1 Experimente und kausale Theorien[1]

Michael R. Waldmann

Kurzbeschreibung
Die Notwendigkeit des Einsatzes formaler Forschungsmethoden wird augenfällig, wenn man sich Studien vergegenwärtigt, die zeigen, wie fehleranfällig und vorurteilsbehaftet unser Denken im Alltag sein kann. Beispiele für solche Befunde werden deshalb im ersten Teil des Beitrags vorgestellt. Im Hauptteil geht es dann um die Beziehung zwischen Theorien und experimenteller Methodik. Die Grundthese des Artikels lautet, dass es psychologische Forschung in der Regel mit der Prüfung kausaler Theorien zu tun hat. Diesem Ziel sind andere Gesichtspunkte wie etwa Repräsentativität untergeordnet. Forscher versuchen vielmehr, natürliche Situationen im Labor so zu verändern, dass sich die empirischen Befunde möglichst eindeutig auf die vermuteten kausalen Beziehungen zurückführen lassen. Anhand einer Reihe von Beispielen wird gezeigt werden, wie man die kausale Interpretierbarkeit der Untersuchungsergebnisse sicherstellen kann und welche Vorannahmen man dabei machen muss. Schließlich wird diskutiert, wie man durch gezielte Überprüfung der Generalität von Theorien zu verbesserten Theorien mit erweitertem Geltungsbereich kommen kann.

Stichwörter: Experimentelle Designs, Kausalität, Konfundierung, theoretische Konzepte, Interaktionen

1.1 Einleitung

Studierende empirischer Wissenschaften wie Psychologie beklagen sich in den ersten Semestern häufig über das Übergewicht von Veranstaltungen zur Methodenlehre. Selbst wenn einmal inhaltliche Theorien zu einem psychologischen Phänomen zur Sprache kommen, dann scheinen auf den ersten Blick Debatten und konkurrierende Ansätze eher das Feld zu bestimmen als gesicherte Erkenntnisse oder wenigstens Einigkeit unter den Wissenschaftlern. Ursprünglich hatte man sich für das Fach doch interessiert, weil man etwas Neues und Interessantes über die Psyche lernen wollte. Wie kommt es zu diesem Übergewicht der Methodenlehre zu Beginn des Studiums? Schließlich sind eine Reihe berühmter Wissenschaftler wie Sigmund Freud und auch Literaten wie Dostojevski zu tief schürfenden Erkenntnissen über die menschliche Psyche gelangt, ohne sich des Methodenrepertoires der experimentellen Psychologie und der Statistik zu bedienen. Warum kann man nicht einfach durch genaue Beobachtung unserer Mitmenschen psychologische Theorien generieren? Die Antwort

[1] Ich danke W. Hager, Y. Hagmayer, P. Sedlmeier und den Herausgebern für hilfreiche Kommentare.

auf diese Frage lautet nicht, dass dies in jedem Fall unmöglich ist, sondern vielmehr, dass es uns diese Vorgehensweise nicht erlaubt, die Gültigkeit der Theorie rational zu beurteilen.

Im zweiten Abschnitt dieses Beitrags wird eine Reihe von psychologischen Studien zum Denken und Hypothesentesten im Alltag diskutiert, die demonstrieren, dass wir in vielen Fällen große Schwierigkeiten damit haben, Theorien und Beobachtungen in adäquater Weise miteinander in Verbindung zu bringen. Ironischerweise zeigt sich häufig genau das, was wissenschaftlichen Theorien vorgeworfen wird, nämlich eine Verzerrung der Daten und ein Festhalten an Vorannahmen, selbst wenn diese durch Beobachtungen längst widerlegt sind. Selbstverständlich sind auch Wissenschaftler nicht vor Fehlern gefeit. Die explizite Verwendung von nachvollziehbaren Methoden erlaubt es aber, solche Voreingenommenheiten rasch zu erkennen und einer kritischen Prüfung zu unterziehen, während sich unser Denken im Alltag vor Kritik oft dadurch schützt, dass es die Grundlagen seiner Schlussfolgerungen nicht offen legt.

Nachdem in diesem Abschnitt gezeigt wurde, dass unser Denken häufig fehleranfällig ist, gibt der restliche Beitrag eine Einführung in die Frage, wie man durch experimentelle Methoden psychologische Theorien überprüfen kann. Psychologische Theorien sind nahezu immer Theorien über kausale Prozesse. Sie beschreiben, wie Menschen in bestimmten Situationen in systematischer Weise reagieren oder handeln, und spezifizieren die mentalen Prozesse, die den beobachtbaren Handlungen zugrunde liegen. Ziel dieses Teils des Beitrags ist es zu zeigen, dass statistische Methoden alleine nicht ausreichen, um kausale Hypothesen zu überprüfen. Statistische Methoden müssen vielmehr in systematischer Weise mit theoretisch begründeten experimentellen Designs gepaart werden, um Schlussfolgerungen auf kausale Vorgänge zuzulassen. Eines der Hauptziele des vorliegenden Beitrags ist es zu zeigen, welche Rolle kausale Theorien bei der Auswahl angemessener experimenteller Designs spielen.

1.2 Quasi-wissenschaftliches Denken im Alltag

Obwohl wir Wissenschaft primär mit akademischen Kontexten in Verbindung bringen, ist es offensichtlich, dass wir auch im Alltag Hypothesen und Theorien bilden. Versuchen wir beispielsweise vorherzusagen, ob ein Kind die Aufnahmeprüfung für das Gymnasium bestehen wird, oder zu erklären, warum eine Freundin an Depressionen leidet, dann nutzen wir Alltagstheorien über psychologische Vorgänge (vgl. Baron, 2000; Hell, Fiedler & Gigerenzer, 1993; Medin, Ross & Markman, 2001; Sternberg & Ben-Zeev, 2001). Manche dieser Theorien schnappen wir in Gesprächen auf oder hören darüber in den Medien. Andere Theorien bilden wir auf der Grundlage von Beobachtungen unserer Mitmenschen. Es stellt sich deshalb die Frage, wie zuverlässig unsere Beobachtungen sind und wie gut wir in der Lage sind, Theorien auf der Basis von Beobachtungen zu evaluieren.

1.2.1 Das Erkennen von Korrelationen

Statistische Zusammenhänge zwischen Ereignissen gehören zu den wichtigsten empirischen Hinweisen auf zugrunde liegende Kausalzusammenhänge. Deshalb ist die Frage interessant, wie gut wir im Alltag darin sind, solche Zusammenhänge zu erkennen. Experimente haben diese Fähigkeit untersucht und eine Reihe interessanter Befunde zu Tage gefördert. In einer frühen Studie hat Smedslund (1963) Krankenschwestern gebeten, 100 Karteikarten durchzusehen, die jeweils Informationen darüber enthielten, ob ein bestimmter Patient ein Symptom zeigte oder nicht und ob der Patient an einer Krankheit litt oder nicht.

Tabelle 1.1 *Kontingenztafel zu einem Zusammenhang zwischen einem Symptom und einer fiktiven Krankheit (nach Smedslund, 1963).*

		Krankheit	
		Anwesend	Abwesend
Symptom	Anwesend	37	33
	Abwesend	17	13

Tab. 1.1 zeigt die statistische Struktur, die den Probandinnen vorgelegt wurde. So fand sich etwa bei 37 Patienten, die die Krankheit hatten, auch das Symptom. Die Tabelle zeigt Daten, die im Wesentlichen dafür sprechen, dass zwischen Krankheit und Symptom kein Zusammenhang besteht. Die Krankheit ist in etwa genauso häufig, wenn das Symptom anwesend ist (53%), wie wenn es abwesend ist (57%). Dennoch gaben 85 Prozent der Krankenschwestern an, dass ein Zusammenhang zwischen Symptom und Krankheit bestehe. Obgleich solche Verzerrungen nicht unter allen Lernbedingungen zu beobachten sind (vgl. Wasserman, 1990), zeigt doch eine große Zahl von Studien, dass es sich um ein allgemeines, häufig zu beobachtendes Phänomen handelt (Arkes & Harkness, 1983; Shaklee & Mims, 1982; Ward & Jenkins, 1965).

Wie kommt es nun zu solchen Fehlurteilen? Eine Erklärung von Smedslunds Befunden ist, dass die Krankenschwestern dazu neigten, bestimmte Informationen zu stark zu gewichten. So findet man häufig, dass der Zelle, in der Symptom und Krankheit beide anwesend sind, besonders starke Beachtung zuteil wird. Die Strategie, nur diese Zelle zu beachten oder sie übermäßig zu gewichten, wird tatsächlich oft beobachtet. Eine andere häufige Strategie ist die zu beurteilen, wie viel Prozent derjenigen, die das Symptom haben, auch krank sind. Diese Strategie berücksichtigt die Einträge in der oberen Zeile („anwesend“), vernachlässigt aber völlig die Informationen in der unteren Zeile („abwesend“), aus der zu entnehmen ist, wie viele Kranke es gibt, die das Symptom nicht haben (vgl. Arkes & Harkness, 1983; Shaklee & Tucker, 1980). Diese Strategie ist unangemessen, weil die Tatsache, dass beispielsweise 80 Prozent der Symptomträger krank sind, keineswegs für eine positive Korrelation sprechen muss. Wenn etwa 100 Prozent aller Personen, die das Symptom

nicht zeigen, ebenfalls krank sind, dann würde dies sogar für eine *negative* Korrelation sprechen.

Ein zweiter Grund, warum wir häufig Korrelationen falsch einschätzen, besteht darin, dass wir dazu neigen, Daten in Richtung Bestätigung unserer Vorannahmen zu verzerren. Diese Frage wurde in einer klassischen Studie von Chapman und Chapman (1969) untersucht. In diesem Experiment wurden den Probanden als Versuchsmaterial Zeichnungen von Personen vorgelegt, die angeblich von Patienten angefertigt wurden. Die Frage, die den Versuchsteilnehmern gestellt wurde, war, ob sich in diesen Zeichnungen Hinweise auf eine psychiatrische Diagnose (Paranoia) finden lassen. Chapman und Chapman konnte zeigen, dass selbst Psychiater Schwierigkeiten hatten, objektiv vorliegende Korrelationen zu erkennen, während sie auf der anderen Seite Korrelationen dort sahen, wo in den Daten in Wirklichkeit keine waren. So gaben die Probanden an, deutliche Zusammenhänge zwischen der Art und Weise, wie die Augen gezeichnet wurden und Paranoia in den Daten zu erkennen, obwohl die objektiven Korrelationen bei Null lagen. Dieses Phänomen wird *illusorische Korrelation* genannt, weil hier Vorannahmen der Versuchsteilnehmer auf die Daten projiziert werden.

Ein weiterer Grund dafür, warum wir Schwierigkeiten insbesondere damit haben, fehlende Zusammenhänge zu sehen, mag darin begründet sein, dass wir zu der Illusion neigen, Kontrolle über unsere Umgebung auszuüben, selbst wenn dies gar nicht der Fall ist. Alloy und Abramson (1979) haben dies in einem Experiment demonstriert, in dem die Probanden hin und wieder auf eine Taste drücken und daraufhin das Verhalten eines Lämpchens beobachten sollten. Es zeigte sich, dass die Versuchsteilnehmer einen Zusammenhang zwischen Tastendruck und Verhalten des Lämpchens sahen, selbst wenn gar keiner vorlag. Sie hatten also die Illusion, Kontrolle auszuüben. Interessanterweise waren lediglich Depressive in der Lage, die objektiven Zusammenhänge korrekt zu erkennen, was in dem Untertitel des Artikels „sadder but wiser“ treffend zum Ausdruck gebracht wurde.

Fairerweise muss man sagen, dass es auch Studien gibt, die belegen, dass wir unter manchen Bedingungen sehr gut in der Lage sind, Korrelationen korrekt zu erkennen und zu nutzen (vgl. Wasserman, 1990; Shanks, Holyoak & Medin, 1996). Wären wir dazu grundsätzlich nicht in der Lage, hätten wir schwerlich überleben können. Dennoch muss man festhalten, dass wir über diese Kompetenz keineswegs stabil in allen Kontexten verfügen, so dass wir uns nicht immer darauf verlassen können. Gerade im Bereich wissenschaftlicher Theorienbildung neigen wir zu Verzerrungen und Fehlurteilen.

1.2.2 Das Testen von Hypothesen

Wissenschaftliches Denken beinhaltet zu einem großen Anteil Prozesse des Hypothesentestens (vgl. Kuhn, 1997). Eine Reihe von Untersuchungen hat sich mit der Frage befasst, welche Strategien wir beim Testen von Hypothesen im Alltag einsetzen. Eine klassische Studie hierzu stammt von Wason (1960). In diesem Experiment wurde Versuchsteilnehmern mitgeteilt, dass die Zahlenreihe 2-4-6 einer bestimmten Regel folge und dass es Aufgabe der Versuchsteilnehmer sei, diese Regel herauszu-

finden. Dabei war es ihnen gestattet, dem Versuchsleiter Sequenzen von drei Zahlen vorzulegen, der dann für jede Sequenz eine Rückmeldung darüber gab, ob sie der Regel folgte oder nicht. Diese Situation war in Analogie zu wissenschaftlichen Studien konstruiert, wobei die Regelhypothesen Theorien und die vorgelegten Zahlensequenzen Experimenten entsprechen sollten. Ein typisches Ergebnis war, dass die Versuchsteilnehmer die Regel „um 2 ansteigende Zahlen" bildeten und dem Versuchsleiter Sequenzen wie „8-10-12", „3-5-7" oder „254-256-258" vorlegten. Bemerkenswert dabei ist, dass es sich durchweg um Beispiele handelte, die der vermuteten Regel folgten, nicht um solche wie „8-9-10", die sie widerlegten. Auf diese Weise war es den Versuchsteilnehmern nicht möglich, andere mit der eingangs vorgelegten Beispielssequenz verträgliche Regeln wie „ansteigende Zahlen" (z.B. 11-12-27) zu entdecken, weil ihre eigene Regel nur Sequenzen zuließ, die gleichzeitig auch dieser allgemeineren Regel folgten. Die Tendenz, nur bestätigende Instanzen auszuwählen, wurde *Bestätigungsbias* („confirmation bias") genannt, weil sie der etwa von Popper (1994) vorgeschlagenen Regel widersprach, Hypothesen dadurch zu testen, dass man sie zu falsifizieren sucht.

Andere Studien haben die Tendenz, Eingangshypothesen zu bestätigen, auch mit realistischerem Material nachgewiesen. Mynatt, Doherty und Tweney (1977) präsentierten Versuchsteilnehmern fiktive Partikel auf einem Computerbildschirm, deren Bewegung die Probanden vorhersagen sollten. Das Display war vergleichsweise komplex und zeigte eine Vielzahl von Objekten an verschiedenen Orten. Die zugrunde liegende Regel war, dass ein Partikel immer dann stoppte, wenn es in der Nähe eines grauen Objekts war. Wieder wurde bei den Probanden eine starke Tendenz deutlich, eher denjenigen Test zu wählen, der konsistent mit der Eingangshypothese war, als den, der sie widerlegen konnte (vgl. auch Klahr, 2000; Snyder & Swann, 1978).

Andere Untersuchungen zeigen, dass Probanden ohne Methodentraining häufig dazu neigen, Daten zu ignorieren, die ihre Hypothesen widerlegen. Kuhn (1997) beispielsweise ließ Probanden herausfinden, welche Faktoren (Humor in der Sendung, Haarlänge der Mitwirkenden, Tag der Sendung, Musik in der Sendung) die Beliebtheit einer Fernsehsendung beeinflussten. Die meisten Versuchsteilnehmer vermuteten, dass Humor der entscheidende Faktor sei. Die (fiktiven) Daten, die ihnen gezeigt wurden, machten allerdings deutlich, dass es keinen Zusammenhang zwischen diesem Faktor und den Einschätzungen der Beliebtheit gab. Nichtsdestoweniger hielten die Versuchsteilnehmer an ihrer Hypothese fest und nannten auf Nachfrage häufig komplizierte Gründe dafür, warum man in den Daten keine klare Evidenz für ihre Hypothese sehen konnte. Die Plausibilität ihrer Vorannahmen hatte also ein stärkeres Gewicht als die statistischen Beziehungen in den beobachteten Daten.

Ähnlich wie bereits bei den Untersuchungen zum Erkennen von Kovariationen zeigt sich hier also wieder die Tendenz bei den Versuchsteilnehmern, auf Vorannahmen zu beharren, auch wenn die empirische Evidenz dagegen spricht. Ähnlich wie dort gibt es auch in der Literatur zum Hypothesentesten Theorien, die argumentieren, dass unsere Strategien nicht so defizitär sind, wie es die berichteten Studien nahelegen. So konnten beispielsweise Klayman und Ha (1987) zeigen, dass die Strategie, nach Daten zu suchen, die die gebildete Hypothese bestätigen, sich unter bestimmten Bedingungen besser eignet, die richtige Regel zu finden, als andere Strate-

gien. Ist die Regel etwa sehr restriktiv, so dass nur relativ wenige Daten mit ihr verträglich sind (z.B. „aufsteigende Primzahlen"), dann ist es besser, nach bestätigenden Daten zu suchen. Klayman und Ha (1987) bezeichnen dies als „positive test strategy". Diese Strategie führt häufig zu Falsifizierungen der eingangs vermuteten Regel, selbst wenn dies nicht das erklärte Ziel der Person war, die die Hypothese testet (vgl. auch Fischhoff & Beyth-Marom, 1983).

1.2.3 Zusammenfassung

Insgesamt zeigen die berichteten Studien zum quasi-wissenschaftlichen Denken im Alltag, dass wir häufig große Schwierigkeiten damit haben, uns von unseren Vorannahmen und Plausibilitätsüberlegungen zu distanzieren und die Beziehung von empirischer Evidenz zu unseren Theorien angemessen zu interpretieren. Häufig scheinen uns Geschichten, die Zusammenhänge zwischen Ereignissen plausibel machen, ein größeres Gewicht zu haben als objektive Daten. Diese Beobachtungen haben viele Psychologen dazu gebracht, menschliches Denken als irrational und fehleranfällig zu sehen (vgl. Dawes, 2001; Kahneman, Slovic & Tversky, 1982). Diese Sicht ist nicht unumstritten. Andere Forscher haben argumentiert, dass wir in bestimmten Situationen, für die uns unsere Evolutionsgeschichte vorbereitet hat, sehr gut in der Lage sind, uns rational zu verhalten (vgl. Gigerenzer, Todd & ABC Research Group, 1999). Es ist nicht das Ziel dieses Artikels, zwischen diesen Positionen zu entscheiden. Sieht man sich aber die Befunde zu den gegenüber stehenden Positionen insgesamt an, dann bleibt unbestreitbar, dass es viele Bedingungen gibt, in denen wir uns wenig rational verhalten. Da wir uns bei der Beurteilung wissenschaftlicher Hypothesen nicht darauf verlassen können, dass wir uns gerade in einer günstigen Situation befinden, die rationales Urteilen erlaubt, müssen wir uns unserer Fehleranfälligkeit ständig gewärtig sein. Wissenschaftliche Methoden bieten uns die Möglichkeit, in nachvollziehbarer und rationaler Weise Hypothesen zu überprüfen und damit vor möglichen intuitiven Fehlertendenzen zu schützen.

1.3 Kausale Theorien und Experimentalmethodik

Die Grundthese dieses Artikels ist, dass Experimente der Überprüfung von Kausalhypothesen dienen. Dabei soll gezeigt werden, dass die Analyse statistischer Beziehungen zwischen beobachteten Ereignissen alleine nicht ausreicht, um kausale Theorien überprüfen. Nur wenn sich diese Analysen auf Beobachtungen beziehen, die im Rahmen sorgfältig geplanter *experimenteller Designs* gewonnen wurden, dann sind die Daten im Hinblick auf die zu prüfenden kausalen Theorien interpretierbar (vgl. auch Pearl, 2000). Die Planung experimenteller Designs ist also ein integraler Bestandteil der Überprüfung kausaler Theorien, ohne deren Kenntnis die Ergebnisse der statistischen Analysen nicht gedeutet werden können. Dieses Zusammenspiel zwischen Design und Analyse soll im Folgenden rekonstruiert werden. Da es sich um einen einführenden Text handelt, sollen nur die Logik, nicht die formalen Details

dargestellt werden (vgl. Pearl, 2000, für formale Ableitungen). Die Darstellung experimenteller Designs kann ebenfalls nur exemplarisch vorgenommen werden (vgl. für ausführlichere Darstellungen Bordens & Abbott, 1999; Bortz & Döring, 1995; Hager, 1987, 1992; Huber, 1995; Kirk, 1995; Shaughnessy & Zechmeister, 1997; Westermann, 2000).

1.3.1 Kausale Theorien

Am Anfang der Planung von Experimenten steht immer die Spezifikation einer Theorie, die es zu überprüfen gilt. Im einfachsten Fall kann so eine Theorie die kausale Beziehung zwischen zwei beobachtbaren Ereignissen spezifizieren, wie etwa die Beziehung zwischen einem Schalter und einem Licht. Eine solche Beziehung wird graphisch durch einen von der Ursache (Schalter) auf den Effekt (Licht) gerichteten Pfeil symbolisiert. In der Psychologie geht es allerdings selten um solche einfachen Beziehungen, sondern um komplexere Kausaltheorien.

Was man genau unter Kausalität verstehen soll, ist ein in der Psychologie und Philosophie heftig umstrittenes Thema (vgl. Eells, 1991; Pearl, 2000; Waldmann, 1996). In einer ersten Annäherung kann man sagen, dass viele Theorien davon ausgehen, dass kausale Einflüsse durch Mechanismen vermittelt werden. Drückt man etwa auf einen Lichtschalter, dann geht das Licht deshalb an, weil es zu einem Stromfluss kommt. Ähnlich nimmt man etwa in der Kognitionspsychologie an, dass zwischen Situation und Verhalten Stufen der Informationsverarbeitung liegen, die der Grund für systematische Beziehungen sind. Ein weiterer wichtiger Aspekt von Wissen über Kausalität ist, dass es uns in die Lage versetzt zu handeln. Die Manipulation einer Ursache bewirkt den Effekt, während das Setzen eines Effekts nicht seine Ursache zustande bringt. Liegt beispielsweise eine Kausalbeziehung zwischen Rauchen und Lungenkrebs vor, dann bedeutet dies, dass wir eine erhöhte Wahrscheinlichkeit haben, an dieser Erkrankung zu sterben, wenn wir uns für regelmäßiges Rauchen entscheiden. Diese inhärente Gerichtetheit von Kausalität wird graphisch durch die Pfeilrichtung in kausalen Diagrammen symbolisiert.

Der Sachverhalt, dass experimentelle Befunde stets im Hinblick auf eine eingangs aufgestellte Theorie interpretiert werden, bedeutet auch, dass man mit Hilfe von Experimenten nicht automatisch zu richtigen Erkenntnissen kommt. Überprüfen wir eine falsche Theorie, dann können uns die besten Experimente nicht die richtige Theorie liefern. Wir können lediglich nachträglich beurteilen, wie gut oder wie schlecht die experimentellen Befunde mit der Theorie, die wir aufgestellt haben, verträglich sind.

Bisher wissen wir nur wenig darüber, wie Wissenschaftler zu ihren Theorien gelangen. Eine Reihe von Faktoren scheint hier zum Tragen zu kommen. So konnte gezeigt werden, dass Theorien oft auf der Grundlage von Analogien zu anderen Inhaltsbereichen generiert werden (Gigerenzer, 1991; Holyoak & Thagard, 1995; Klahr, 2000; Thagard, 2000). In der modernen Kognitionspsychologie werden beispielsweise häufig Theorien vorgeschlagen, die eine gewisse Analogie zu technischen Entwicklungen wie dem Computer aufweisen. Die vielleicht wichtigste Quelle von Theorien ist die Auseinandersetzung mit anderen Theorien, die in einem Gebiet

bereits postuliert wurden. Phänomene, die durch diese Theorien nicht erklärt werden können, oder konkurrierende Theorien über den gleichen Gegenstandsbereich sind häufig ein wichtiger Ausgangspunkt für die Entwicklung neuer Theorien.

Leider gibt es keine automatisierbare Methode, wie man zu richtigen Theorien gelangt. Man sollte aber versuchen, Theorien formal zu präzisieren, um sie für andere Wissenschaftler nachvollziehbar zu machen und logische Widersprüche zu vermeiden (vgl. Westermann, 2000, für eine Einführung in die logische Analyse). Die Formulierung von Theorien als Computermodelle ist eine wichtige Methode, um formal präzise Theorien zu bilden und Vorhersagen der häufig komplexen Theorien ableiten zu können (vgl. O'Reilly & Munakata, 2000).

1.3.2 Experimentelle Methodik

Das Ziel von experimentellen Designs besteht darin, Bedingungen herzustellen, die eine kausale Interpretation der empirischen Befunde erlauben. Im Folgenden werden einige zentrale Merkmale von Designs diskutiert, um die Beziehung zwischen statistischer Analyse und kausalen Theorien deutlich zu machen.

Manipulation

Die Unmöglichkeit, kausale Hypothesen auf statistische Muster zu reduzieren, lässt sich am klarsten am Beispiel zweier kovariierender Ereignisse veranschaulichen (vgl. Abb. 1.1). Beobachtet man etwa, dass zwei Ereignisse wie Frustration und Aggression miteinander kovariieren, dann kann man daraus nicht ableiten, ob Frustration eine Ursache für Aggression ist oder umgekehrt. (Wir nehmen vereinfachend an, dass es keine weiteren Kausaleinflüsse gibt.) Kovariationen sind anders als Kausalbeziehungen nämlich nicht gerichtet. Zeigt sich etwa, dass Frustrierte mit höherer Wahrscheinlichkeit aggressiv reagieren als Nicht-Frustrierte, dann gilt auch das Umgekehrte, dass Aggressive mit höherer Wahrscheinlichkeit frustriert sind als Nicht-Aggressive (vgl. Waldmann, 1996, 1997). Dieser symmetrische Aspekt von Kovariation wird in Abb. 1.1A durch die gekrümmte Linie wiedergegeben.

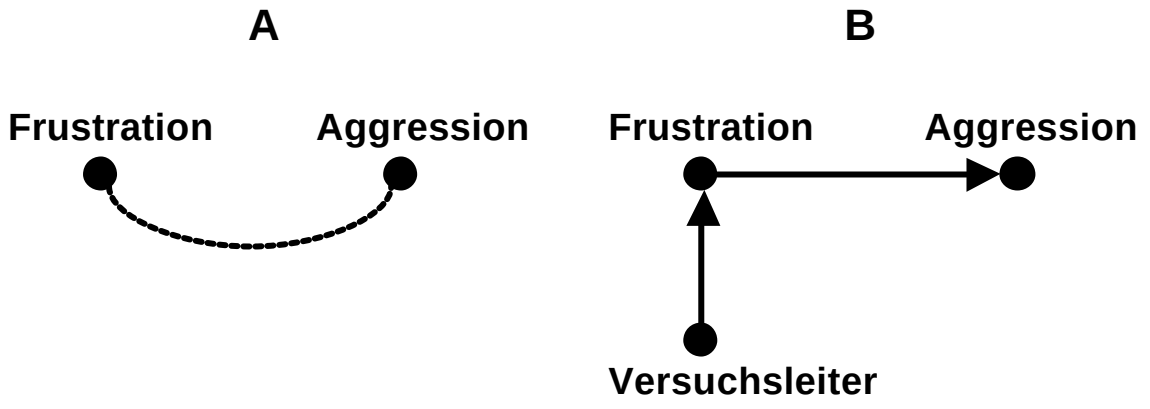

Abbildung 1.1 *Natürliche Beobachtung (A) vs. experimentelle Manipulation (B).*

Wie kann man nun die Kausalrichtung empirisch erkennen? Führt man eine nicht-experimentelle Beobachtungsstudie durch, dann bleibt einem in dieser Situation mit zwei Ereignissen nichts anderes übrig, als auf Plausibilitätsüberlegungen zurückzugreifen. Anders als in dem Beispiel in Abb. 1.1 gibt es auch Ereignisse, bei denen nur eine Richtung plausibel ist. Beobachtet man beispielsweise eine Kovariation zwischen dem Wetter und Angaben zum psychischen Wohlbefinden, dann liegt es auf der Hand, dass das Wetter die Ursache und nicht der Effekt psychischen Wohlbefindens ist. In vielen psychologisch interessanten Fällen ist die Richtung allerdings weniger eindeutig vorab festlegbar.

Im Gegensatz zu Beobachtungsstudien bieten Experimente die Möglichkeit, die Kausalrichtung durch einen Eingriff des Experimentators festzulegen. In echten Experimenten werden bestimmte Ereignisse, die *unabhängigen Variablen* (UV), durch *Manipulation* des Versuchsleiters variiert, während die Auswirkungen der UV auf die *abhängige Variable (AV)* beobachtet werden (Abb. 1.1B). Auf diese Weise lässt sich überprüfen, ob die UV eine Ursache der AV ist. So könnte man in einem Experiment etwa die Hälfte der Versuchsteilnehmer durch eine experimentelle Manipulation frustrieren, um festzustellen, ob dieser Eingriff zu einem höheren Ausmaß an Aggression in der Gruppe der Frustrierten im Vergleich zur Gruppe der Nicht-Frustrierten führt. In dieser Studie kann Frustration nicht der Effekt von Aggression sein, einfach deshalb, weil Frustration durch Außeneinwirkung des Experimentators zustande gekommen ist, der nun die Rolle der Ursache für die UV einnimmt (Abb. 1.1B). Es ist natürlich möglich, dass man Unrecht hatte und dass nicht Frustration die Ursache von Aggression ist, sondern vielmehr Aggression die Ursache von Frustration. Dies sollte sich in einem weiteren Experiment bestätigen lassen, in dem man Aggression als UV manipuliert und Frustration als AV beobachtet. Denkbar ist natürlich auch, dass Kausalbeziehungen in beide Richtungen vorliegen, so dass sich Frustration und Aggression wechselseitig beeinflussen.

Manipulation ist eines der wichtigsten Merkmale von echten Experimenten. Manche Fragestellungen lassen aber keine Manipulation der UV zu. Möchte man etwa die Gedächtnisleistung älterer mit der jüngerer Menschen vergleichen oder die Depressivität Geschiedener im Vergleich zu Verheirateten, dann muss man mit vorgegebenen *natürlichen Gruppen* arbeiten. Die Merkmale, die die natürlichen Gruppen charakterisieren, sind nicht durch Manipulation während des Versuchs herstellbar. Solche Studien nennt man *Quasi-Experimente*, weil hier Schlussfolgerungen über den kausalen Status der untersuchten Variablen stärker von Vorannahmen abhängig sind, als es bei echten Experimenten der Fall ist.

Kovariation

Das wichtigste Ziel statistischer Verfahren bei der Analyse von Experimenten ist die Untersuchung der Frage, ob die unabhängigen und die abhängigen Variablen kovarieren. Wenngleich sich Kovariationen nicht ohne weitere Zusatzvorkehrungen kausal interpretieren lassen, so gilt doch das Umgekehrte, dass das Vorliegen von Kausalbeziehungen zu Kovariationen zwischen den betroffenen Ereignissen führen sollte (vgl. Cheng, 1997). Um Kovariationen in Experimenten beobachten zu können, muss man

mindestens zwei Bedingungen vergleichen, eine Bedingung, in der die UV anwesend ist (Experimentalbedingung), und eine Bedingung, in der sie abwesend ist (Kontrollbedingung). Beobachtet man etwa, dass mehr Frustrierte (Experimentalbedingung) aggressiv sind als Nicht-Frustrierte (Kontrollbedingung) oder dass die Intensität aggressiven Verhaltens in der Experimentalbedingung höher ist als in der Kontrollbedingung, dann liegt Kovariation vor. Beobachtet man hingegen lediglich, dass 80 Prozent der Frustrierten aggressiv sind, dann lässt sich gar nichts über die Kovariation aussagen. Sollte sich nämlich herausstellen, dass ebenfalls 80 Prozent der Nicht-Frustrierten aggressiv sind, dann läge die Kovariation bei Null. Frustration würde das Ausmaß an Aggression gegenüber der Kontrollgruppe nicht steigern. Es könnte sogar sein, dass sich in diesem Fall Frustration hemmend auf Aggression auswirkt, wenn nämlich das Ausmaß von Aggression in der Kontrollgruppe der Nicht-Frustrierten *höher* ausfallen sollte als in der Experimentalgruppe (z.B. 90%). Dann könnte man durch Frustration die Aggressionsneigung reduzieren.

In *Designs mit unabhängigen Versuchsgruppen* vergleicht man mindestens eine Experimental- mit einer Kontrollgruppe, der jeweils verschiedene Versuchsteilnehmer zugewiesen werden. Je nach Fragestellung kann man auch Designs wählen, in denen mehrere Stufen der UV mit einer oder mehreren Kontrollgruppen verglichen werden. So ließen sich beispielsweise mehrere Experimentalbedingungen verwirklichen, in denen man die Intensität der Frustration variiert oder verschiedene Formen von Frustrationserzeugung miteinander vergleicht.

Bei manchen Fragestellungen kann man Experimental- und Kontrollgruppen auch bei den gleichen Versuchsteilnehmern realisieren. In solchen *Messwiederholungsdesigns*[2] nehmen die gleichen Versuchsteilnehmer an verschiedenen Bedingungen teil. So könnte man in unserem Beispiel Versuchsteilnehmer mehrmals an verschiedenen Tagen zur Teilnahme an einem Experiment einladen. An manchen Tagen, die den Experimentalbedingungen entsprechen, werden sie dann beobachtet, nachdem sie vorher frustriert wurden, an anderen Tagen (den Kontrollbedingungen) werden sie nicht frustriert. Man kann dann bei jeder Versuchsperson getrennt untersuchen, welchen Einfluss Frustration auf Aggression hat (vgl. auch das Beitrag von Sedlmeier). Auch hier geht es darum, Kovariationen zwischen der UV und der AV zu erfassen (vgl. Kirk, 1995; Huber, 1995; Shaughnessy & Zechmeister, 1997, für ausführlichere Diskussionen dieser Designs).

Nach Dawes (2001) gehört die Vernachlässigung von Kontrollgruppen zu den wichtigsten Gründen für irrationales Denken im Alltag. Hören wir beispielsweise, dass viele Depressive angeben, in der Kindheit von den Eltern schlecht behandelt worden zu sein, dann tendieren wir fast automatisch dazu, einen Zusammenhang zu sehen. Diese Tendenz wird dadurch unterstützt, dass wir uns eine plausible Geschichte ausdenken können, die diesen Zusammenhang erklärt. Um zu einem rationalen Urteil zu kommen, sollten wir uns aber vielmehr fragen, wie Nicht-

[2] In manchen Texten wird zwischen echten Messwiederholungsdesigns, bei denen die gleiche abhängige Variable zu verschiedenen Zeitpunkten gemessen wird (z.B. bei Längsschnittstudien), und „Within-Subjects"-Designs, bei denen die abhängige Variable bei den gleichen Versuchspersonen unter verschiedenen Bedingungen (d.h. Stufen der UV) untersucht wird, unterschieden. Hier werden der Einfachheit halber beide Designs unter der gemeinsamen Bezeichnung Messwiederholungsdesigns diskutiert.

len Urteil zu kommen, sollten wir uns aber vielmehr fragen, wie Nicht-Depressive von ihren Eltern behandelt wurden und ob es tatsächlich einen Unterschied in der Depressionsneigung bei Personen, die eher gut oder eher schlecht von ihren Eltern behandelt wurden, gibt.

Das Problem der Konfundierung

Ein weiterer Grund, warum beobachtete statistische Kovariationen nicht ohne weiteres Rückschlüsse auf Kausalbeziehungen zulassen, besteht darin, dass man immer mit der Möglichkeit rechnen muss, dass die Kovariationen durch weitere Ereignisse verursacht wurden. Die Kovariation von Fieber und Halsschmerzen bei Erkältungskrankheiten beruht nicht auf einer direkten Kausalbeziehung zwischen diesen Symptomen, sondern ist vielmehr eine Folge einer Virusinfektion, die sowohl Fieber als auch Halsschmerzen verursacht. Abb. 1.2 zeigt den einfachsten Fall einer solchen Situation, bei der die Beziehung zwischen UV und AV durch eine Drittvariable, die konfundierende Variable bzw. Störvariable (KV), verzerrt wird.

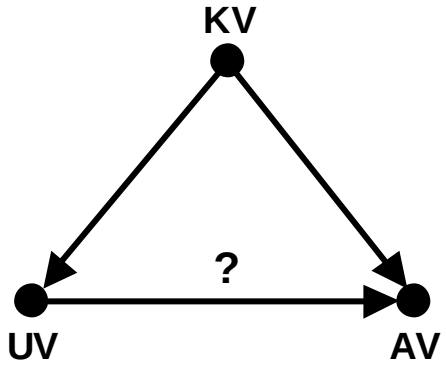

Abbildung 1.2 *Beispiel für ein Kausalmodell, bei dem eine konfundierende Variable (KV) sowohl die unabhängige (UV) als auch die abhängige Variable (AV) kausal beeinflusst.*

Als erstes Beispiel für eine Konfundierung mag wieder die Frustrations-Aggressions-Hypothese dienen. Angenommen, Geschlecht (KV) wäre ein Faktor, der sowohl Frustration (UV) als auch Aggression (AV) kausal beeinflusst. Ein extremer Fall wäre etwa, dass nur Männer frustriert sind, aber nicht Frauen, und dass nur Männer aggressiv sind, aber nicht Frauen. Das Geschlecht der Probanden kovariiert also perfekt sowohl mit der UV als auch mit der AV. Beobachtet man nun im Rahmen einer nicht-experimentellen Studie die Kovariation von Frustration und Aggression, dann zeigt sich auch zwischen diesen beiden Variablen eine perfekte Kovariation. Alle Personen, die frustriert sind, sind auch aggressiv, und alle Personen, die nicht frustriert sind, verhalten sich nicht aggressiv. Beachtet man die Konfundierung mit dem Geschlecht nicht, so sieht dies auf den ersten Blick wie eine eindeutige Bestätigung der Frustrations-Aggressions-Hypothese aus. Tatsächlich handelt es sich aber um

eine Scheinkorrelation, die vollständig durch die beiden deterministischen Kausalpfeile erklärt wird, die von der KV ausgehen. Zwischen UV und AV muss dabei keinerlei kausale Beziehung angenommen werden.

In diesem extremen Beispiel wird die Konfundierung sofort augenfällig, wenn man sich das Geschlecht der Untersuchungsteilnehmer ansieht. Die Brisanz des Konfundierungsproblems wird aber deutlich, wenn man die Möglichkeit in Betracht zieht, dass es konfundierende Faktoren gibt, die nicht so augenfällig sind oder die man noch nicht kennt.

Konfundierungen müssen nicht durch einen einzelnen kausalen Faktor generiert werden, es kann auch zu Scheinkorrelationen aufgrund des Zusammenspiels mehrerer KV in einem komplexen kausalen Netzwerk kommen. Schließlich gibt es auch potentielle Konfundierungen, die dadurch zustande kommen, dass Ursachen (bzw. UVs) Ereignisse sind, die stets durch ein Bündel von Merkmalen beschrieben werden können. Da sich diese Merkmale auf das gleiche Ereignis beziehen, sind sie notwendigerweise miteinander korreliert. Testet man als UV etwa ein Medikament, um dessen Wirksamkeit auf den Gesundungsprozess (AV) zu untersuchen, dann lassen sich unterschiedliche Merkmale der UV isolieren, die für den etwaigen Effekt verantwortlich sein können. So ist es beispielsweise denkbar, dass der Effekt des Medikaments nicht aufgrund seiner Wirksubstanz zustande kommt, sondern deshalb, weil das Wissen bei den Probanden darüber, dass sie ein Medikament einnehmen, selbst bereits einen Effekt ausüben kann (Placebo-Effekt).

Tabelle 1.2 *Beispiel für Simpsons Paradox (vgl. Text).*

	Männer	Frauen	Gesamt
Medikament	0/4 (0%)	16/36 (44%)	16/40 (40%)
Kein Medikament	4/36 (11%)	4/4 (100%)	8/40 (20%)

Konfundierungen können dazu führen, dass die beobachteten Kovariationen das Gegenteil von der tatsächlich zugrunde liegenden kausalen Beziehung ausdrücken. In der methodologischen Literatur sind solche Situationen unter der Bezeichnung *Simpsons Paradox* bekannt geworden, das von Pearson 1899 entdeckt wurde (vgl. Pearl, 2000; Simpson, 1951; Waldmann & Hagmayer, in press). Simpsons Paradox beschreibt eine Situation, in der eine Kovariation, die in der Gesamtgruppe zu beobachten ist, in *jeder* Teilgruppe verschwindet oder gar ins Gegenteil verkehrt wird. Ein Beispiel für eine solche Situation gibt Tab. 1.2. Diese Tabelle beschreibt die Wirkung eines fiktiven Medikaments auf die Gesundungsrate der Probanden. Blickt man auf die Wirkung in der Gesamtgruppe (rechts), dann zeigt sich, dass 40 Prozent der Probanden, die das Medikament einnahmen, gesund wurden, während sich nur 20 Prozent der Probanden in der Kontrollgruppe besser fühlten. Das Medikament scheint also eindeutig eine positive Wirkung auszuüben. Blickt man nun aber auf die Teilgruppen der Männer und Frauen getrennt, dann scheint das Medikament negativ zu wirken. Elf Prozent der Männer und 100 Prozent der Frauen gesundeten ohne

Medikament, während keiner der Männer und nur 44 Prozent der Frauen gesund wurden, die das Medikament eingenommen hatten. Obwohl das Medikament der Gesamtgruppe zu helfen scheint, schadet es eindeutig der Gruppe der Männer und der Frauen, in die sich die Gesamtgruppe vollständig zerlegen lässt.

Wie kommt es nun zu diesem paradoxen Befund? Der Grund ist ein extremer Fall von Konfundierung, wie er in Abb. 1.2 dargestellt ist. Das Geschlecht (KV) ist ein kausal wirksamer Faktor, der sowohl die Medikamenteneinnahme (UV) als auch die Gesundungsrate (AV) beeinflusst. Wie man der Tabelle entnehmen kann, neigen Frauen in der beobachteten Situation eher dazu, das Medikament einzunehmen als Männer, was dazu führt, dass sich mehr Frauen als Männer in der Medikamenten-Bedingung befinden. Außerdem ist der Gesundheitszustand von Frauen unabhängig von der Medikamenteneinnahme besser als der von Männern. Diese beiden statistischen Kovariationen zusammen bewirken eine Umkehrung der Kovariation zwischen Medikament und Gesundungsrate in der Gesamtgruppe gegenüber der wahren negativen Wirkung des Medikaments.

Es ist wichtig zu sehen, dass die beobachteten Muster kein Fehler der Untersuchung sind, sondern dass es sich durchaus um reale Kausalzusammenhänge handeln kann, die man bei freien (nicht-experimentellen) Beobachtungen von Ereignissen vorfinden würde. Dennoch sollte klar sein, dass konfundierende Variablen, insbesondere dann, wenn man sie nicht kennt und somit nicht bei der Analyse der Daten berücksichtigen kann, ein fundamentales Problem für die Prüfung von Kausalhypothesen darstellen. In der Regel werden empirische Befunde durch Konfundierung uninterpretierbar.

Situationen, die Simpsons Paradox entsprechen, sind nicht etwa eine Erfindung von Methodikern, sondern treten auch in realen empirischen Untersuchungen auf. So konnten Appleton, French und Vanderpump (1996) in einer Re-Analyse einer Studie zum Rauchen eine Situation nachweisen, die Simpsons Paradox entspricht. In dieser Studie mit einer Stichprobe von 1314 Frauen aus dem teils städtischen, teils ländlichen Distrikt von Whickham (Newcastle-upon-Tyne) wurde zunächst im Jahr 1972 und dann 20 Jahre später 1992 eine Erhebung durchgeführt. Überraschenderweise stellte sich heraus, dass die Raucherinnen eine längere Lebenserwartung hatten als die Nicht-Raucherinnen. Die Re-Analyse zeigte aber, dass sich in der Raucherinnengruppe im Durchschnitt jüngere Frauen befanden als bei den Nicht-Raucherinnen. Da jüngere Frauen unabhängig davon, ob sie rauchen, eine längere Lebenserwartung als ältere Frauen haben, kam es zu der Umkehrung der üblichen Beziehung zwischen Rauchen und Lebenserwartung. In diesem Beispiel ist also das Lebensalter der Probandinnen die konfundierende Variable, die zu einer Verzerrung der Kovariation zwischen Rauchen und Lebenserwartung führte.

Kontrolle und interne Validität

Ziel der Planung von Experimenten ist es nun, Designs zu realisieren, die die kausale Interpretierbarkeit der beobachteten statistischen Kovariationen sicherstellen. Campbell und Stanley (1963) haben zur Beschreibung dieses Idealzustands das Konzept der *internen Validität* eingeführt. Untersuchungen sind dann intern valide, wenn die

Variation der AV eindeutig auf Veränderungen der UV zurückführbar ist und die beobachtete statistische Kovariation zwischen der UV und der AV nicht durch Störvariablen (KV) verzerrt wird, die mit der UV konfundiert sind. Ein intern valides Experiment ist also frei von Konfundierungen. Da man in der Regel nicht alle relevanten Einflussfaktoren kennt, ist interne Validität ein Ziel, das man zu erreichen versucht, ohne dass man sich immer sicher sein kann, dass nicht doch irgendeine Art von Konfundierung vorliegt.

Wie kann man nun versuchen, das Ziel der Abwesenheit von Konfundierung zu erreichen? Wir haben gesehen, dass Kausalmodelle, in denen sich, wie in Abb. 1.2, mehrere Ereignisse wechselseitig beeinflussen, durchaus häufig in der Realität vorkommen, so dass man bei freien Beobachtungsstudien mit Konfundierungen rechnen muss. In Experimenten wird man deshalb in die beobachteten Situationen in bestimmter Weise aktiv eingreifen. Durch *Kontrolltechniken* versucht man, die natürlich auftretenden Konfundierungen zu durchbrechen und die beobachtete Situation so zu verändern, dass die kausale Interpretierbarkeit der statistischen Effekte sichergestellt wird.

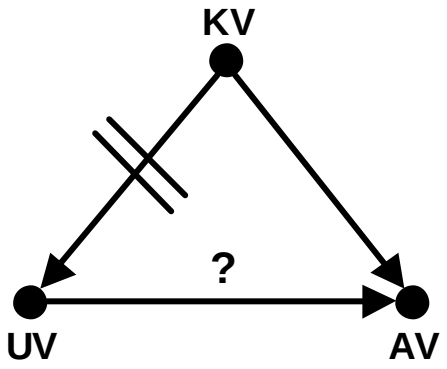

Abbildung 1.3 *Ziel von Kontrolltechniken bei der Planung von Experimenten.*

Abb. 1.3 veranschaulicht das Ziel sämtlicher Kontrolltechniken. Um die Kovariation zwischen UV und AV kausal interpretierbar zu machen, wird versucht, die bei natürlichen Beobachtungen eventuell vorliegende Kovariation zwischen den KV und der UV auf Null zu setzen. Dies heißt nicht, dass man den kausalen Zusammenhang zwischen KV und UV leugnet, sondern dass man ihn in dem aktuellen Experiment ausblendet, um die Hypothese prüfen zu können, dass die UV und die AV kausal verknüpft sind. Weitere Studien können sich dann mit der kausalen Beziehung zwischen der KV und der UV beschäftigen, wenn dies von wissenschaftlichem Interesse ist. Es steht eine Reihe von Techniken zur Verfügung, mit denen man das Ziel der Kontrolle erreichen kann:

Elimination und Konstanthalten. Die nahe liegendste Methode besteht darin, das Experiment so zu planen, dass bestimmte Störvariablen gar nicht auftreten. Führt man etwa ein Wahrnehmungsexperiment in einem schallisolierten Raum durch, dann kann die potentiell konfundierende Variable Lärm gar nicht zum Tragen kommen. Elimination von Störvariablen ist aber nicht immer möglich. Eine weitere Technik

besteht deshalb darin, dass man die konfundierende Variable konstant hält. Dabei wählt man eine Ausprägung dieser Variable aus und verändert diese nicht zwischen den Bedingungen des Experiments. So kann man beispielsweise alle Bedingungen vom gleichen Versuchsleiter durchführen lassen, man kann den gleichen Untersuchungsraum verwenden, das Experiment nur bei einer Altersstufe oder bei Versuchsteilnehmern gleichen Geschlechts oder gleicher Bildungsstufe durchführen. Konstanthalten ist auch eine Technik, die sich anbietet, wenn man Konfundierung durch verschiedene Merkmale der UV vermutet wie bei dem diskutierten Beispiel der Medikamentenstudie. Gibt man sowohl den Versuchsteilnehmern der Experimentalgruppe als auch denen der Kontrollgruppe eine Pille, wobei diese nur in der ersten Gruppe den Wirkstoff trägt, dann hat man die potentielle Störvariable Medikamenteneinnahme konstant gehalten (Placebo-Kontrolle). In all diesen Beispielen wird die potentielle Störbeziehung zwischen KV und UV dadurch auf Null gesetzt, dass die KV durchweg eine konstante Ausprägung einnimmt und deshalb wegen ihrer fehlenden Variation nicht mit der UV kovariieren kann. Grundsätzlich gibt es natürlich eine unendliche Zahl möglicher Merkmale, die man konstant halten könnte (z.B. Raumtemperatur, Uhrzeit der Untersuchung). Man wird deshalb nur solche Faktoren konstant halten, für die man einen theoretisch oder empirisch begründeten Verdacht hat, dass es sich um Störvariablen handeln könnte.

Balancierung. Nicht immer möchte oder kann man sich auf eine Ausprägung der potentiellen Störvariable beschränken. So lassen sich manche Merkmale wie etwa Persönlichkeitsmerkmale der Probanden gar nicht oder nur sehr schwer konstant halten. Eine mögliche Technik in diesen Situationen besteht dann darin, dass man die KV zwar variieren lässt, dass man das Experiment aber so anlegt, dass sie mit Null mit der UV kovariiert. So könnte man in der fiktiven Demonstration zu Simpsons Paradox (Tab. 1.2) etwa eine experimentelle Studie so planen, dass sowohl in der Experimentalbedingung (Medikamenteneinnahme) als auch in der Kontrollbedingung (kein Medikament) je gleich viele Männer und Frauen teilnehmen. Man greift also in die natürliche Situation ein und weist Männer und Frauen zu gleichen Anteilen nach Zufall den beiden Bedingungen zu, auch wenn die Versuchsteilnehmer spontan vielleicht lieber in die andere Gruppe gegangen wären. In ähnlicher Weise kann man auch unterschiedliche Versuchsleiter so einsetzen, dass jeder Versuchsleiter in allen experimentellen Bedingungen gleich häufig eingesetzt wird. Auf diese Weise kann es zu keiner Kovariation zwischen Versuchsleiter (KV) und Versuchsbedingung (UV) kommen. Andere Faktoren, die man balancieren kann, sind der Untersuchungszeitpunkt oder der Versuchsraum.

Balancierung ist auch eine besonders wichtige Technik bei Messwiederholungsdesigns, bei denen die Versuchsteilnehmer nacheinander mehrere Bedingungen durchlaufen. Würde man etwa die Experimentalbedingung stets vor der Kontrollbedingung durchführen, dann kann es zu sogenannten Konfundierungen mit *Übungseffekten* kommen, also Einflüssen, die durch die Reihenfolge der Bedingungen zustande kommen. Untersucht man etwa eine neue Lehrmethode immer als erste Bedingung und vergleicht sie mit der Standardbedingung, die als zweites durchgeführt wird, dann ist es denkbar, dass die Leistung in der zweiten Bedingung teilweise durch die erste Bedingung beeinflusst wird. Es ist möglich, dass manche Ver-

suchsteilnehmer auch hier die neue Lehrmethode einsetzen, was zu einer künstlichen Verringerung des Unterschieds führt. Bei vielen nacheinander durchgeführten Versuchsbedingungen kann es auch zu Ermüdungs- oder Langeweileeffekten kommen. Auch Kontrasteffekte sind denkbar. Soll beispielsweise die Attraktivität von Personen eingeschätzt werden, dann wird die Einschätzung dadurch beeinflusst werden, wie attraktiv die zuvor eingeschätzte Person war. Schließlich ist es denkbar, dass frühere Bedingungen zu Erwartungen führen, die sich in die eine oder andere Richtung störend auswirken können.

Um diese Störfaktoren zu kontrollieren, muss man deshalb in Messwiederholungsdesigns die Reihenfolge der Untersuchungsbedingungen balancieren. Eine Balancierungsmöglichkeit besteht daran, den Probanden alle möglichen Reihenfolgen von Bedingungen nacheinander vorzugeben. Beispiele für solche Studien sind etwa Wahrnehmungsexperimente, bei denen Töne verschiedener Lautstärke verglichen werden. Hier ist es möglich, jeder Versuchsperson eine Vielzahl solcher Töne in unterschiedlicher Reihenfolge nacheinander vorzulegen.

Häufig verbietet das Versuchsmaterial allerdings die Realisierung mehrerer Reihenfolgen bei der gleichen Versuchsperson, da es sonst zu Lerneffekten und anderen Störungen kommen kann. Eine alternative Möglichkeit besteht deshalb darin, dass jede Versuchsperson in dem Experiment nur eine einzelne Reihenfolge von Bedingungen durchläuft, wobei aber die Reihenfolge zwischen verschiedenen Probanden variiert wird. Ideal wäre es, wenn in dem Experiment alle möglichen Reihenfolgen von Bedingungen untersucht würden. Dies ist aber nur realistisch, wenn die Anzahl der Bedingungen vergleichsweise klein ist. Bereits bei fünf Bedingungen gibt es schon 120 mögliche Reihenfolgen. Für diesen Fall wurde eine Reihe von alternativen Balancierungstechniken entwickelt, die zwar nicht eine vollständige Kontrolle von potentiellen Reihenfolgeeffekten erlauben, dennoch aber die wichtigsten Aspekte balancieren (vgl. auch Shaughnessy & Zechmeister, 1997; Kirk, 1995). *Lateinische Quadrate* sind beispielsweise Versuchsanordnungen, bei denen jede Bedingung gleich häufig an jeder Position vorkommt und gleich häufig vor und nach jeder anderen Bedingung auftritt. Tab. 1.3 gibt ein Beispiel für ein Lateinisches Quadrat mit vier Bedingungen A, B, C und D. Jede Zeile, die einer anderen Reihenfolge von Bedingungen entspricht, wird von einer anderen Gruppe von nach Zufall der Bedingung zugewiesenen Versuchspersonen durchlaufen. Wie man sieht, erfüllt dieses Design die Kriterien von Lateinischen Quadraten.

Tabelle 1.3 *Beispiel für ein Lateinisches Quadrat mit vier Versuchsbedingungen.*

Reihenfolge	Position			
	Erste	Zweite	Dritte	Vierte
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

Parallelisierung von Versuchsgruppen. Merkmale der Versuchsteilnehmer (Geschlecht, Intelligenz, Temperament, Motivation, Gesundheitszustand usw.) sind eine wichtige Quelle möglicher Konfundierungen. Normalerweise versucht man durch zufällige Zuteilung der Versuchspersonen auf die unterschiedlichen Versuchsbedingungen mögliche Kovariationen dieser Merkmale (KV) mit der UV zu unterbinden. Gelegentlich führt man aber Experimente durch, bei denen man nur wenige Probanden untersuchen kann, so dass man damit rechnen muss, dass es zu zufälligen Unterschieden zwischen den Bedingungen kommt. So hat man es in klinischen Studien mit Patienten häufig mit kleinen Stichproben zu tun. Eine mögliche Kontrolltechnik in diesen Situationen besteht darin, die Versuchsgruppen nach den relevanten Merkmalen zu parallelisieren. So kann man in einer Studie zur Wirksamkeit eines Herzmedikaments die Versuchsteilnehmer nach ihrem Alter in verschiedene Klassen aufteilen und dann von jeder Klasse nach Zufall gleich viele Probanden den verschiedenen Bedingungen zuweisen. Auf diese Weise ist sichergestellt, dass die Gruppen im Hinblick auf die potentiell konfundierende Variable Alter gleichartig sind, so dass es zu keiner Kovariation zwischen der KV Alter und der UV kommen kann. Diese Technik ist nur dann sinnvoll, wenn man eine Störvariable kennt, die potentiell hoch mit der abhängigen Variable kovariiert, da ansonsten Kontrolle unnötig wäre. Ein extremer Fall dafür wäre eine Studie, bei der man nach der abhängigen Variable parallelisiert. Man könnte etwa in einer Studie, die die Wirksamkeit eines Blutdruck senkenden Medikaments überprüft, die Gruppen nach ihrem Blutdruck parallelisieren. Häufig ist diese Strategie, die AV als Parallelisierungsvariable zu verwenden, aufgrund möglicher Übungseffekte nicht praktikabel. Untersucht man beispielsweise Problemlösekompetenzen bei bestimmten Problemtypen, dann kann man die Gruppen nicht aufgrund dieses Merkmals parallelisieren, weil man den Probanden dadurch ja bereits die Aufgabe vorher vorlegen müsste, die man später untersuchen möchte. Hier muss man sich mit einer anderen kovariierenden Störvariable (z.B. Intelligenz) behelfen.

Randomisierung. Die vielleicht wichtigste Kontrolltechnik, die Experimente von reinen Beobachtungsstudien unterscheidet, ist die Randomisierung. Bisher mag der Eindruck geweckt worden sein, dass Beobachtungsstudien aufgrund der inhärenten Konfundierungsproblematik nicht oder nur schwer interpretierbar sind. Dies stimmt nicht generell. Kennt man die relevanten Einflussfaktoren, dann ist es möglich, deren Einfluss durch statistische Verfahren konstant zu halten. So kommt man in dem Beispiel für Simpsons Paradox (Tab. 1.2) etwa zur richtigen Einschätzung der kausalen Wirksamkeit des Medikaments, wenn man die Kontingenztafel getrennt nach Geschlecht betrachtet und damit diesen Faktor statistisch konstant hält (vgl. auch Waldmann & Hagmayer, in press). Komplexe statistische Verfahren, die sich der Modellierung von Kausaleinflüssen widmen (LISREL), berücksichtigen den Einfluss von konfundierenden Faktoren mit Hilfe statistischer Kontrolltechniken (vgl. Bollen, 1989; Loehlin, 1998). So weit, so gut. Diese Techniken funktionieren aber nur, wenn man alle relevanten Einflussfaktoren kennt. Gibt es weitere unbekannte konfundierende Faktoren, dann kann es zu der bereits bei der Diskussion von Simpsons Paradox aufgezeigten Problematik der Uninterpretierbarkeit der beobachteten Kovariationen kommen. In Beobachtungsstudien wird bei der Analyse deshalb routinemäßig die empirisch nicht überprüfbare Annahme gemacht, dass etwaige weitere, unbe-

kannte Einflussfaktoren zwar existieren mögen, aber dass diese mit den beobachteten Variablen *nicht* kovariieren. Die für eine kausale Interpretation notwendige Annahme der Nullkorrelation mit unbekannten Störvariablen (Abb. 1.3) muss also hier als *Vorannahme* bei der statistischen Analyse zugrunde gelegt werden.

Demgegenüber bieten Experimente den unschätzbaren Vorteil, dass sie es ermöglichen, auch den Einfluss *unbekannter Faktoren* zu kontrollieren. Die bisher diskutierten Kontrolltechniken zielten auf das bewusste Herstellen einer Nullkorrelation zwischen bekannten Störvariablen und der UV. Randomisierung ist eine Technik, die dies auch für unbekannte Faktoren ermöglicht. Unter Randomisierung versteht man beispielsweise die Zuordnung von Probanden auf die verschiedenen Versuchsbedingungen nach Zufall. Dies bedeutet, dass jeder Versuchsteilnehmer bei einem Experiment mit gleicher Wahrscheinlichkeit den verschiedenen Bedingungen zugewiesen werden kann. Ist die Stichprobe groß genug, dann sollte es auf diese Weise dazu kommen, dass die verschiedenen Versuchsgruppen im Hinblick auf ihre Merkmale vergleichbar sind, selbst wenn man viele der relevanten Merkmale gar nicht kennt. Nehmen wir beispielsweise am Beispiel des Simpson-Paradoxes in Tab. 1.2 an, dass die konfundierende Variable nicht ein augenfälliges Merkmal wie das Geschlecht der Versuchsteilnehmer ist, sondern ein physiologisches Merkmal, das schwer zu messen ist oder dessen Relevanz man noch nicht kennt. Weist man in einem Experiment die Probanden nach Zufall den Bedingungen zu, dann sollte bei ausreichend großer Zahl die Wahrscheinlichkeit hoch sein, dass sich die Probanden im Hinblick auf dieses unbekannte Merkmal gleich über die Vergleichsgruppen verteilen. Dies wäre in einer Beobachtungsstudie nicht sichergestellt, wenn diese Variable mit der UV kovariieren sollte.

Randomisierung der Probanden ist nur ein Beispiel für den Einsatz dieser Technik. Randomisieren lassen sich auch andere Aspekte eines Versuchs wie die Reihenfolge der Bedingungen oder die Auswahl des Versuchsmaterials. Untersucht man beispielsweise die Geschwindigkeit, mit der Wörter unterschiedlicher Häufigkeit erkannt werden, dann wird man sich nicht auf bestimmte Wörter konzentrieren, sondern eine Zufallsstichprobe von Wörtern verschiedener Häufigkeit einsetzen. Auf diese Weise kann man sicherstellen, dass die gefundenen Effekte nicht aufgrund einer Konfundierung mit weiteren relevanten Merkmalen der gewählten Wörter zustande kommen, die man vielleicht noch gar nicht kennt.

Zusammen mit Manipulation gehört Randomisierung zu den wichtigsten Merkmalen echter Experimente. Quasi-Experimente, bei denen natürliche Gruppen (z.B. Alter, Geschlecht) verglichen werden, lassen keine randomisierte Zuteilung der Probanden in Bezug auf diese Gruppenmerkmale zu und teilen mit Beobachtungsstudien deshalb das Problem potentieller Konfundierung durch Faktoren, die mit den Gruppenmerkmalen korreliert sind.

Theorienüberprüfung und Distinktheit von Vorhersagen

Um die Diskussion zu vereinfachen, wurde bisher so getan, als handle es sich bei Variablen wie Frustration und Aggression um direkt beobachtbare Ereignisse. Tatsächlich beziehen sich psychologische Theorien vorwiegend auf *theoretische Kon-*

zepte bzw. theoretische Variablen (TV), die nicht direkt beobachtbar sind, deren Auswirkung im Verhalten man aber erkennen kann. Die theoretischen Konzepte bezeichnen dabei erschlossene psychische Ereignisse, die häufig nicht alleine stehen, sondern Teil eines komplexen Netzwerks theoretischer Konzepte sein können. Wie man nun theoretische Konzepte mit beobachtbaren Verhalten in Zusammenhang bringen kann, ist eine in der Wissenschaftstheorie heftig umstrittene Frage (vgl. Westermann, 2000). In diesem Beitrag wird die Position vertreten, dass sich theoretische Konzepte ebenso auf kausal wirksame Ereignisse beziehen wie Konzepte, die direkt auf beobachtbare Ereignisse referieren. Dabei können die durch theoretische Konzepte beschriebenen Ereignisse sowohl andere theoretisch postulierte als auch direkt beobachtbare Ereignisse kausal beeinflussen. Da theoretische Konzepte nicht direkt beobachtbare Ereignisse beschreiben, kann man diese nur auf der Basis von Kovariationen zwischen beobachtbaren Ereignissen erschließen. Frustration beispielsweise wird als psychischer Zustand konzeptualisiert, der sich erschließen lässt, wenn Personen auf bestimmte Aktionen des Versuchsleiters (etwa Beschimpfungen) mit einem bestimmten Verhalten (Unmutsbezeugungen) reagieren, das auf Frustration schließen lässt. Die Beschimpfungen spielen hier die Rolle der beobachtbaren Ursache von Frustration, wobei das Verhalten ein beobachtbarer Effekt dieses postulierten psychischen Zustands darstellt.

Ein etwas komplexeres, aber immer noch extrem vereinfachtes Beispiel zeigt Abb. 1.4. Am Beispiel einer einfachen Variante eines Gedächtnismodells mit drei zeitlich geordneten Phasen kann man sehen, wie die beobachteten Variablen (UV und AV) mit einem Netzwerk kausal verknüpfter, von der Theorie postulierter Konstrukte (TV) in Verbindung stehen. Beobachtbare Variablen (UV, AV) werden in solchen Diagrammen durch Rechtecke, theoretische Konzepte (TV) durch Ellipsen symbolisiert.

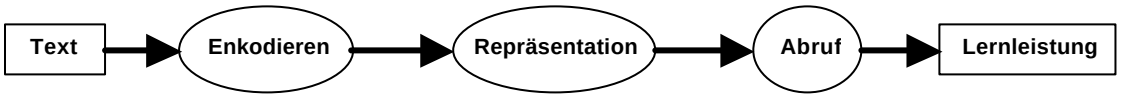

Abbildung 1.4: *Beispiel eines einfachen Gedächtnismodells mit drei theoretischen, kausal verknüpften Ereignissen.*

Wie kann man nun kausale Theorien, deren Elemente nur zum Teil direkt beobachtet werden können, empirisch überprüfen? Dies ist natürlich ein sehr komplexes Thema, das bisher wissenschaftstheoretisch nur teilweise befriedigend gelöst ist. Dennoch gibt es eine Reihe praktischer für die Versuchsplanung wichtiger Gesichtspunkte, die hier angesprochen werden sollen.

Theoretische Konzepte und theoretische Modelle gewinnen ihre pragmatische und empirische Nützlichkeit dadurch, dass sie nicht nur Zusammenhänge zwischen einer einzelnen UV und einer einzelnen AV vorhersagen, sondern eine Reihe ganz verschiedener Vorhersagen zulassen. Gedächtnistheorien beispielsweise werden mit dem Ziel postuliert, Gedächtnisleistungen in ganz verschiedenen Aufgabenkontexten zu erklären. Es ist also möglich, dass eine Theorie, die bislang zu richtigen Voraussagen geführt hat, in einem neuen Anwendungsbereich, für den sie Geltung bean-

sprucht, versagt. Die Konsequenz einer solchen Situation kann sein, dass man die Theorie modifiziert oder ihren Anwendungsbereich einschränkt (vgl. Westermann, 2000).

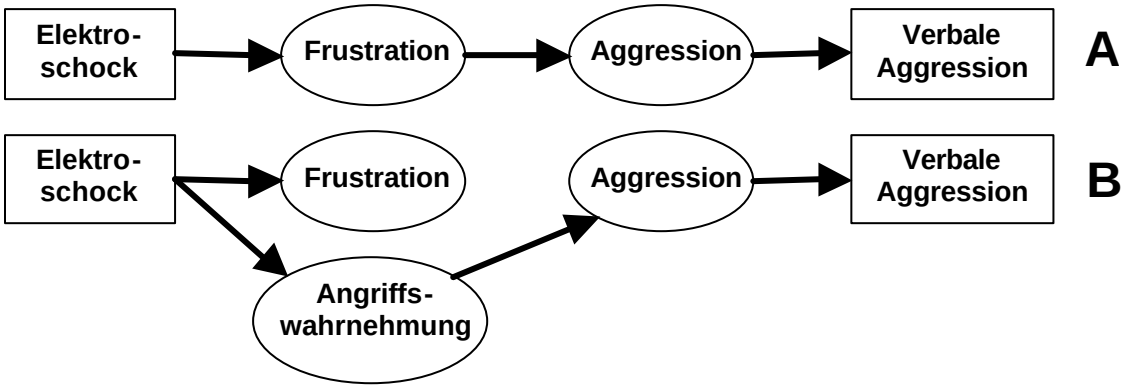

Abbildung 1.5: *Alternative Modelle für den Zusammenhang von Frustration und Aggression (nach Bollen,1989).*

Selten gibt es nur eine einzige Theorie für einen Phänomenbereich. In der Regel konkurrieren mehrere Theorien miteinander. Selbst wenn eine Theorie zunächst alleine steht, geht es in wissenschaftlichen Diskussionen nahezu immer darum, ob es nicht eine alternative Theorie gibt, die die Daten besser erklärt und vielleicht weitere Vorzüge wie größere Einfachheit hat. Bollen (1989, S. 74ff.) diskutiert diesen Sachverhalt am Beispiel der Frustrations-Aggressions-Hypothese. Eine einfache Variante dieser Theorie mag postulieren, dass Elektroschocks (UV) zu Frustration (TV) führen, was wiederum Aggression (TV) auslöst. Dieser innere Zustand manifestiert sich schließlich in verbalen aggressiven Äußerungen (AV). Die Beobachtung, dass Elektroschocks zu verbalen aggressiven Äußerungen führen, wäre demnach konsistent mit der postulierten Theorie (vgl. Abb. 1.5A). Ein Problem entsteht aber, wenn ein anderer Wissenschaftler das Modell in Abb. 1.5B vorschlägt, bei dem Elektroschocks zwar zu Frustration führen, aber auch zu einem Gefühl, angegriffen worden zu sein. Nach diesem Alternativmodell ist dieses Gefühl dann die Ursache für Aggression. Dieses Modell ist *empirisch äquivalent*, da es, so wie es hier formuliert ist, identische Vorhersagen wie Modell 1.5A macht. Es handelt sich hier also um einen weiteren, bisher nicht diskutierten Typ von Konfundierung, der durch alternative Theorien ins Spiel kommt. Die zwischen der UV und der AV beobachtete Kovariation kann durch die in Modell 1.5A oder in Modell 1.5B hypothetisch angenommenen unterschiedlichen kausalen Ketten generiert werden. Diese Art von Konfundierung lässt sich *nicht* durch Kontrolltechniken wie Randomisierung in den Griff bekommen. Die diskutierten Kontrolltechniken zielen darauf ab, den Einfluss *alternativer Ursachen* (vgl. Abb. 1.3) zu kontrollieren, während Konfundierung durch äquivalente Modelle sich auf die zwischen UV und AV vermittelnden Kausalglieder bezieht. Kontrolltechniken zielen darauf ab, alternative Ursachen (KV) statistisch unabhängig von der interessierenden Ursache (UV) zu machen, während dies natürlich nicht das Ziel bei den Stufen der kausalen Kette sein kann, die zwischen der UV und der AV vermittelt. Würde man in analoger Weise wie bei Kontrolle versuchen, die statistische Bezie-

hung zwischen UV und den kausal vermittelnden theoretischen Variablen (TV) auf Null zu setzen, würde dies auch die interessierende Verbindung zwischen der UV und der AV kappen. Deshalb muss man hier zu anderen Methoden der Dekonfundierung greifen.

Wie geht man nun mit diesem Typ von Konfundierung um? Ein interessanter Aspekt dieses Beispiels ist, dass es deutlich macht, dass es nicht genügt, aus Modellen deduktive Vorhersagen zu machen und empirisch zu testen. Nicht jede Vorhersage ist interessant, da viele Vorhersagen, wie in dem in Abb. 1.5 gezeigten Beispiel, auch von alternativen Theorien gemacht werden. Wonach Wissenschaftler also suchen sollten, sind *distinkte Vorhersagen*, also Vorhersagen, die nur von der eigenen Theorie gemacht werden, aber nicht von der Alternativtheorie. So könnte man etwa die beiden in Abb. 1.5 skizzierten Theorien erweitern und dadurch gegeneinander testen, dass man versucht, einen Zustand von Frustration herzustellen, der nicht zu einem Gefühl von Angegriffensein führt. Sollte sich dann dennoch Aggression im Verhalten zeigen, deutet dies eher auf Modell 1.5A, ansonsten wäre Modell 1.5B die überlegene Variante (zumindest in Bezug auf dieses eine Experiment).

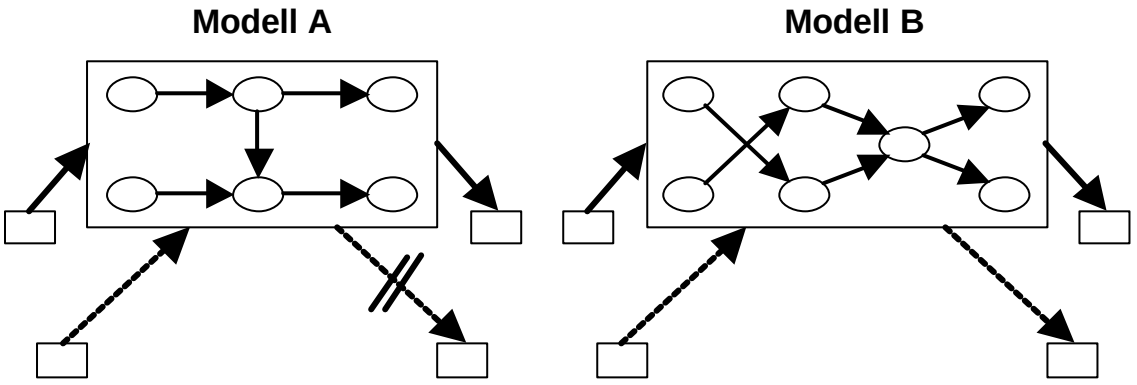

Abbildung 1.6: *Zwei konkurrierende Modelle (A und B) mit gleichartigen und distinkten Vorhersagen.*

Abb. 1.6 zeigt die abstrakte Struktur distinkter Vorhersagen. Verglichen werden zwei alternative Kausalmodelle, die sich auf theoretisch postulierte Zustände beziehen. Vorhersagen, die beide Modelle gleichermaßen machen (durchgehende Pfeile zwischen Modell und den beobachteten Variablen), unterscheiden zwischen den Theorien nicht. Die Prüfung dieser gemeinsamen Vorhersagen lässt keine Diskrimination zwischen den konkurrierenden Theorien zu, wenngleich ein möglicher negativer Befund immerhin die Schlussfolgerung stützt, dass *beide* Modelle unangemessen sind. Interessanter für die experimentelle Überprüfung der Theorien sind allerdings die gestrichelten Pfeile. Hier zeigt sich, dass Modell B eine Vorhersage macht, die von Modell A nicht geteilt wird. Eine empirische Überprüfung dieser distinkten Vorhersage kann also ein erster Schritt in Richtung der Entscheidung zwischen den konkurrierenden Theorien darstellen.

Generalisierbarkeit und die Bedeutung multifaktorieller Designs

Neben der internen Validität wurde von Campbell und Stanley (1963) auch das Konzept der *externen Validität* als weiteres Gütekriterium für Experimente vorgeschlagen. Unter externer Validität versteht man das Ausmaß, in dem die Ergebnisse der Studie auf andere Personen, Situationen und Operationen verallgemeinert werden können. Repräsentativität der Stichprobe und der Untersuchungssituation wird hierbei häufig als Voraussetzung für externe Validität diskutiert.

Die Bedeutung externer Validität ist in der Methodenliteratur allerdings heftig umstritten (vgl. Frick, 1998; Hager, 1992; Mook, 1983, Westermann, 2000, für kritische Positionen). So zeigt bereits ein kursorischer Überblick über die Forschung in der Psychologie, dass Repräsentativität der Stichprobe nicht nur nicht realisiert, sondern meist nicht einmal angestrebt wird. In der Regel werden Experimente mit gerade verfügbaren Studierenden der ersten Semester in künstlichen Laborsituationen mit eigens für die experimentelle Fragestellung konstruierten Aufgaben durchgeführt. Auch bei Experimenten, die im WWW durchgeführt werden, wird keine Prozedur angewendet, die die Repräsentativität der Stichprobe herbeiführt. Ist dies nun ein Defizit dieser Studien?

Die Position, die in diesem Beitrag bezogen wird, verneint diese Frage. Das Ziel von psychologischen Experimenten ist die Untersuchung von Kausalbeziehungen. Wie wir gesehen haben, ist es dazu in der Regel nötig, künstliche, nichtrepräsentative Laborsituationen herzustellen, die in die Struktur natürlicher Ereigniszusammenhänge eingreifen. So lässt sich die kausale Wirksamkeit des Medikaments in dem Beispiel für Simpsons Paradox (Tab. 1.2) nur untersuchen, wenn man die Untersuchungssituation gegenüber der natürlichen (repräsentativen!) Situation so verändert, dass die Kovariation zwischen Geschlecht (KV) und Medikamenteneinnahme (UV) auf Null gesetzt ist. Auch die Realisierung des Ziels, distinkte Vorhersagen zu testen, um zwischen alternativen Theorien entscheiden zu können, macht es häufig nötig, eher untypische alltagsferne Aufgaben zu wählen. Ähnlich wie in Studien aus der Physik werden also bewusst künstliche Versuchsbedingungen geschaffen, um eine bestimmte Kausalhypothese besser untersuchen zu können.

Es gibt zwar Studien in den Sozialwissenschaften, bei denen Repräsentativität der Stichprobe wichtig ist, etwa wenn man eine Vorhersage über das Wahlverhalten machen möchte. In solchen Studien ist es tatsächlich notwendig, dass das in der Stichprobe beobachtete Verhalten dem der zugrunde liegenden Population maximal ähnelt. Wissenschaftliche Theorien gehen aber einen Schritt weiter; sie interessieren sich für die kausalen Prozesse, die hinter den beobachteten Zusammenhängen stehen.

Dies bedeutet natürlich nicht, dass Laborexperimente immer richtiges und vollständiges Kausalwissen erzeugen. Es ist geradezu der Regelfall, dass man nur einen Teil der relevanten Randbedingungen und Kausalfaktoren kennt, die einen Effekt beeinflussen. Dieses Defizit kann aber nicht dadurch überwunden werden, dass man repräsentative Stichproben von Versuchsteilnehmern oder Aufgaben zieht, sondern dadurch, dass man systematisch weitere, theoriengeleitete Studien macht, die das bereits vorhandene Kausalwissen erweitern. Damit wird die Untersuchung der *Generalität* von Effekten zu einem der wichtigsten Ziele experimenteller Forschung (vgl.

Abelson, 1995). Die Generalität von Effekten kann mit verschiedenen Verfahren untersucht werden.

Replikation. Eine wichtige Methode, um die Zuverlässigkeit und die Generalität eines Effekts sicherzustellen, sind Replikationen von Experimenten. Eine exakte Replikation wird dabei so gut wie nie möglich sein. Versuchsteilnehmer, Versuchsleiter und andere Aspekte der Versuchssituation werden nahezu immer variieren. Häufig sind auch konzeptuelle Replikationen, bei denen man alternative unabhängige oder abhängige Variablen aus der Theorie ableitet und untersucht. Schließlich findet man in der Literatur auch oft partielle Replikationen, bei denen einige Bedingungen einer früheren Studie wiederholt und mit neuen Bedingungen gepaart werden. So kann man bei erfolgreicher Replikation zum einen sicherstellen, dass die eigenen Versuchsbedingungen den in anderen Studien gefundenen entsprechen; zum anderen kann man aber auch die Auswirkung eines neuen Einflussfaktors untersuchen und somit die kausale Theorie erweitern.

Das Problem interaktiver Effekte. Bisher hatten wir gesagt, dass man durch Kontrolltechniken wie Konstanthalten und Randomisierung mögliche Konfundierungen durch weitere Variablen in den Griff bekommen kann, indem man ihre statistische Beziehung zur UV auf Null setzt (vgl. Abb. 1.3). Streng genommen setzen diese Verfahren aber die Erfüllung einer weitere Voraussetzung voraus, die wir noch nicht diskutiert haben. Sie setzen nämlich voraus, dass die kausalen Einflüsse der UV und der KV auf die AV *unabhängig* voneinander sind. Unabhängigkeit der Wirkung ist nicht das Gleiche wie Unabhängigkeit des Auftretens von UV und KV. Es ist möglich, durch Kontrolltechniken die Unabhängigkeit des Auftretens von UV und KV herzustellen, ohne dass dadurch sichergestellt ist, dass diese beiden Faktoren auch unabhängig auf die untersuchte AV einwirken.

Dies lässt sich am besten anhand eines Beispiels verständlich machen. Angenommen wir untersuchen den Einfluss von Frustration auf Aggression bei männlichen und weiblichen Versuchsteilnehmern und vermuten, dass Geschlecht eine konfundierende Variable ist. Durch Balancierungstechniken können wir nun erreichen, dass die UV (Frustration) und die KV (Geschlecht) unkorreliert sind (Abb. 1.3). Nun gibt es dennoch verschiedene Möglichkeiten, wie KV und UV zusammenspielen in ihrem kausalen Einfluss auf die AV. Abb. 1.7 zeigt eine Reihe unterschiedlicher Datenmuster, die bei Untersuchungen mit zwei zweistufigen Faktoren auftreten können. So könnte in unserem Beispiel Faktor A die Variable Frustration bezeichnen (A1=abwesend, A2=anwesend) und Faktor B das Geschlecht der Probanden (B1=Frauen, B2=Männer). Die Y-Achse repräsentiert die AV, also in unserem Beispiel die Intensität der Aggression. Zunächst besteht die Möglichkeit, dass nur einer der beiden Faktoren (oder keiner) kausal wirksam ist. Abb. 7A gibt ein Beispiel für eine Situation, in der Frustration Aggression beeinflusst (A2 liegt höher als A1), es aber keinen Unterschied in der Aggressivität zwischen den beiden Geschlechtsgruppen gibt (B1 und B2 liegen übereinander). Es ist aber auch denkbar, dass beide Faktoren wirksam sind. So könnte es beispielsweise sein, dass Frustration die Aggressivität erhöht und zusätzlich, dass Männer aggressiver sind als Frauen. Wirken diese beiden Faktoren unabhängig auf den Effekt ein, dann könnte sich etwa ein Daten-

muster wie in Abb. 7B ergeben. Die beiden Linien verlaufen parallel, was ein Zeichen für die Unabhängigkeit des kausalen Einflusses ist. Der Unterschied zwischen A2 und A1 ist immer gleich groß, gleichgültig, ob man die Männergruppe (B2) betrachtet oder die Frauengruppe (B1). Diese Unabhängigkeit ist der Grund dafür, warum man mit Hilfe der Kontrolltechniken den beobachteten Effekt kausal interpretieren kann, obwohl es doch einen weiteren Faktor gibt, der die Ausprägung der AV mitbeeinflusst. Gleichgültig, ob man das Geschlecht konstant hält (z.B. nur die Frauen oder nur die Männer untersucht) oder balanciert, der Effekt für die UV wird immer in etwa gleich groß sein.

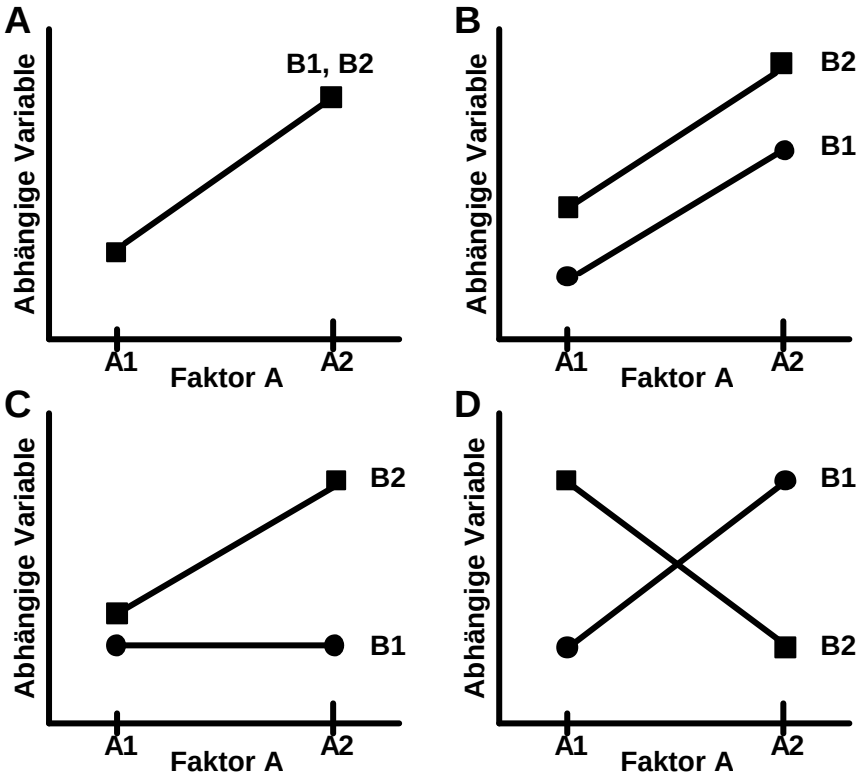

Abbildung 1.7 *Einige mögliche Befunde von Designs mit zwei Faktoren.*

Dies ist nicht so, wenn eine *Interaktion* zwischen den Ursachen in Bezug auf den Effekt vorliegt. Eine Interaktion liegt immer dann vor, wenn sich das Ausmaß des Effekts eines Faktors über die Ausprägungen anderer Faktoren verändert. Abb. 1.7C zeigt einen Fall, bei dem Faktor A nur einen Effekt zeigt, wenn Faktor B die Ausprägung B2 hat, aber nicht, wenn er die Ausprägung B1 hat. So könnte es sein, dass Frustration nur bei Männern (B2), aber nicht bei Frauen (B1) die Aggressionsneigung erhöht. Dies ist nur ein Beispiel für eine mögliche Interaktion. Alle Arten von Verläufen, bei denen die Linien nicht parallel liegen, deuten auf eine Interaktion. Eine Interaktion liegt also immer dann vor, wenn die Differenz der Mittelwerte bei den Ausprägungen einer UV (z.B. A2-A1) signifikant variiert zwischen den Teilgruppen, die den Ausprägungen anderer Faktoren entsprechen (z.B. bei Gruppe B1 im Vergleich zu Gruppe B2). Während sich in Abb. 1.7C die Linien nicht überschneiden, was man als ordinale Interaktion bezeichnet, zeigt Abb. 1.7D einen ex-

tremen Fall einer sogenannten disordinalen Interaktion. Ein konkretes Beispiel für eine solche disordinale Interaktion wäre etwa der mögliche Sachverhalt, dass Frustration Aggression bei Frauen erhöht, aber bei Männern senkt, wobei in diesem Beispiel nicht frustrierte Männer aggressiver sind als nicht frustrierte Frauen.

Liegen interaktive Effekte vor, dann liefern die diskutierten Kontrolltechniken keine Kovariationen, die korrekt die kausalen Beziehungen widerspiegeln. Hält man etwa in der Situation in Abb. 1.7C das Geschlecht konstant, dann wird sich eventuell gar kein Effekt zeigen (wenn man nämlich nur Frauen untersucht). Selbst wenn man Frauen und Männer balanciert, entspricht die beobachtete Kovariation nur dem Mittelwert zwischen dem Nicht-Effekt bei Frauen und dem starken Effekt bei Männern und nicht den tatsächlichen kausalen Verhältnissen. Am besten sieht man das Problem interaktiver Effekte bei disordinalen Interaktionen. Balanciert man im Beispiel in Abb. 1.7D den Faktor B (z.B. Geschlecht), dann würde man in den Daten gar keinen Effekt sehen. Aufgrund der Umkehrung der Effekte bei den Gruppen B1 und B2 wären nämlich die Mittelwerte von A1 und A2 gleich, obwohl tatsächlich natürlich zwei deutliche, allerdings gegenläufige Kausaleffekte zu beobachten sind.

Interaktive Effekte bilden auch ein Problem bei Messwiederholungsdesigns, bei denen man generell mit Konfundierungen durch Übungseffekte rechnen muss. Auch hier gilt, dass Balancierungstechniken nur dann diese Konfundierungen kontrollieren, wenn es keine Interaktionen zwischen der Position der Bedingung und dem Übungseffekt gibt. Ein Beispiel für eine solche Interaktion wäre etwa eine Situation, in der ein Medikament A, das nach Medikament B eingenommen wird, das Wohlbefinden erhöht, während es nach Medikament C zu einer Verschlechterung des Gesundheitszustands führt. Vermutet man solche interaktiven Reihenfolgeeffekte in einer geplanten Untersuchung, dann sollte man in diesem Fall besser auf ein Messwiederholungsdesign verzichten.

Multifaktorielle Designs. Interaktionen sind der wichtigste Hinweis darauf, dass eine kausale Beziehung nicht über alle kausalen Kontexte generalisierbar ist, sondern dass die Wirkung einer Ursache nur vorhersagbar ist, wenn man andere Faktoren kennt, mit denen diese Ursache interagiert. Vermutet man eine Interaktion oder möchte man sicherstellen, dass keine Interaktion zwischen zwei Faktoren vorliegt, muss man Designs mit mehreren Faktoren wählen, bei denen man die Wirksamkeit mehrerer Faktoren gleichzeitig untersucht. Tab. 1.4 gibt ein Beispiel für ein 2×2-Design mit zwei Faktoren (bzw. zwei UVs), die jeweils zwei Ausprägungen (A, B) haben. In dieser Studie würde man die Versuchsteilnehmer nach Zufall den vier Versuchsgruppen zuteilen. Mit diesem Design könnte man beispielsweise eine kausale Struktur wie in Abb. 1.3 untersuchen. An der Stelle der KV steht allerdings ein zweiter Faktor, dessen Wirksamkeit man nun explizit untersuchen möchte. Durch die zufällige und gleichmäßige Zuteilung auf die vier Versuchszellen wird sichergestellt, dass die beiden Faktoren (analog zur UV und KV in Abb. 1.3) nicht kovariieren, so dass es zu keiner Konfundierung zwischen den manipulierten Faktoren kommen kann. Kontrolltechniken wie Konstanthalten und Randomisieren muss man auch bei diesen Designs anwenden, um mögliche Konfundierungen durch weitere (möglicherweise unbekannte) Variablen auszuschalten. Multifaktorielle Designs greifen also ebenso wie einfaktorielle Designs in die Struktur natürlicher Situationen ein, um die kausale

Interpretierbarkeit der Daten zu garantieren. Mit Hilfe eines 2×2-Designs kann man Interaktionen, wie sie in Abb. 1.7 gezeigt wurden, explizit untersuchen und auf diese Weise die Randbedingungen der postulierten Theorie explorieren (vgl. Rosnow & Rosenthal, 1989, für eine kritische Diskussion der Bedeutung von Interaktionseffekten).

Tabelle 1.4 *Beispiel für ein multifaktorielles Design mit zwei Faktoren (2 ´ 2-Design).*

		Geschlecht	
		Männlich	Weiblich
Frustration	Ja	Gruppe 1	Gruppe 2
	Nein	Gruppe 3	Gruppe 4

2×2-Designs sind der einfachste Fall multifaktorieller Designs. Sollte man komplexere Interaktionen vermuten, bieten sich Designs mit mehr Stufen der einzelnen Faktoren (z.B. 3×4-Designs mit zwei Faktoren und 12 Bedingungen) oder mehr Faktoren (z.B. $2 \times 2 \times 2$-Designs mit drei Faktoren und acht Bedingungen) an. Dreifaktorielle Designs beispielsweise gestatten die Untersuchung der Frage, ob eine für eine Ausprägung des Drittfaktors vorhergesagte Interaktion zwischen zwei Faktoren bei anderen Ausprägungen dieses Drittfaktors in einer spezifischen von einer Theorie vorhergesagten Weise variiert (Dreiweg-Interaktion). Schließlich erlauben multifaktorielle Designs auch die Kombination von Faktoren mit unabhängigen Versuchsgruppen oder natürlichen Gruppen mit Messwiederholungsfaktoren (vgl. Kirk, 1995, für eine ausführliche Darstellung komplexer Designs).

Während die kausale Interpretierbarkeit einfaktorieller Designs voraussetzt, dass die UV nicht mit den konfundierenden Variablen interagiert, kann man bei komplexen Designs explizit der Frage nachgehen, ob Faktoren miteinander interagieren oder unabhängig auf den Effekt einwirken. Wieso macht man dann nicht generell multifaktorielle Designs mit vielen Faktoren? Die Antwort wird klar, wenn man sich vergegenwärtigt, dass man bei N Faktoren 2^N verschiedene Versuchsgruppen untersuchen müsste. Dupré (1993) hat darauf hingewiesen, dass bereits bei 31 Faktoren die Kapazität der Menschheit überschritten wäre, jede Zelle mit einem Probanden zu füllen. Komplexe Designs sollten also so geplant werden, dass die Frage, ob mehrere Faktoren interagieren, theoretisch begründbar und von wissenschaftlichem Interesse ist.

1.4 Schlussdiskussion

Die Grundthese dieses Artikels lautet, dass es psychologische Forschung in der Regel mit der Prüfung kausaler Theorien zu tun hat. Diesem Ziel sind andere Gesichtspunkte wie etwa Repräsentativität untergeordnet. Forscher versuchen vielmehr, natürliche Situationen im Labor dahingehend zu verändern, dass sich die empirischen

Befunde möglichst eindeutig auf die vermuteten kausalen Beziehungen zurückführen lassen. Dabei werden manche in natürlichen Kontexten vorkommende Beziehungen künstlich zerstört und Aufgaben zum Teil artifiziell abgewandelt, um die interessierende Kausalrelation besser isolieren zu können. So zielen sämtliche Kontrolltechniken darauf, die Kovariation zwischen Störvariablen und den unabhängigen Variablen auf Null zu setzen, um potentielle Konfundierungen ausschließen zu können. In komplexen Designs werden Kovariationen zwischen den interessierenden Faktoren künstlich verhindert, um die kausale Interpretierbarkeit der Daten zu garantieren. Es ist zwar möglich, auch mit Beobachtungsstudien und Quasi-Experimenten kausale Theorien zu überprüfen, aber bei diesen Untersuchungen gehen ungleich mehr Vorannahmen in die Analyse der Daten ein als bei echten Experimenten. So muss in der Regel angenommen werden, dass unbekannte Faktoren mit den bekannten nicht kovariieren. Außerdem muss man häufig Vorannahmen darüber machen, welche Ereignisse potentielle Ursachen sind und welche Effekte. Demgegenüber bieten die für Experimente charakteristischen Methoden der Manipulation und der Randomisierung weitergehende Möglichkeiten. Durch Manipulation lässt sich empirisch überprüfen, ob ein Ereignis eine Ursache anderer Ereignisse ist oder nicht. Randomisierung erlaubt es, Unabhängigkeit zwischen der unabhängigen Variablen und alternativen unbekannten Ursachen aktiv herzustellen anstatt lediglich stipulieren zu müssen.

Dennoch haben wir gesehen, dass auch die Befunde von Experimenten ohne Vorannahmen nicht interpretierbar sind. So muss bei Schlussfolgerungen aus statistischen Befunden der Daten angenommen werden, dass die untersuchte unabhängige Variable nicht mit weiteren unbekannten Variablen bei der Produktion des Effekts interagiert. Vermutet man eine solche Interaktion, dann kann man das Problem allerdings in den Griff bekommen, indem man eine multifaktorielle Studie durchführt, die diese Interaktion zum Gegenstand hat, wenngleich auch diese Studie nur interpretierbar ist, wenn man davon ausgeht, dass es nicht Interaktionen höherer Ordnung mit weiteren unbekannten Faktoren gibt.

An dieser Stelle wird häufig von Studierenden die Frage gestellt, wie man denn unter diesen Bedingungen sicher sein kann, dass man es mit einem stabilen interpretierbaren Befund zu tun hat. Eine Antwort darauf ist, dass es Grund zur Annahme gibt, dass komplexe (insbesondere disordinale) Interaktionen eher eine Seltenheit darstellen und nicht die Regel sind. Wir wären kaum in der Lage gewesen zu überleben, gäbe es nicht einigermaßen stabile kausale Zusammenhänge, die in verschiedenen Kontexten zumindest der Richtung nach gleichartige Effekte produzieren (Dawes, 1979). Ordinale Interaktionen mögen häufiger vorkommen, Dawes (1979) weist aber darauf hin, dass man solche Interaktionen durch Modelle mit unabhängiger Wirkung der Ursachen häufig gut approximieren kann, so dass der Fehler, den man begeht, nicht allzu groß ist. Vernachlässigt man beispielsweise in dem Beispiel in Abb. 1.7C den Faktor Geschlecht, dann wird man immer noch zu der qualitativ richtigen Schlussfolgerung kommen, dass Frustration dazu tendiert, die Aggressivität zu erhöhen, auch wenn der quantitative Effekt nicht genau der zugrundeliegenden Kausalbeziehung entspricht.

Die zweite Antwort auf diese Frage ist, dass es geradezu das Wesen von Wissenschaft ist, bisheriges Wissen in Frage zu stellen und nach Randbedingungen und weiteren Faktoren zu suchen, die eine kompletteres Bild der zugrunde liegenden kausa-

len Struktur liefern. Dabei macht es wenig Sinn, in blindem Pessimismus immer den schlimmsten Fall anzunehmen und grundsätzlich die Geltung der erhobenen Befunde in Frage zu stellen. In vielen Bereichen wie etwa der Allgemeinen Psychologie kann man aufgrund unseres Vorwissens davon ausgehen, dass sich die basalen kognitiven Funktionen zwischen Individuen nicht qualitativ unterscheiden, so dass etwa Interaktionen mit Personenmerkmalen eher nicht plausibel angenommen werden müssen. Anders sieht es natürlich bei Themen der Differentiellen Psychologie aus, die sich gerade für Aufgaben interessiert, in denen sich Individuen systematisch unterscheiden. Annahmen über Hintergrundfaktoren können und müssen also durch unser theoretisches Vorwissen begründet werden. Legt unsere Theorie die Vermutung nahe, dass bestimmte Faktoren interagieren, dann haben wir die Möglichkeit, Designs zu wählen, die dieser Möglichkeit Rechnung tragen. Der Einsatz von experimentellen und auch nicht-experimentellen Methoden bietet keine Garantie dafür, dass unsere Theorien richtig sind. Sie liefern aber das Rüstzeug dafür, vorhandene Theorien so gut wie möglich zu überprüfen, gleichzeitig schaffen sie aber auch Bewusstsein darüber, mit welchen Risiken und ungeprüften Vorannahmen die einzelnen Schlussfolgerungen behaftet sind.

Literatur

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Alloy, L. B. & Abramson, L. J. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108*, 441-485.

Appleton, D. R., French, J. M. & Vanderpump, M. P. (1996). Ignoring a covariate: An example of Simpson's paradox. *American Statistician, 50*, 340-341.

Arkes, H. R. & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General, 112*, 117-135.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). Cambridge: Cambridge University Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Bordens, K. S. & Abbott, B. B. (1999). *Research design and methods* (4th ed.). London: Mayfield.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research in teaching. In N. L. Gage (Ed.), *Handbook of research in teaching* (pp. 171-246). Chicago: Rand McNally. Nachdruck als Monographie: (1966). *Experimental and quasi-experimental designs for research in teaching*. Chicago: Rand McNally.

Chapman, L. J. & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid diagnostic signs. *Journal of Abnormal Psychology, 74*, 271-280.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.
Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist, 34*, 571-582.
Dawes, R. M. (2001). *Everyday irrationality.* Boulder: Westview Press.
Dupré, J. (1993). *The disorder of things. Metaphysical foundations of the disunity of science.* Cambridge, MA: Harvard University Press.
Eells, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.
Fischhoff, B. & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*, 239-260.
Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments & Computers, 30*, 527-535.
Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*, 1-13.
Gigerenzer, G., Todd, P. M. & ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: University Press.
Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hrsg.), *Allgemeine Experimentelle Psychologie* (S. 43-264). Stuttgart: Fischer.
Hager, W. (1992). *Jenseits von Experiment und Quasi-Experiment. Zur Struktur psychologischer Versuche und zur Ableitung von Vorhersagen*. Göttingen: Hogrefe.
Hell, W., Fiedler, K. & Gigerenzer, G. (1993). *Kognitive Täuschungen*. Heidelberg: Spektrum Akademischer Verlag.
Holyoak, K. J. & Thagard, P. (1995). *Mental leaps*. Cambridge, MA: MIT Press.
Huber, O. (1995). *Das psychologische Experiment* (2. Aufl.). Bern: Huber.
D. Kahneman, P. Slovic, & A. Tversky (Eds.)(1982), *Judgment under uncertainty: Heuristics and biases.* Cambridge, MA: Cambridge University Press.
Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Monterey, CA: Brooks/Cole.
Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.
Klayman, J. & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.
Kuhn, D. (1997). Is good thinking scientific thinking? In D. R. Olson & N. Torrance (Eds.), *Modes of thought: Explorations in culture and cognition* (pp. 261-281). New York: Cambridge University Press.
Loehlin, J.C. (1998). *Latent variable models. An introduction to factor, path, and structural analysis (3rd ed.).* Mahwah, NJ: Lawrence Erlbaum.
Medin, D. L., Ross, B. H. & Markman, A. B. (2001). *Cognitive psychology* (3rd ed.). Fort Worth : Harcourt College Publishers.
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379-387.
Mynatt, C. R., Doherty, M. E. & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85-95.

O'Reilly, R. C. & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.
Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.
Popper, K. R. (1994). *Logik der Forschung* (10. Aufl.). Tübingen: Mohr.
Rosnow, R. L. & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin*, *105*, 143-146.
Shaklee, H. & Mims, M. (1982). Sources of error in judging event covariation: Effects of memory demands. *Journal of Experimental Psychology: Human Learning and Memory, 8*, 208-224.
Shaklee, H. & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory & Cognition, 8,* 459-467.
Shanks, D. R., Holyoak, K. J. & Medin, D. L. (1996), *The psychology of learning and motivation, Vol. 34: Causal learning*. San Diego: Academic Press.
Shaughnessy, J.J. & Zechmeister, E.B. (1997). *Research methods in psychology*. Boston: McGraw Hill.
Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B (Methodological), 13*, 238-241.
Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4,* 165-173.
Snyder, M. & Swann, W. B. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology, 14,* 148-162.
Sternberg, R. J. & Ben-Zeev, T. (2001). *Complex cognition. The psychology of human thought*. New York: Oxford University Press.
Thagard, P. (2000). *How scientists explain disease.* Princeton: Princeton University Press.
Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47-88). San Diego: Academic Press.
Waldmann, M. R. (1997). Wissen und Lernen. *Psychologische Rundschau, 48,* 84-100.
Waldmann, M. R. & Hagmayer, Y. (in press). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition.*
Ward, W. C. & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19*, 231-241.
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.
Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 26* (pp. 27-82). New York: Academic Press.
Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur Psychologischen Methodenlehre*. Göttingen: Hogrefe.