schluss, dass Wissenschaft nicht im Elfenbeinturm, sondern in Spezialistengruppen, den wissenschaftlichen Gemeinschaften, stattfindet, und dass man charakteristische Gemeinsamkeiten zwischen den Mitgliedern spezifizieren

kann, die er als Paradigma bezeichnet. Wissenschaftliche Gemeinschaften zeichnen sich dadurch aus, dass ihre Mitglieder intensiv miteinander kommunizieren, dass sie in ihren Urteilen über Aspekte ihres wissenschaftlichen Untersuchungsgegenstandes in hohem Maße übereinstimmen und dass es viele gemeinsame Elemente in der Ausbildung des Nachwuchses und in den Kenntnissen der Mitglieder gibt. Ein Paradigma umfasst allgemein akzeptierte theoretische Annahmen, Gesetze und empirische Generalisierungen, erfolgreich bewertete Anwendungen, häufig eingesetzte Hilfsmittel und Apparaturen sowie allseits akzeptierte Methoden und Begriffsbildungen.

Diese Beschreibung der Wissenschaftsgemeinde hätte aber sicherlich nicht gereicht, um Kuhn berühmt zu machen. Das erreichte er durch einige erstaunliche und provozierende Thesen (z.B. Kuhn, 1981). Diese Thesen haben damit zu tun, was Wissenschaftler tatsächlich tun, wenn sie auf sogenannte Anomalien stoßen, d.h. auf wissenschaftliche Ergebnisse, die nicht in ein Paradigma passen. Zunächst stellte er fest, dass Wissenschaftler so gut wie nie strikt dem Falsifikationsprinzip folgen und ihre Theorien verwerfen, wenn die Vorhersagen nicht eintreffen, sondern zunächst versuchen, sie durch Zusatzannahmen oder Zusatzerklärungen zu retten.[5] Am bekanntesten ist aber wohl Kuhns These, dass Wissenschaft nicht kumulativ betrieben wird (neue Erkenntnisse bauen notwendigerweise auf alten auf), sondern eher in Sprüngen, die er „wissenschaftliche Revolutionen" nennt. Wissenschaftliche Revolutionen sind aber relativ selten und treten meist erst nach langen „normalwissenschaftlichen Forschungsperioden" auf. In den normalwissenschaftlichen Forschungsperioden werden existierende Paradigmen modifiziert und erweitert, wenn Anomalien auftreten. Wenn sich Anomalien allerdings längere Zeit solchen Auflösungsversuchen widersetzen, kann das zu einer Krise der normalen Wissenschaft und in eine Phase außerordentlicher Wissenschaft führen. Diese Phase ist dadurch gekennzeichnet, dass verschiedene miteinander konkurrierende Theorien – Modifikationen des bestehenden Paradigmas – auftauchen, die aber letztlich alle nicht voll befriedigend sind. Das führt dazu, dass die bis dahin akzeptierten Methoden, Regeln und Normen zur Diskussion gestellt werden. Aber auch in dieser Phase ist es noch möglich, dass die Anomalie durch eine Modifikation des herrschenden Paradigmas aufgelöst wird und es wieder zu einer Periode normalwissenschaftlicher Forschung kommt. Lässt sich eine Anomalie aber nicht auflösen, dann kann sie in eine *wissenschaftliche Revolution* münden, bei der das alte Paradigma durch ein neues ersetzt wird. Das neue Paradigma baut aber nicht auf dem alten auf, und altes und neues Paradigma sind meist in vielen Punkten nicht vergleichbar.[6] Trotzdem geht Kuhn davon aus, dass die Wissenschaft Fortschritte macht: Neue Theorien (nach der Revolution) sind genauer, spezialisierter und können mehr erklären.

5 Kuhn (1981, 211) zitiert als Kronzeugen dafür auch Max Planck, der die Ansicht vertrat, dass (schlechte) Theorien erst mit ihren Gründern sterben.

6 Ein Beispiel, das Kuhn anführt, ist die Ablösung des Ptolemäischen Weltbilds (Erde als Mittelpunkt, Sonne, Mond, Merkur, Venus, Mars, Jupiter, und Saturn als „Planeten", die sich um die Erde drehen) durch das Kopernikanische Weltbild (Sonne als Mittelpunkt und als neu eingeführte Kategorie Merkur, Venus, Erde, Mars, Jupiter und Saturn als „Planeten", die sich um die Sonne drehen, und Monde als ebenfalls neu eingeführte Kategorie der „Satelliten"). Bei diesen beiden Paradigmen ist beispielsweise die Kategorie „Planeten" nicht in gleicher Weise verwendbar. Ein anderes Beispiel ist die Ablösung der Newton'schen Physik durch Einsteins Relativitätstheorie. Auch diese beiden Theorien sind in weiten Teilen nicht vergleichbar. So verwenden sie beispielsweise beide den Begriff „Energie", der jedoch unterschiedliche und nicht vergleichbare Bedeutungen hat.

2.2.4 Methodologie wissenschaftlicher Forschungsprogramme (Lakatos)

Die von Imre Lakatos (1974) vorgeschlagene *Methodologie wissenschaftlicher Forschungsprogramme* wird gemeinhin als Versuch angesehen, Poppers Grundidee der Falsifizierbarkeit von Theorien so zu erweitern, dass sie mit Kuhns Paradigmenkonzept vereinbar wird. Lakatos, ein ungarischer Mathematiker, Physiker und Wissenschaftstheoretiker, argumentiert, dass es nicht sinnvoll sei, eine Theorie zu falsifizieren, solange keine bessere vorhanden sei. Ebenso wenig sei es sinnvoll, isolierte Theorien zu betrachten, da Theorien immer in einen Kontext eingebettet seien und als Theorienreihen T_1, T_2, T_3, … aufträten. Die jeweils nächste Theorie entsteht dabei als Reaktion auf eine Anomalie oder einen mit der vorhergehenden Theorie nicht zu vereinbarenden empirischen Befund. Durch dieses Aufeinanderbezogensein wird die Theorienreihe zu einem *wissenschaftlichen Forschungsprogramm*. Lakatos argumentiert zudem, dass Theorien aus einem harten Kern bestehen, in dem sich die Theorien einer Reihe nicht unterscheiden und der auch – weil sich die Mitglieder eines Forschungsprogramms dazu entschieden haben – nicht falsifizierbar ist, sowie aus einem „Schutzgürtel" von Hilfshypothesen, die falsifiziert und gegebenenfalls modifiziert werden können. Die Entscheidung, die getroffen werden muss, bezieht sich also nicht auf zwei Theorien (ein Entscheidungsexperiment ist in Lakatos' Ansatz nicht sinnvoll), sondern auf zwei konkurrierende Forschungsprogramme. Wann soll nun ein Forschungsprogramm zugunsten eines anderen aufgegeben werden? Dies soll nur geschehen, wenn ein Programm stagniert, d.h. wenn seine Erklärungskraft (die Postulate 1 bis 3 in ▶ Abbildung 2.4) nicht mehr wächst und wenn zudem ein Forschungsprogramm mit einer größeren Erklärungskraft existiert. Nach Lakatos ist es durchaus möglich, schon verworfene Forschungsprogramme wieder zu reaktivieren, sollte sich nachträglich herausstellen, dass Anomalien zufriedenstellend durch eine Erweiterung des Programms erklärt werden können. Außerdem plädiert er dafür, „junge" Theorien nicht vorschnell aufzugeben.

2.2.5 Wirklichkeit als Konstruktion

Das Anliegen der konstruktivistischen Ansätze ist vielleicht am ehesten zu verstehen, wenn man sie mit den konventionellen Ansätzen kontrastiert. Die Kritik der konstruktivistischen Ansätze richtet sich in erster Linie nicht gegen einen speziellen konventionellen Ansatz, sondern gegen eine zentrale Grundannahme, die Grundlage des Logischen Empirismus (siehe oben) ist, die aber alle konventionellen Ansätze mehr oder weniger teilen, und die unter dem Begriff *Positivismus* bekannt ist. Wir werden zunächst die Position des Positivismus skizzieren und uns dann mit der Kritik daran auseinandersetzen. Anschließend beschreiben wir exemplarisch einen konstruktivistischen Ansatz, der sich explizit auf die Psychologie bezieht.

Positivismus und Positivismus-Kritik

Der Name Positivismus wurde von Auguste Comte, dem Begründer der Soziologie, geprägt. Positivismus als wissenschaftstheoretische Position bedeutet, dass man das „Positive" zum Prinzip allen wissenschaftlichen Wissens macht. Positiv meint hier aber nicht das Gegenteil von „negativ" im abwertenden Sinn, sondern bezeichnet das Gegebene, Tatsächliche oder unbezweifelbar Vorhandene. In der Erkenntnistheorie

nimmt der Positivismus eine Mittelstellung ein zwischen dem Materialismus (alles ist Materie und unsere Sinne spiegeln die Außenwelt getreu wider) und der Transzendentalphilosophie (wir können über die Dinge an sich im Prinzip nichts sagen, sie bleiben jenseits der Sinneswahrnehmungen). Im Unterschied zu den Materialisten legen sich die Positivisten nicht bezüglich der Natur der Außenwelt fest (letztlich weiß man nichts darüber), stützen sich jedoch auf das Positive, also die Sinnesdaten. Diese Sinnesdaten werden dann „denkökonomisch" interpretiert, das heißt, so, dass man keine unnötigen „Instanzen und Wesenheiten" mit ins Spiel bringen muss. Ein Beispiel soll den Unterschied der drei Positionen verdeutlichen: Gibt es Gott? Im Materialismus heißt die Antwort eindeutig „Nein" – hier gibt es nur Materie. Die Transzendentalisten halten dagegen einen Raum für Gott offen. Die Positivisten sagen jedoch, dass man – so wie das Konzept „Gott" definiert ist – mithilfe von Sinnesdaten keine vernünftige Aussage darüber machen kann: Die Frage nach der Existenz Gottes ist ein „Scheinproblem" und kann somit nicht wissenschaftlich bearbeitet werden.

Als weitere Charakteristika des Positivismus – bezogen auf die Psychologie wird Positivismus hier meist mit „von den Naturwissenschaften übernommener konventioneller Ansatz" gleichgesetzt – gelten (z.B. Ashworth, 2003, 11):

- Es gibt eine einheitliche reale Welt, in der die Ereignisse, die für die Psychologie interessant sind, stattfinden.
- Das Individuum ist Teil dieser realen Welt, genauso wie Gedächtnisprozesse, Emotionen und Gedanken; und alle diese Vorgänge haben überdauernde Eigenschaften.
- Der Zweck von Wissenschaft ist es, experimentelle Situationen zu erzeugen, in denen sich die Eigenschaften psychologischer Prozesse offenbaren; und das erlaubt es wiederum, diese Prozesse nachzubilden.
- Die Welt kann als Gefüge von messbaren Variablen[7] beschrieben werden, die miteinander auf gesetzmäßige Weise interagieren können.
- Die Modelle (wenn möglich in mathematischer Formulierung) sollen zeigen, wie Variablen zusammenwirken, insbesondere, wie sie dies in einer Ursache-Wirkungs-Beziehung tun.
- Der Zweck von Forschung ist es, Hypothesen darüber zu testen, wie Variablen zusammenwirken und – durch immer engere Approximation – zu Theorien zu gelangen, die man nach und nach als wissenschaftliche Gesetzmäßigkeiten betrachten kann.

Mit den genannten Charakteristika würde sich wohl die große Mehrheit der in der wissenschaftlichen Psychologie tätigen Forscher (mit einigen kleineren Einschränkungen oder Zusätzen) identifizieren. Die Konstruktivisten bezweifeln aber schon den ersten angeführten Punkt: Ihnen zufolge gibt es *keine* unabhängig von uns existierende Welt. Jeder Mensch *konstruiert* sich seine eigene Welt und die Aufgabe der Wissenschaft ist es, diese Welt zu entdecken. Relevante Forschungsfragen sind dann: Wie sieht diese Welt aus? Wodurch wird sie erzeugt? Wie und wodurch wird sie beeinflusst und gesteuert? Konstruktivisten bemängeln auch, dass die Konzeption der Welt

7 Der Begriff „Variable" bezeichnet hier Merkmale im weitesten Sinn, die bei allen oder den meisten Menschen vorhanden sind, die aber unterschiedliche (variable) Werte annehmen können. Beispielsweise kann die Variable „Alter" – wenn man die Werte in vollen Jahren angibt – Werte zwischen 0 und ca. 125 annehmen. Siehe hierzu auch *Kapitel 3* und *5*.

als „Variablenstruktur“ dem Verständnis menschlichen Erlebens und Verhaltens nicht gerecht werden kann. Deshalb sind sie mit dem konventionellen Methodenspektrum nicht zufrieden und spezialisieren sich auf „qualitative Methoden“. „Qualitativ“ steht hier im Kontrast zu „quantitativ“ und drückt aus, dass die Ergebnisse von Studien nicht in Zahlen ausgedrückt werden müssen, sondern dass die Daten auch Texte sein können, die entsprechend interpretiert werden (*Kapitel 32*). Wie schon eingangs erwähnt, gibt es kein einheitliches „konstruktivistisches Lager“. Die inhaltliche und weltanschauliche Variation der Ansätze, die in diesem Buch als „konstruktivistisch“ bezeichnet werden, ist tatsächlich deutlich größer als die der vorherrschenden konventionellen Ansätze. Einer dieser Ansätze – die *diskursive Psychologie* – soll im Folgenden exemplarisch beschrieben werden (siehe auch *Kapitel 32*).

Diskursive Psychologie

Die diskursive Psychologie ist ein relativ junger Ansatz (Potter & Wetherell, 1987), der in der Tradition der „Diskursanalyse“ steht, deren Entstehung wiederum meist auf Wittgenstein zurückgeführt wird. Die zentrale Annahme aller diskursanalytischen Ansätze ist, dass die Sprache – in einem weiten Sinne verstanden – unterschiedliche Versionen sozialer Realität konstruiert und zum Erreichen sozialer Absichten und Ziele eingesetzt wird.[8] Die diskursive Psychologie befasst sich hauptsächlich mit Letzterem: Wie gebrauchen Menschen Sprache (Diskurs), um ihre Interessen in sozialen Interaktionen durchzusetzen? Was *tun* Menschen mit ihrer Sprache? Worauf zielen sie ab? Auch die diskursive Psychologie ist vielleicht am leichtesten zu verstehen, wenn man sich vergegenwärtigt, was sie am konventionellen Ansatz kritisiert. Die folgende Aufzählung soll die Position der diskursiven Psychologie anhand ihrer Kritik (*in kursiver Schrift*) an einigen allgemein akzeptierten Ansichten illustrieren (siehe Willig, 2003):

- Kognitionen basieren auf Wahrnehmungen (Kognitionen sind Repräsentationen realer Objekte, Ereignisse und Prozesse).

  Kritik: Objekte und Ereignisse werden durch die Sprache erst konstruiert.

- Eine objektive Wahrnehmung der Realität ist theoretisch möglich (Fehler werden durch zeitsparende, aber im Prinzip kontrollierbare Heuristiken erzeugt).

  Kritik: Sprache konstruiert die soziale Realität, aber repräsentiert sie nicht unbedingt; eine objektive Wahrnehmung der Realität ist also prinzipiell nicht möglich.

- Es herrscht Konsens über die Existenz von sozialen Objekten und Ereignissen (die Menschen stimmen darin überein, worüber sie reden [z.B. die Europäische Währungsgemeinschaft oder den Zusammenbruch der Sowjetunion], obwohl sie möglicherweise unterschiedliche Erklärungen [Attributionen] dafür haben und auch unterschiedliche Einstellungen dazu).

  Kritik: Soziale Objekte und Ereignisse werden durch Sprache konstruiert (Attributionen und Einstellungen sind Aspekte dieser Konstruktionen).

8 Neben der diskursiven Psychologie existieren noch zahlreiche andere diskursanalytische Ansätze, die sich teilweise deutlich voneinander unterscheiden. Die diskursive Psychologie scheint jedoch der einzige Ansatz zu sein, der innerhalb der Psychologie entwickelt wurde.

- Es gibt relativ überdauernde kognitive Strukturen (Glaubensinhalte, Einstellungen, Attributionen usw. können zwar durch sie beeinflussende Variablen geändert werden, sind aber einigermaßen stabil).

  Kritik: Vorurteile, Identität, Gedächtnis, Vertrauen usw. ist etwas, was Menschen TUN*, nicht was sie haben.*

Die zentrale Frage, die sich die diskursive Psychologie stellt, ist also, wie Menschen Sprache gebrauchen (einen Diskurs führen), um ihre Interessen in sozialen Interaktionen durchzusetzen. Man kann sich nun fragen, ob Sprache – selbst im weitesten Sinne interpretiert – menschliches Erleben und Verhalten in seiner Gänze abbildet und ob somit die diskursive Psychologie als alternativer Ansatz zur konventionellen psychologischen Forschung brauchbar ist. Die konstruktive Funktion von Sprache wird zwar auch von der „Mainstream"-Psychologie anerkannt (siehe z.B. *Abschnitt 1.2*), aber die diskursive Psychologie und vergleichbare konstruktivistische Ansätze betrachten das Problem nicht als empirisch, sondern als erkenntnistheoretisch. Wenn man keine unabhängig von uns existierende Realität postulieren kann, dann muss dies Auswirkungen auf die Forschungspraxis und auf die Erkenntnismöglichkeiten der Wissenschaft haben.

Trotz dieser provozierenden Postulate gibt es bisher keine nennenswerten Reaktionen des vorherrschenden Wissenschaftsbetriebs in der Psychologie auf die konstruktivistischen Ansätze (ganz im Gegensatz zu anderen Disziplinen, wie z.B. der Soziologie). Zu einem Teil ist dies sicher darauf zurückzuführen, dass in den hauptsächlich gelesenen Fachzeitschriften so gut wie keine konstruktivistisch motivierten Studien zu finden sind. Das wiederum hängt wohl zum einen damit zusammen, dass qualitative Methoden (*Kapitel 32*), derer sich Konstruktivisten bedienen, in der psychologischen Methodenausbildung bisher ein Schattendasein führen, zum anderen aber auch damit, dass die Resultate der Konstruktivisten bislang von der psychologischen Forschergemeinschaft als wenig relevant erachtet werden. Wie brauchbar konstruktivistische Ansätze sind und auf welche Gegenstandsbereiche sie angewandt werden können, sind nach wie vor offene Fragen (siehe auch hierzu *Kapitel 32*).

2.3 Spezialprobleme der Psychologie

Bevor wir uns nun wieder den Theorien im engeren Sinn zuwenden, befassen wir uns noch kurz mit den Besonderheiten der Psychologie. Keiner der wissenschaftstheoretischen Ansätze, die wir beschrieben haben, ist (mit Ausnahme der diskursiven Psychologie, die aber auch eine Modifikation eines schon existierenden Ansatzes ist) direkt für die Psychologie entwickelt worden. Die konventionellen Ansätze sind sämtlich in Auseinandersetzung mit den Naturwissenschaften entstanden. Die Psychologie hat diese Ansätze weitgehend unverändert übernommen, unter anderem vielleicht, um vom Prestige der Naturwissenschaften zu profitieren. Dies ist in einigen Bereichen der Psychologie gerechtfertigt – Wahrnehmungs- und auch komplexere Informationsverarbeitungsprozesse sind bei vielen Tieren so ähnlich wie beim Menschen, was eine Trennung zwischen Psychologie und Naturwissenschaften (wenn man Biologie als Naturwissenschaft betrachtet) nicht leicht macht. Trotzdem ist menschliches Erleben und Verhalten ein Forschungsgegenstand mit Besonderheiten, die über die Besonderheiten von Wasser, Elektrizität und auch über die eines Diamanten, einer Rose oder

eines Pferdes hinausgehen. Diese Besonderheiten gilt es in der wissenschaftlichen Psychologie zu berücksichtigen. Zwei davon scheinen uns besonders bedeutsam: das Postulat von sogenannten *latenten Variablen* und das besondere Verhältnis von Forscher und Forschungsgegenstand, das auch in den Sozialwissenschaften eine bedeutende Rolle spielt.

2.3.1 Latente Variablen

Auf ein großes Problem der Psychologie werden viele Menschen entweder nie oder erst auf den zweiten Blick aufmerksam: Wenn wir über Intelligenz, Gedächtnis, Ärger, Freude, Liebe usw. reden, dann behandeln wir diese Ausdrücke, als bezeichneten sie etwas, was tatsächlich als Entität vorhanden ist und worauf wir direkten Zugriff haben. In gewisser Weise haben wir Zugriff – jeder macht Einschätzungen über die Intelligenz von sich und anderen und kann sich an Situationen erinnern, in denen er ärgerlich war oder sich gefreut hat. Wenn man aber versucht, Intelligenz, Gedächtnis, Ärger und Freude genauer zu definieren, merkt man, dass das nicht so einfach ist. Am einfachsten scheint es noch mit der Intelligenz zu sein, aber auch hier lautet eine – ursprünglich nicht so ernst gemeinte, aber mittlerweile durchaus akzeptierte – Definition: „Intelligenz ist das, was der Intelligenztest misst". Sieht man sich jedoch in der Welt der Intelligenztests um, dann wird schnell deutlich, dass die in unterschiedlichen Tests enthaltenen Aufgaben sich teilweise deutlich voneinander unterscheiden. Das ursprüngliche Intelligenzkonzept ist überdies um Konzepte wie emotionale Intelligenz oder soziale Intelligenz stark erweitert worden. Allgemein werden die Aufgaben in den Tests als Indikatoren für etwas „Dahinterliegendes" benutzt, etwas, das man eben nicht direkt messen kann: die Intelligenz. Je mehr Aufgaben richtig gelöst werden, desto höher ist die diagnostizierte Intelligenz. Ähnlich misst man auch Gedächtnisgüte oder die Stärke von Emotionen wie Ärger und Freude. Dieses Dahinterliegende, von dem man jeweils annimmt, dass es mit entsprechenden Testaufgaben abgebildet werden kann, bezeichnet man in der Psychologie als *latente Variable*, *Konstrukt*, oder *Faktor*. Offensichtlich bestimmt die Art, wie man solche latenten Variablen misst, auch deren Inhalt. Zur Bestimmung von latenten Variablen sind in der Psychologie mittlerweile ausgeklügelte Verfahren entwickelt worden, wie beispielsweise die Faktorenanalyse (siehe *Kapitel 25*). Man sollte sich als Psychologin oder Psychologe jedoch stets bewusst sein, dass die Inhalte von latenten Variablen immer vorläufig sind und sich ändern können. Im Zweifelsfall wird man sich die Verfahren, mithilfe derer eine latente Variable definiert wurde, genau ansehen müssen. Ein ähnliches Problem gibt es natürlich auch in den Naturwissenschaften. Auch dort können manche Objekte mit den heute zur Verfügung stehenden Mitteln nicht direkt beobachtet werden (z.B. Elementarteilchen), selbst wenn die entsprechenden latenten Variablen präzise definiert sind. Die Bedeutsamkeit von latenten Variablen ist aber in der Psychologie deutlich größer, da viele zentrale Konzepte in diese Rubrik fallen.

2.3.2 Verhältnis zwischen Forscher und „Erforschten"

In den Naturwissenschaften gibt es eine klare Trennung zwischen Forscher und Forschungsgegenstand. Selbst in der Biologie, in der auch der Mensch Forschungsgegenstand sein kann, ist die Trennung eindeutig. In Psychologie und Sozialwissenschaften haben wir es jedoch mit der Besonderheit zu tun, dass die Erforschten im Prinzip

auch Forscher sein können. Tatsächlich war es in der Frühzeit der Psychologie, vor allem in Deutschland, durchaus üblich, dass Professoren als Versuchsteilnehmer gearbeitet haben – etwa bei Studien darüber, wie Denk- und Gedächtnisprozesse ablaufen (man nahm an, dass Professoren aufgrund ihrer Übung im Denken ihre eigenen Denkprozesse unvoreingenommener wahrnehmen können). Auch während des Psychologiestudiums findet beispielsweise ein solcher Rollenwechsel statt: Studierende in den Anfangssemestern müssen in der Regel nachweisen, dass sie eine festgelegte Anzahl von Stunden an psychologischen Studien teilgenommen haben. Spätestens bei der Durchführung ihrer Bachelor- oder Masterarbeit befinden sie sich aber in der Regel selbst in der Rolle von Forschern.

Warum kann die Umkehrbarkeit der Rollen problematisch sein? Versuchsteilnehmer bilden automatisch Erwartungen über das, was andere – wie beispielsweise Forscher – von ihnen erwarten und sind demgemäß selbst auch wieder Forscher mit dem Versuchsleiter als Forschungsgegenstand („Was könnte der von mir wollen?“ – siehe Kasten „Versuchsteilnehmer als Forscher“). Die Erwartungen, die auf beiden Seiten des Forschungsprozesses – Forscher und Erforschte – vorhanden sind, können Verhalten und Erleben stark beeinflussen. Der Einfluss von Erwartungen ist auch bei Forschern nachgewiesen worden und wir werden uns damit und mit den Möglichkeiten, solche Einflüsse zu minimieren, in *Kapitel 4* und *5* befassen. Festzuhalten bleibt, dass die prinzipielle Austauschbarkeit der Rollen von Forscher und Erforschten und die Möglichkeit der „Messobjekte“, auf die Messung zu reagieren, ein potenzielles Problem in der psychologischen und sozialwissenschaftlichen Forschung ist, dessen man sich immer bewusst sein sollte.

Versuchsteilnehmer als „Forscher“

Norenzayan und Schwarz (1999) baten Studierende um ihre Meinung zu einem Massenmord, der tatsächlich kurze Zeit vor der Studie in den USA stattgefunden hatte. Zwei Gruppen von Studierenden erhielten einen Zeitungsartikel über den Massenmord und sollten dann auf einem Fragebogen die Gründe angeben, die ihrer Meinung nach dazu geführt hatten. Die Informationen, die beide Gruppen bekamen, unterschieden sich nur darin, was auf dem Briefkopf des Fragebogens stand. Bei der einen Gruppe war als durchführende Institution „Institut für Persönlichkeitsforschung“ und bei der anderen Gruppe „Institut für Sozialforschung“ angegeben. Die von den Studierenden angeführten Gründe für den Massenmord unterschieden sich erheblich in Abhängigkeit der Gruppenzugehörigkeit. Wenn der Fragebogen vermeintlich vom Institut für Persönlichkeitsforschung stammte, dann wurden Gründe, die in der Person des Täters lagen, deutlich häufiger angeführt, als wenn der Fragebogen vom vermeintlichen Institut für Sozialforschung kam. Mit Gründen, die in der sozialen Situation, in der der Mord stattfand, lagen, verhielt es sich dagegen umgekehrt. Solche Gründe wurden deutlich häufiger genannt, wenn der Briefkopf vom vermeintlichen Institut für Sozialforschung stammte. Diese Studie verdeutlicht, dass Menschen keine physikalischen „Messobjekte“ sind, sondern selbst als „Forscher“ tätig werden können und die „Forschungsergebnisse“ (z.B. „Die wollen anscheinend was zu den Gründen wissen, die in der *Person* des Täters liegen“) in ihrem Verhalten berücksichtigen.

2.4 Woher kommen Theorien?

Während die Überprüfung von Theorien einen großen Stellenwert sowohl in der wissenschaftstheoretischen als auch der Methodik-Literatur einnimmt, findet man zum Thema „Woher kommen Theorien?“ relativ wenig. Wenn man Wissenschaftler fragt, wie sie auf eine gute Idee oder Theorie gekommen sind, dann ist eine häufige Antwort, dass sie ihnen einfach „eingefallen“ sei: Theorien werden also – zumindest im Anfangsstadium – oft nicht systematisch entwickelt. Trotzdem kann man einige begünstigende Bedingungen für das Entstehen von Theorien etwas genauer spezifizieren: *bed, bathroom and bicycle* (*Abschnitt 2.4.1*). Darüber hinaus existieren aber auch systematische (qualitative) Ansätze zur Entwicklung von Theorien. Egal, wie man zu Theorien kommt, eine Eigenschaft ist sicher ganz zentral für einen erfolgreichen Wissenschaftler: grenzenlose Neugierde.

2.4.1 Bed, Bathroom and Bicycle

In der englischsprachigen Literatur findet man auf die Frage, woher denn Theorien kommen, häufig die Antwort: *bed, bathroom, bicycle*. Diese Antwort drückt die Überzeugung aus, dass einem gute Theorien nicht dann einfallen, wenn man mit voller Konzentration danach sucht, sondern eher, wenn man entspannt ist oder nicht ans Arbeiten denkt. Berühmte Beispiele (mit nicht eindeutig geklärtem Wahrheitsgehalt) sind Archimedes, der in der Badewanne das Archimedische Prinzip (der hydrostatische Auftrieb eines Körpers ist gleich dem Gewicht der von ihm verdrängten Flüssigkeit) entdeckte und daraufhin sein berühmtes „Heureka!“ rief, und Kekulé, dem beim Träumen im Lehnstuhl der Benzolring einfiel. Tatsächlich ist es bei vielen Theorien schwierig, ihre Entstehungsgeschichte genau nachzuvollziehen. Intuition scheint eine wichtige Rolle zu spielen. Manchmal werden auch induktive Verfahren zu Hilfe genommen und gar nicht so selten sind Theorien durch eine Art Analogieschluss entstanden: Man benutzt Begriffe aus einem anderen Bereich – z.B. aus den Computerwissenschaften – und überträgt sie auf menschliches Verhalten und Erleben.

Intuition

Wenn man annimmt, dass die oben angeführten Beispiele von Archimedes und Kekulé zutreffen – was sollte dann ein guter Wissenschaftler tun? Häufig baden und tagträumen? Auf keinen Fall aber zu viel arbeiten, denn dann ist es schwieriger, sich zu entspannen. Diese Strategie könnte natürlich in dem einen oder anderen Fall funktionieren (die Autoren kennen allerdings kein bekanntes Beispiel dafür), aber meistens versucht man, der Intuition etwas auf die Sprünge zu helfen. Intuition bedeutet ja nichts anderes, als dass Ideen scheinbar spontan entstehen. Diese Spontaneität scheint allerdings häufig an ein hohes Maß Vorarbeit gekoppelt zu sein. Eine Methode, der Intuition auf die Sprünge zu helfen, ist sicher: viel lesen oder sich mit Menschen unterhalten, die sich auf einem Gebiet schon auskennen. Wenn durch Literaturstudium ein grundlegendes Verständnis für einen inhaltlichen Bereich geschaffen wurde, erhöht das die Chancen dafür, dass ein neuer Eindruck – z.B. das Lesen eines neuen Artikels oder eine zufällige Beobachtung – zu einer guten Idee führt. Manchmal hilft es auch, ein Thema, an dem man länger erfolglos gearbeitet hat, einfach beiseite zu legen und eine Zeit lang etwas ganz anderes zu tun. Wirklich interessante Themen

scheinen oft „selbstständig“ weiterzuarbeiten und es kommt vor, dass einem eines Morgens beim Aufwachen plötzlich etwas Interessantes einfällt.

Induktion

Induktiv vorgehen bedeutet häufig, von etwas Besonderem auf etwas Allgemeines oder von Daten auf Theorien zu schließen. Das machen wir im Alltag andauernd. Wir machen die Erfahrung, dass in einem China-Imbiss der Döner nicht besonders gut schmeckt und schließen daraus, dass Chinesen weniger gute Döner produzieren als Türken. Wir gehen also von einer Erfahrung oder Beobachtung aus und verallgemeinern sie oder spinnen sie weiter, wobei wir unser schon vorhandenes Wissen (oder unsere schon vorhandenen Vorurteile) benutzen. Je fundierter dieses Grundwissen ist, desto fundierter werden auch die abgeleiteten Theorien sein. Auch Alltagspsychologie und Vorurteile können eine gute Grundlage für wissenschaftliche Theorien bilden. Solche „Alltagstheorien“ sind meist intuitiv entstanden und müssen nicht falsch sein. Die Problematik mit Alltagstheorien ist nur – wie in *Kapitel 1* beschrieben – die Art und Weise, wie sie überprüft werden. Wenn man beispielsweise die sozialpsychologische Literatur durchforstet, trifft man auf viele Theorien, die einem aus dem Alltag bekannt vorkommen und die durch die Verbindung von Alltagstheorie und der zu einem Thema existierenden Literatur entstanden sein könnten (wie schon erwähnt ist es im Nachhinein äußerst schwierig, herauszufinden, wie eine Theorie zustande kam – manchmal selbst für deren Urheber). Häufig sind jedoch schon vorhandene Theorien die Ausgangsbasis für neue. Dies ist insbesondere der Fall, wenn ein Forscher in einer Studie Ergebnisse erhält, die er mit seiner Theorie nicht richtig erklären kann. Manchmal führen Wissenschaftler, wenn sie eine interessante Idee haben, nicht gleich eine groß angelegte Studie durch, sondern explorieren diese Idee erst einmal in einer oder mehreren kleineren Erkundungs- oder Pilotstudien. Die Ergebnisse aus der Pilotstudie werden dann für die Verbesserung oder Präzisierung der so entstehenden Theorie benutzt.

Metaphern

Eine nicht zu unterschätzende Quelle für Theorien in der Psychologie bilden Metaphern (siehe Gigerenzer, 1991). Damit ist gemeint, dass man einen Mechanismus oder ein Modell aus einem – oft technischen – Bereich als Analogie für die Beschreibung psychischer Prozesse benutzt. Oft scheint ein solcher Analogieschluss selbst den Begründern der entsprechenden Theorien nicht völlig bewusst zu sein. Das vielleicht bekannteste Beispiel ist die Dampfmaschinenmetapher, die Sigmund Freud bei der Beschreibung der Psychoanalyse benutzt hat (das Es ist beispielsweise der Dampf, der im Kessel Druck erzeugt). Eine weitere Quelle für Metaphern ist die Statistik: Es gibt nicht wenige Theorien in der Psychologie – z.B. Kelleys (1967) Attributionstheorie –, die dem statistischen Verfahren der Varianzanalyse (*Kapitel 14* und *15*) ähneln. Auch das am häufigsten benutzte inferenzstatistische Verfahren in der Psychologie, der Signifikanztest (*Kapitel 12*), hat Pate gestanden für eine bekannte Theorie: die sogenannte Signal-Entdeckungstheorie, die zur Beschreibung von Wahrnehmungsvorgängen benutzt wird. Vor allem in der kognitiven Psychologie findet man nicht selten Anklänge an eine weitere Metapher – den Computer. In entsprechenden Theorien (z.B. Anderson, 1993) gibt es ein Arbeitsgedächtnis, das dem Arbeitsspeicher entspricht, und andere Arten von Gedächtnis, die analog zum aktuell ausgeführten Com-

puterprogramm oder zu einem Festplattenspeicher betrachtet werden (siehe auch *Kapitel 31*).

Es ist sicher nie so, dass Metaphern eins zu eins auf die Beschreibung menschlicher Informationsverarbeitungsprozesse angewandt werden, aber die Ähnlichkeit zwischen Theorie und Metapher ist manchmal verblüffend. Meist ist die Beziehung zwischen Metapher und Theorie keine Einbahnstraße. Die Konzeption technischer Systeme, wie etwa des Computers, sind auch durch Vorstellungen darüber, wie Menschen Informationen verarbeiten, beeinflusst worden. Es spricht auch in keiner Weise etwas gegen das Benutzen von Metaphern. Alle Hilfsmittel, die zu einer guten oder einer noch besseren Theorie führen, sollten genutzt werden.

2.4.2 Die systematische Suche nach Theorien

Wie wir zu Beginn dieses Kapitels schon angedeutet haben, werden die sogenannten qualitativen Methoden in der Psychologie bislang kaum angewandt (mit Ausnahme von Forschern in Großbritannien und dessen ehemaligen Kolonien). Manche dieser Methoden sind speziell für das Erstellen von Theorien entwickelt worden. Dabei sind die Unterschiede zwischen positivistischen und konstruktivistischen Ansätzen – beide Positionen findet man bei qualitativ arbeitenden Forschern, oft sogar als Mischformen – nicht so groß, wie man meinen könnte. Qualitative Verfahren zielen darauf ab, aus der Analyse von Daten einen größeren Sinnzusammenhang herzustellen. Ausgangsdaten sind in der Regel Beobachtungen und mündliche oder schriftliche Äußerungen. Die größte Schwierigkeit für den Forscher besteht nun darin, diese Daten nicht sofort auf der Grundlage des eigenen Erfahrungs- und Wissenshintergrundes zu interpretieren, sondern gewissermaßen einen Schritt zurückzutreten und – in der Sprache der Phänomenologie ausgedrückt – zu versuchen, zu den „Dingen selbst“ zu gelangen. Wie man nun von den Daten zu einer Theorie kommt, darin unterscheiden sich die einzelnen Ansätze. Immer wird jedoch versucht, diesen Weg auf eine systematische Weise durch die Vorgabe von Arbeitsschritten zu gehen. Man könnte also von einer „geleiteten Induktion“ sprechen.

Das vielleicht bekannteste Verfahren zur Theoriegewinnung ist die von Glaser und Strauss (2005/1967) entwickelte *Grounded Theory*. Theorien sollen demnach immer auf Daten gegründet (daher „grounded“) sein. In der *Grounded Theory* versucht man, über sukzessives Kodieren von – meistens aus Interviews gewonnenen – Texten zu immer abstrakteren Kategorien zu kommen, deren systematische Verbindung dann letztlich zu einer Theorie führen soll. Sehr wichtig ist bei diesem Verfahren ein permanenter Rückbezug auf die Daten, die tatsächlichen Äußerungen. Um eine höhere Verallgemeinerbarkeit der so entstandenen Theorien zu erreichen, werden dabei „Informanten“ danach ausgewählt, wie hoch ihr Potenzial zu einer Verbesserung der gegenwärtig vorliegenden Fassung einer Theorie ist. Wenn man beispielsweise Zweifel hat, ob die Ideen, die man nach der Befragung von einigen Akademikern gewonnen hat, wirklich für die gesamte Bevölkerung zutreffen, wird man vielleicht im nächsten Schritt ein Interview mit einem Arbeiter durchführen. In der Psychologie gibt es bislang wenige Anwendungsbeispiele für die *Grounded Theory*, wohingegen dieser Ansatz in den Sozialwissenschaften relativ verbreitet ist.

2.5 Von Theorien zu Hypothesen

Wenn eine Theorie – egal, wie sie entstanden ist – überprüft werden soll, dann erfolgt das in der Regel in einer bestimmten Abfolge von Maßnahmen. Diese Abfolge haben Sie schon in *Kapitel 1* kennengelernt: die wissenschaftliche Methode. Dort haben wir diese Methode in einer relativ abstrakten Weise und nur sehr kurz vorgestellt. Hier sollen Teile daraus etwas ausführlicher und anhand von Beispielen erläutert werden. Zunächst möchten wir aber noch kurz reflektieren, was in der wissenschaftlichen Psychologie unter „Theorie" verstanden wird (diese Ausführungen gelten weitgehend auch für die Sozialwissenschaften).

2.5.1 Wie sehen Theorien in der Psychologie aus?

Ob man nun darüber glücklich ist oder nicht: So umfassende Theorien wie in den Naturwissenschaften gibt es in der Psychologie noch nicht. Zwei Bemühungen in diese Richtung sind die Evolutionspsychologie und der Informationsverarbeitungs-Ansatz im weitesten Sinne. Der evolutionspsychologische Ansatz knüpft an die Evolutionstheorie an und versucht, im Prinzip alle Aspekte menschlichen Verhaltens und Erlebens als Resultat evolutionärer Variations- und Selektionsprozesse zu erklären. Evolutionspsychologen gehen davon aus, dass Verhalten nicht durch bereichsübergreifende Mechanismen (z.B. generelle Lernmechanismen) erklärt werden kann, sondern sie postulieren mehrere unterschiedliche bereichsspezifische Mechanismen (z.B. einen Mechanismus dafür, wie man Betrüger entdeckt, und einen anderen dafür, wie man entdeckt, ob andere Menschen mit einem selbst verwandt sind). Informationsverarbeitende Ansätze gehen dagegen meistens davon aus, dass solche allgemeinen, bereichsunabhängigen Mechanismen der Informationsverarbeitung existieren, und integrieren auch emotionale und motivationale Prozesse. Die erfolgreichsten Ansätze dieser Art münden meist in Computermodelle (siehe *Kapitel 31*). Solche Computermodelle sind notwendigerweise sehr präzise Theorien, die versuchen, traditionell separat behandelte Inhalte wie Gedächtnis, Lernen, Motivation, Emotion, Aktion und zunehmend auch soziale Prozesse in ein gemeinsames theoretisches Modell zu integrieren.

Bisher haben wir über Theorien gesprochen, als wäre klar, wovon wir reden. Tatsächlich ist es so, dass, wenn man in der Psychologie von einer Theorie spricht, sehr unterschiedliche Dinge damit gemeint sein können. Manchmal sind Theorien äußerst komplex und umfangreich, wie etwa die Evolutionspsychologie oder der Informationsverarbeitungs-Ansatz im weiteren Sinne. Manche Forscher würden hier eher von Theoriensystemen sprechen. Das andere Ende der Komplexität wird eingenommen von Theorien, die nur eine isolierte Aussage über einen eng umgrenzten inhaltlichen Bereich machen. Solche Theorien, die man getrost auch Hypothesen nennen könnte, finden sich besonders häufig in der Sozialpsychologie (siehe Vallacher & Novak, 1994). Die meisten Theorien in der Psychologie haben – zumindest nachdem sich die Forschung einige Zeit mit ihnen beschäftigt hat – einen mittleren Komplexitätsgrad, was bedeutet, dass man sie in der Regel nicht direkt prüfen kann. Prüfbar sind aber die aus einer Theorie abgeleiteten Hypothesen: relativ präzise, aber eingegrenzte Vorhersagen. Die Überprüfung der Hypothesen gibt dann wieder Aufschluss über die Güte der jeweiligen Theorie.

2.5.2 Von der Theorie zur Hypothesenprüfung: Grundlegende Vorgehensweise

Den groben Ablaufplan für die Überprüfung von Theorien haben Sie schon in *Kapitel 1* kennengelernt: die wissenschaftliche Methode, abgebildet als „Trichtermodell" (*Abbildung 1.6*). Hier sehen wir uns die ersten drei Bestandteile des Trichtermodells, „Theorie", „Forschungshypothese" und „Präzisierung der Hypothese", noch einmal etwas genauer an und vernachlässigen für den Moment weitgehend den Rest des Trichters.

Theorie

Zur Illustration zunächst ein fiktives Beispiel: Angenommen es existiert eine Theorie über die nachhaltigen Auswirkungen der Radon-Strahlung aus dem Erzgebirge auf die Bewohner Sachsens. Ohne hier auf die hypothetischen Mechanismen genauer einzugehen, postuliert die Theorie unter anderem, dass die Strahlung bei längerem Einwirken zu genetischen Veränderungen führt, die sich auch auf das Gehirn Neugeborener auswirken. Die Veränderungen im Gehirn wiederum sollen mehrere Effekte haben, unter anderem den, dass sich die Intelligenz erhöht.

Forschungshypothese

Eine Hypothese (manchmal als *Forschungshypothese* oder als *wissenschaftliche Hypothese* bezeichnet), die man daraus ableiten könnte, wäre:

Die Sachsen sind intelligenter als die anderen Deutschen.

Manchmal werden Hypothesen auch als Fragen (*Forschungsfragen*) formuliert:

Sind die Sachsen intelligenter als die anderen Deutschen?

Obwohl wir nun nur einen Aspekt der Theorie herausgegriffen und als Hypothese formuliert haben, ist immer noch unklar, wie wir tatsächlich bei der Prüfung vorgehen sollen. Wir müssen die Hypothese also so präzisieren, dass sie empirisch überprüfbar wird. Oder, in anderen Worten, wir müssen die Hypothese *operationalisieren*, das heißt, die Operationen angeben, die sie tatsächlich überprüfbar machen.

Präzisierung der Forschungshypothese

Sehen wir uns erst einmal an, was an der Hypothese noch nicht eindeutig formuliert ist. Zunächst ist das der Begriff „Sachsen". Wenn wir wieder auf unsere Theorie zurückgreifen, sollten für diese Hypothese nur Personen als Sachsen gelten, deren Eltern der Radonstrahlung ausgesetzt waren, also eine bestimmte Zeit vor der Geburt ihres Kindes in Sachsen gelebt haben. Als Minimalforderung könnte gelten, dass die Eltern sich bei der Geburt in Sachsen aufgehalten haben, aber man könnte auch fordern, dass die Eltern beispielsweise mindestens zehn Jahre vor der Geburt kontinuierlich in Sachsen gelebt haben, damit das Kind als „Sachse" im Sinne der Hypothese klassifiziert werden kann. Zu überlegen wäre auch noch, ob man Altersbeschränkungen einführen sollte. Sollen alle Sachsen, vom Kind bis zum Greis untersucht werden? Oder soll man sich vielleicht auf alle erwachsenen Sachsen (etwa ab 18 Jahren) beschränken? Bei diesen Überlegungen wird schon deutlich, dass sich aus der Forschungshypothese eine Vielzahl präzisierter Hypothesen – manchmal als *empirische*

Hypothesen bezeichnet – ableiten lassen, je nachdem für welche Operationalisierung man sich entscheidet. Der zweite Bestandteil der Hypothese, den man sich genauer ansehen muss – das „intelligenter" –, erfordert gleich zwei Operationalisierungen. Zunächst muss man bestimmen, was Intelligenz ist. Machen wir es uns diesmal einfach: Das Ausmaß der Intelligenz wird durch das Ergebnis in einem verbreiteten Intelligenztest bestimmt (da kommen wir aber nicht umhin, uns für einen bestimmten Test zu entscheiden). Nun haben wir auch noch den Komparativ zu berücksichtigen: Wann ist jemand intelligent*er* als jemand anders? Hier liegt entweder ein Vergleich mit bundesdeutschen Durchschnittswerten nahe (bei den meisten Intelligenztests ein IQ von 100) oder ein direkter Vergleich zwischen „Sachsen" und „anderen Deutschen". Schließlich bleibt noch festzulegen, was wir unter „anderen Deutschen" verstehen. Dabei treten wieder mehr oder weniger die gleichen Probleme auf, die wir bei der Operationalisierung von „Sachsen" hatten. Hier sind einige Beispiele für – durch unterschiedliche, teilweise durchaus anfechtbare Operationalisierungen – präzisierte Hypothesen:

1. Der Anteil von Erwachsenen (ab 18 Jahren), deren Eltern vor ihrer Geburt mindestens ein Jahr in Sachsen gelebt haben und deren IQ-Wert (im Test X) über dem Durchschnittswert (100) liegt, ist um mindestens 5% höher als der entsprechende Anteil am Bundesdurchschnitt (= 50%).
2. Erwachsene (ab 18 Jahren), deren Eltern vor ihrer Geburt mindestens ein Jahr in Sachsen gelebt haben, erzielen im Durchschnitt mindestens 5 IQ-Punkte (im Test X) über dem bundesdeutschen Durchschnittswert.
3. Erwachsene (ab 18 Jahren), deren Eltern vor ihrer Geburt mindestens ein Jahr in Sachsen gelebt haben, erzielen im Durchschnitt mindestens 5 IQ-Punkte (im Test X) mehr als Erwachsene (ab 18 Jahren), deren Eltern nie in Sachsen gewohnt haben.

Statistische Hypothesen

Oft werden solche präzisierten Hypothesen noch weiter im Hinblick auf Werte, die man erwartet, formalisiert. Die Hypothesen in der psychologischen Forschung beziehen sich meist auf aggregierte (zusammengefasste) Werte wie Mittelwerte oder Anteile und sie beziehen sich fast immer auf die Population (hier: alle Sachsen), nicht nur auf die Personen in der Stichprobe. Solche Werte werden *Populationsparameter* genannt und Hypothesen, die sich darauf beziehen, häufig *statistische Hypothesen* (in *Kapitel 12* erfahren Sie mehr zu statistischen Hypothesen). Populationskennwerte werden in der Statistik meist mit griechischen Buchstaben bezeichnet (z.B. steht π für einen Anteil in der Population und μ für einen Populationsmittelwert). Ausgedrückt als statistische Hypothesen würden die drei obigen Hypothesen so aussehen:

1. $\pi_{Sachsen\ mit\ IQ > 100} \geq 55\%$
2. $\mu_{IQ\text{-}Sachsen} \geq 105$ IQ-Punkte
3. $\mu_{IQ\text{-}Sachsen} \geq \mu_{IQ\text{-}andere\ Deutsche} + 5$ IQ-Punkte

Solche statistischen Hypothesen werden häufig mithilfe von Signifikanztests überprüft (siehe auch hierzu *Kapitel 12*).

Mittlerweile sollte deutlich geworden sein, dass der Weg von der Theorie zur Hypothese zu immer präziseren Aussagen führt, dass aber der Geltungsbereich der immer präziseren Hypothesen immer kleiner wird. Das Beispiel sollte zudem gezeigt haben, dass bei der Ableitung der empirisch überprüfbaren Hypothese auch subjektive Entscheidungen mit im Spiel sind. Die potenziellen Nachteile, die mit der Verengung des Geltungsbereichs einer Hypothese einhergehen, lassen sich jedoch dadurch ausgleichen, dass man aus einer Theorie mehrere empirisch überprüfbare Hypothesen ableitet und der Prüfung aussetzt.

Weitere Vorgehensweise

Wenn die Hypothese präzisiert ist, kann sie im Prinzip einfach überprüft werden. Man führt die in der Hypothese festgelegten Operationen durch, erhebt eine repräsentative Stichprobe – am besten eine Zufallsstichprobe –, prüft, soweit möglich, ob die Annahmen tatsächlich erfüllt sind, und sieht sich die Ergebnisse an. Wenn die Ergebnisse der Hypothese entsprechen, geht diese – und damit auch die zugrunde liegende Theorie – gestärkt aus der Untersuchung hervor (ganz im Sinne des Kritischen Rationalismus: die Hypothese hat einen Falsifikationsversuch erfolgreich überstanden). Wenn die Ergebnisse von den Vorhersagen abweichen, gibt es mehrere Möglichkeiten. Man könnte – wie in der strengen Form des Kritischen Rationalismus – zu dem Schluss kommen, dass die Hypothese – und damit die zugrunde liegende Theorie – falsch ist; das geschieht manchmal nach mehreren abweichenden Ergebnissen, aber fast nie nach einer einzigen Studie. Stattdessen wird meist zunächst untersucht, ob nicht möglicherweise die Operationalisierung der Hypothesen mangelhaft war, oder ob es Gründe bei der Durchführung des Experiments gegeben haben könnte, die zu dem abweichenden Ergebnis geführt haben. Ein möglicher Grund für ein abweichendes Ergebnis könnte sein, dass man zufällig eine atypische Stichprobe gezogen hat. Dies kann durch (mehrfaches) Wiederholen der Studie überprüft werden. Wenn man auch dadurch keine Erklärung für das (konsistent) abweichende Ergebnis finden kann, versucht man häufig – ganz im Sinne des von Kuhn und auch Lakatos beschriebenen Umgangs mit Anomalien – die Theorien so abzuändern, dass sie mit dem Ergebnis konform sind. Solche Modifikationen sind dann wieder die Ausgangsbasis für erneute, modifizierte Hypothesen.

2.5.3 Von der Theorie zur Hypothesenprüfung: Beispiele

Nach dem hypothetischen Beispiel wollen wir zur Illustration zwei bekannte Experimente vorstellen, die tatsächlich durchgeführt worden sind (beide in vielen Varianten). Die Theorien werden dabei jeweils nur kurz dargestellt.

Konformitätsdruck in Gruppen (Asch, 1955)

Stellen Sie sich vor, Sie sind mit einer Gruppe von Leuten zusammen. Sie diskutieren über ein Thema und bald wird Ihnen klar, dass sich die Meinung aller anderen Gruppenmitglieder von Ihrer Meinung unterscheidet. Später in der Diskussion fragt Sie jemand nach Ihrer Meinung. Was sagen Sie? Das ist die Ausgangssituation für Aschs Theorie darüber, wie soziale Normen in einer Gruppe Meinungsäußerungen von einzelnen Gruppenmitgliedern beeinflussen können. Asch hatte noch keine sehr elaborierte Theorie zur Verfügung, aber einige sehr interessante Ideen. Eine von meh-

reren Forschungshypothesen, die er daraus ableitete, kann folgendermaßen zusammengefasst werden:

> *Wenn die Wahrnehmung eines Gruppenmitglieds von der Gruppenwahrnehmung abweicht, tendiert dieses Individuum dazu, sich wider besseres Wissen der Gruppenmeinung anzuschließen.*

Eine Operationalisierung der Forschungshypothese ist in ▶ Abbildung 2.6 illustriert. Die Wahrnehmungsaufgabe bestand darin, festzustellen, welcher Vergleichsreiz dieselbe Länge hatte wie der Standardreiz. Die „Gruppenwahrnehmung" wurde dadurch operationalisiert, dass der Versuchsteilnehmer zunächst die Meinung der anderen Gruppenmitglieder erfuhr, bevor er selbst gefragt wurde. Er saß dabei mit sieben vermeintlichen anderen Versuchsteilnehmern – die aber tatsächlich Mitarbeiter des Versuchsleiters waren – an einem Tisch und jeder dieser anderen „Versuchsteilnehmer" gab dieselbe Einschätzung ab. Manchmal war diese „Gruppenwahrnehmung" richtig (in ▶ Abbildung 2.6: „Der Standardreiz hat dieselbe Länge wie der Vergleichsreiz B"), manchmal war sie falsch (z.B.: „Der Standardreiz hat dieselbe Länge wie Vergleichsreiz C"). Interessant ist nun zu überprüfen, ob sich die Urteile des Individuums in Abhängigkeit von der Gruppenwahrnehmung (richtig vs. falsch) ändern.

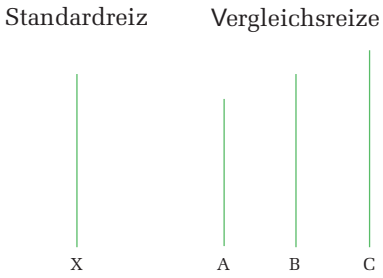


Abbildung 2.6: Versuchsmaterial im Experiment von Asch (nach Asch, 1955).

Tatsächlich schlossen sich 75% aller Versuchsteilnehmer in der experimentellen Bedingung (Abweichung zwischen Mehrheitsmeinung und tatsächlichem Sachverhalt) in mindestens einem von mehreren Durchgängen der Mehrheitsmeinung an. Die statistische Hypothese ließe sich in diesem Fall als eine Hypothese über Populationsanteile (Anteil der Personen, die sich einer abweichenden Gruppenmeinung anschließen) formulieren.

Die Entstehung von Emotionen (Schachter & Singer, 1962)

Ein zentraler Punkt der Emotionstheorie von Schachter und Singer (1962) ist, dass für die Entstehung von Emotionen zwei Faktoren zusammenwirken müssen: Zum einen muss ein physiologischer Erregungszustand vorhanden sein und zum anderen eine Kognition, z.B. die Interpretation einer Situation, die dann die physiologische Erregung zu einer bestimmten Emotion werden lässt. Die spezifische Emotion entsteht also dadurch, dass die Erregung durch die gerade verfügbare Kognition „interpretiert" wird. Wenn eine andere Erklärung für einen Erregungszustand vorliegt (z.B. körperliche Anstrengung), entsteht keine Emotion. Eine aus diesen Annahmen ableitbare Forschungshypothese ist:

Wenn ein physiologischer Erregungszustand nicht erklärt werden kann, entscheidet die Interpretation des sozialen Kontexts über die Art der Emotion.

In mehreren Experimenten operationalisierten die Forscher den physiologischen Erregungszustand ihrer Versuchsteilnehmer durch die Injektion von Adrenalin (erhöhter Erregungszustand) oder Kochsalzlösung (Placebo). Konzentrieren wir uns auf eines der Experimente, in denen alle Teilnehmer eine Adrenalininjektion erhielten (▶ Abbildung 2.7). Nur eine von zwei Teilnehmergruppen erhielt eine Erklärung für den Erregungszustand (Auswirkung der Injektion), die andere bekam die falsche Information, dass es sich bei dem Mittel um ein mildes und harmloses Vitaminpräparat handl[9]e. Eine Gruppe erwartete also körperliche Symptome wie Herzklopfen, Zittern oder schnelleren Atem, konnte sich also den Erregungszustand erklären, die andere Gruppe aber nicht. Die Interpretation des sozialen Kontextes wurde durch einen Vertrauten des Versuchsleiters manipuliert, der in der einen Bedingung wütendes Verhalten zeigte („Wütend" in ▶ Abbildung 2.7) oder – in einer anderen Bedingung – herumalberte („Fröhlich" in ▶ Abbildung 2.7).

Tatsächlich gaben die Teilnehmer in der Gruppe, die ihren Erregungszustand nicht erklären konnte, eine deutlich höhere Intensität ihrer Emotionen in die erwartete Richtung an (ärgerlicher in der „Wütend"-Bedingung und fröhlicher in der anderen Bedingung) als die Teilnehmer, die sich ihren Erregungszustand durch die Wirkung des Medikaments erklären konnten. Die statistische Hypothese bezog sich in dieser Studie auf (Populations-)Unterschiede in den Mittelwerten in einem Befindlichkeitsfragebogen.

		ERWARTUNGEN DES VERSUCHSTEILNEHMERS	
		Erwartet Symptome	Erwartet keine Symptome
VERHALTEN DES „KOMPLIZEN"	Wütend	Versuchsteilnehmer wird nicht durch „Komplizen" beeinflusst	Versuchsteilnehmer wird ärgerlich
	Fröhlich	Versuchsteilnehmer wird nicht durch „Komplizen" beeinflusst	Versuchsteilnehmer wird fröhlich

Abbildung 2.7: Bedingung in einer der Versuchsvarianten von Schachter und Singer (1962). In den Zellen stehen die erwarteten (und tatsächlichen) Ergebnisse.

9 Diese Manipulation könnte heutzutage nicht mehr durchgeführt werden, da sie gegen mittlerweile fest etablierte ethische Grundsätze aller psychologischen Verbände verstößt.

2.5.4 Hypothesenprüfung und Wissenschaftstheorie

Wie wir in den vorigen Abschnitten gesehen haben, sind Theorien und daraus abgeleitete Hypothesen untrennbar miteinander verbunden. In der Überschrift zu diesem Kapitel wird jedoch angedeutet, dass auch die Wissenschaftstheorie etwas mit Theorien und Hypothesen zu tun haben muss. Und das hat sie in der Tat. Die wissenschaftliche Methode liefert im Prinzip auch ohne Rekurs auf eine spezifische Form von Wissenschaftstheorie sinnvolle Ergebnisse, sieht man einmal ab von extremen konstruktivistischen Ansätzen. Aber wie man bestimmt, was eine Theorie ist, was man im Prinzip überprüfen kann und welche Schlussfolgerungen man aufgrund eines empirischen Ergebnisses für die getestete Theorie ziehen kann, das hängt zumindest von einer impliziten Wissenschaftstheorie ab. So gut wie alle wissenschaftstheoretischen Ansätze haben das Postulat des Logischen Empirismus übernommen, dass eine Theorie so präzise wie möglich formuliert werden soll. In der Frage, was Gegenstand der Wissenschaft sein kann, gibt es Divergenzen. Diese hängen teilweise davon ab, ob man – wie die große Mehrzahl aller Forscher – annimmt, dass es eine unabhängig von uns existierende Welt gibt oder nicht. Zentral für die Interpretation von Ergebnissen in Bezug auf eine Theorie ist auch, ob man annimmt, dass eine Theorie verifiziert werden kann oder nicht. Mittlerweile teilt die Mehrheit der Wissenschaftler das zentrale Postulat des Kritischen Rationalismus, dass eine Theorie zwar falsifizierbar, aber nicht verifizierbar ist.

Die Wissenschaftstheorie dient dazu, dem Forscher klarzumachen, welche Annahmen er in seinem Erkenntnisbestreben voraussetzt und welche Schlussfolgerungen aufgrund dieser Annahmen möglich sind. Sie dient auch dazu, Kriterien für die Wissenschaftlichkeit von Theorien zu definieren und sich klar zu werden über Möglichkeiten und Grenzen wissenschaftlicher Erkenntnis.

ZUSAMMENFASSUNG

Grundannahmen, die ein Wissenschaftler treffen muss, auch wenn er sich ihrer nicht immer bewusst ist, beziehen sich auf den Forschungsgegenstand (Was ist die Welt?) und die generelle Zugriffsmethode darauf (Wie können wir sie erkennen?). Diesen Fragen widmet sich die Wissenschaftstheorie. Die meisten wissenschaftstheoretischen Ansätze, die in Psychologie und Sozialwissenschaften eine Rolle spielen, gehen davon aus, dass es eine unabhängig von uns existierende Welt gibt, geben aber teilweise unterschiedliche Antworten darauf, wie wir sie erkennen können, und behandeln auch unterschiedliche Aspekte des Zugriffs auf die Wirklichkeit.

Sehr stark vereinfacht gesagt, befasst sich der *Logische Empirismus* (Carnap u.a.) damit, wie Theorien aussehen sollten, der *Kritische Rationalismus* (Popper) und die *Methodologie wissenschaftlicher Forschungsprogramme* (Lakatos) damit, wie man sie überprüfen sollte, und die *historisch-soziologische Analyse* (Kuhn) damit, was in der Wissenschaft tatsächlich geschieht. Neben diesen vier in Psychologie und Sozialwissenschaften wohl bedeutsamsten konventionellen Ansätzen gibt es auch vereinzelte Bestrebungen, die psychologische Forschung auf eine konstruktivistische Grundlage zu stellen, dort im Vergleich mit den Sozialwissenschaften bislang allerdings mit wenig Erfolg.

Eine Besonderheit der Psychologie ist das (immer noch nicht umfassend gelöste) Leib-Seele-Problem, also die Beziehung zwischen physischen und mentalen Zuständen. Eine wichtige Rolle spielen zudem, auch in den Sozialwissenschaften, nicht beobachtbare Variablen wie etwa „Intelligenz" oder „Gedächtnis" und die reaktive Rolle des „Untersuchungsgegenstands": Der Erforschte kann im Prinzip selbst auch wieder Forscher sein. Diese Besonderheiten unterscheiden Psychologie und Sozialwissenschaften von den Naturwissenschaften und haben Auswirkungen auf die Forschungsmethodik.

Wie Theorien entstehen, ist im Wesentlichen nach wie vor unklar, darüber wie sie überprüft werden sollen, stimmen jedoch die meisten Wissenschaftler überein. Aus Theorien werden Hypothesen abgeleitet, die so präzisiert werden, dass sie empirisch überprüfbar sind. Die Ergebnisse der Hypothesenprüfung werden benutzt, um Rückschlüsse auf die Güte der Theorie zu ziehen, diese zu modifizieren oder, wenn starke negative Evidenz vorhanden ist, sie zu verwerfen. Es besteht ein allgemeiner Konsens darüber, dass Theorien nie bewiesen oder verifiziert werden können, sondern sich nur dadurch in ihrer Güte verbessern, dass sie Falsifizierungsversuche unbeschadet überstehen.

ZUSAMMENFASSUNG

Weiterführende Literatur

Bunge, M. & Ardila, R. (1990). *Philosophie der Psychologie.* Tübingen: Mohr.

Ein Buch, das von einem Psychologen (Ardila) und einem Physiker, der zum Philosophen wurde (Bunge), verfasst wurde. Wissenschaftstheoretische Aspekte der Psychologie werden umfassend beschrieben. Die Autoren sympathisieren stark mit einer materialistischen Ansicht (alles Verhalten und Erleben ist auf Gehirnprozesse reduzierbar).

Dienes, Z. (2008). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Palgrave Macmillan.

Ein Buch, das nicht nur eine kluge Einführung in die wissenschaftstheoretischen Ansätze von Popper, Kuhn und Lakatos gibt, sondern auch die grundlegenden Ideen der statistischen Verfahren erläutert, die auch wir später behandeln werden. Der enge Bezug zwischen Wissenschaftstheorie und Statistik gibt dem Buch ein „Alleinstellungsmerkmal".

Gadenne, V. (2004). *Philosophie der Psychologie.* Bern: Huber.

Ein lesenswertes Buch von einem Psychologen, der einen Lehrstuhl für Philosophie und Wissenschaftstheorie innehat, mit einem Schwerpunkt darauf, philosophisches Denken für Psychologen verständlich zu machen. Es enthält auch ein eigenes Kapitel zum konstruktivistischen Ansatz.

Huber, O. (2005). *Das psychologische Experiment* (4. Aufl.). Bern: Huber.

Elementare und – u.a. durch die vom Autor selbst gezeichneten witzigen Karikaturen – sehr aufgelockerte und gut lesbare Einführung in das psychologische Experimentieren.

Smith, J. A. (Ed.). (2003). *Qualitative psychology: A practical guide to research methods.* London: Sage.

Einführung in verschiedene Ansätze qualitativer Methoden mit Beschreibung der entsprechenden wissenschaftstheoretischen (konstruktivistischen) Hintergründe (vor allem Kapitel 2 und 8).

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik: Ein Lehrbuch zur psychologischen Methodenlehre.* Göttingen: Huber.

Geschrieben von einem methodisch und allgemeinpsychologisch ausgerichteten Psychologen. Es behandelt die in diesem Kapitel angesprochenen Themen etwas ausführlicher, gibt einen guten Überblick über alle wichtigen Aspekte der Wissenschaftstheorie für Psychologen und geht auch ausführlich auf das Verhältnis zwischen Theorien und Hypothesen ein.

Für eine schnelle Orientierung zu wissenschaftstheoretischen Ansätzen siehe auch die entsprechenden Einträge in der Internetdatenbank Wikipedia (*http://de.wikipedia.org/wiki/*).

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.

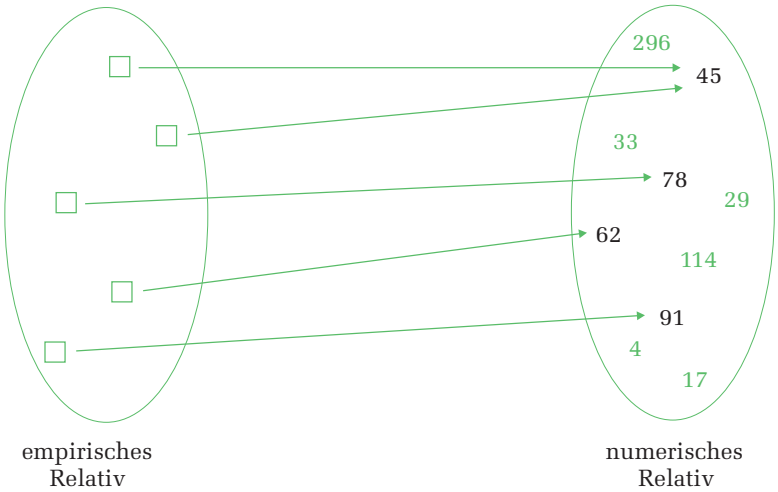


Abbildung 3.1: Abbildung eines empirischen Relativs in ein numerisches Relativ beim Messen.

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 schwarz, 2 = weiblich und weiß, 3 = weiblich und schwarz. 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 Schwarzen 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 Schwarzen, 20 von weiblichen Weißen und 10 von weiblichen Schwarzen 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 Schwarze vertauschen würden (0 = männlich und schwarz, 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 durch ordinalskalierte Messwerte zum Ausdruck gebracht. Ordinalskalierte Daten

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

erlauben somit noch keine Aussage über die Größe des Unterschieds zwischen zwei Messobjekten.

Ein Beispiel sind die Download-Charts, mit denen der Verkaufserfolg von Musiktiteln auf Ordinalskalenniveau gemessen wird. Dabei wird dem Titel 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 Ranginformationen.

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.

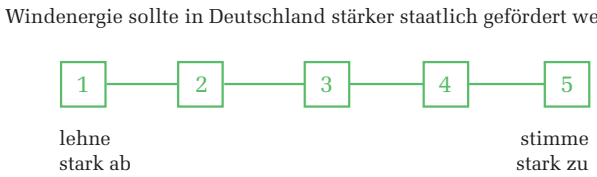


Abbildung 3.2: Ein Beispiel für eine Rating-Skala.

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 Tempe-

ratur 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

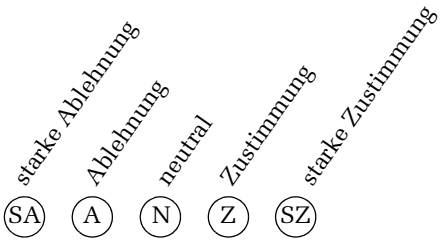
- Item 1:

  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, et al. 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 Inter-

pretationen 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 Testergebnisse 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:

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