methevalreport
2(1) 2000

Christof Nachtigall, Ute Suhl & Rolf Steyer

Einführung in die Konfundierungsanalyse

metheval
psychologische methodenlehre
und evaluationsforschung jena

Impressum

methevalreport
erscheint seit 1999
in unregelmäßigen Abständen
als „graue“ Schriftenreihe des Lehrstuhls für
Psychologische Methodenlehre und Evaluationsforschung
am Institut für Psychologie der Friedrich-Schiller-Universität Jena

Herausgeber:
Prof. Dr. Rolf Steyer
Sekr.: +49 (3641) 945 230
Durchwahl: +49 (3641) 945 231
Fax: +49 (3641) 945 232

rolf.steyer@uni-jena.de

Redaktion:
Dipl.Psych. Friedrich Funke
sff@uni-jena.de

Typographie:
cand.psych. Silke Zachariae
zachariae@web.de

Standort:
Thüringer Universitäts- und Landesbibliothek
Lesesaal Zweigstelle Psychologie

Internet
http://www.uni-jena.de/svw/metheval/report/

Bestellungen:
Methodenlehre und Evaluationsforschung
Institut für Psychologie
Steiger 3 Haus 1
D-07743 Jena
Deutschland

Einführung in die Konfundierungsanalyse

Christof Nachtigall, Ute Suhl & Rolf Steyer

0 Zusammenfassung

Ein häufiges Hindernis, welches der kausalen Interpretation von Effekten einer psychologischen Behandlung (Treatment) im Wege steht, ist die mögliche Verfälschung dieser Effekte aufgrund einer Störvariable. Man spricht von einer *Konfundierung*. In diesem Beitrag wird aufgezeigt, was genau eine Konfundierung ausmacht und welche Konsequenzen eine mögliche Konfundierung haben kann. Aufgrund dieser genauen Analyse des Konfundierungsbegriffs ergeben sich Wege, eine Konfundierung empirisch zu testen. Ziel dieses Beitrages ist es, eine einfache Darstellung der Kernideen der Konfundierungsanalyse zu geben.

1 Einleitung und Problemstellung

Das vornehmliche Ziel in den empirischen Wissenschaften ist es, Zusammenhänge zwischen interessierenden Merkmalen zu ermitteln und möglichst exakt zu beschreiben. In der Psychologie möchten wir beispielsweise wissen, ob es einen Zusammenhang zwischen der Durchführung einer bestimmten Therapie und einer Befindlichkeitsverbesserung bei den Patienten gibt. Zur Ermittlung eines solchen Zusammenhanges könnte etwa der Anteil der nach der Therapie geheilten Patienten mit dem entsprechenden Anteil der im selben Zeitraum unbehandelt gebliebenen und trotzdem geheilten Patienten verglichen werden. Ist die Heilungsquote bei therapierten Patienten höher als in der Kontrollgruppe der unbehandelten Patienten, so würden wir dazu neigen, diese Therapie als wirksam zu betrachten. Doch eine solch einfache Betrachtungsweise kann manche Überraschung bergen, wie das folgende Beispiel zeigt:

Beispiel 1: Betrachten wir eine experimentelle Untersuchung zur Wirksamkeit einer bestimmten Therapie. Das Merkmal *Teilnahme an einer Therapie* hat zwei Ausprägungen (*ja, nein*), ebenso wie das Merkmal *Heilung* (*ja, nein*). *Teilnahme an einer Therapie* wird als *Treatment*- und *Heilung* als *Responsevariable* bezeichnet. An der Untersuchung mögen insgesamt 100 Patienten teilnehmen. Die Hälfte von ihnen wird therapiert (Therapiebedingung), die anderen bleiben unbehandelt (Kontrollbedingung). Am Ende des Therapiezeitraumes werden die folgenden Ergebnisse ermittelt:

		Heilung	
		ja	*nein*
Therapie	*ja*	*30*	*20*
	nein	*30*	*20*

Tabelle 1: *Ergebnisse der n=100 Patienten. Die Zahlen geben die Häufigkeit der Patienten unter den entsprechenden Bedingungen an. So wurden insgesamt 50 Patienten therapiert, 30 davon waren anschließend geheilt. Insgesamt haben therapierte und nicht-therapierte Patienten die gleichen Heilungsquoten von 30/50, also 60%.*

Die Zahlen in Tabelle 1 geben jeweils die Häufigkeiten an, mit denen Patienten mit der entsprechenden Merkmalskombination auftauchen. So gab es z. B. 30 therapierte und geheilte Patienten sowie 20 therapierte und nicht geheilte Patienten. Liest man die Daten der Tabelle zeilenweise, so kann man die Ergebnisse der therapierten und der nicht-therapierten Patienten vergleichen.

Auf den ersten Blick zeigt sich kein Unterschied zwischen Therapie- und Kontrollbedingung: Unter beiden Bedingungen beträgt die Heilungsquote 30/50, also 60%. Demnach scheint es egal zu sein, ob eine Therapie durchgeführt wird oder nicht: Auf den ersten Blick ist der Effekt der Therapie gleich Null. Man spricht in diesem Zusammenhang auch von dem sogenannten *Prima Facie Effekt* (Effekt auf den ersten Blick).

Ist damit das letzte Wort über die Wirksamkeit der Therapie gesprochen? Ein neugieriger Forscher betrachtet diese Ergebnisse noch einmal im Detail. Für alle Patienten wurde nämlich zu Beginn der Untersuchung zusätzlich die *Therapiemotivation* (in den zwei Ausprägungen *motiviert* und *nicht-motiviert*) gemessen. Der Zusammenhang von *Therapie* und *Heilung* wird nun getrennt für motivierte und nicht-motivierte Patienten untersucht. Dabei zeigt sich folgendes Ergebnis:

Motivierte Patienten (n=60)

		Heilung	
		ja	*nein*
Therapie	*ja*	*12*	*6*
	nein	*26*	*16*

Nicht-motivierte Patienten (n=40)

		Heilung	
		ja	*nein*
Therapie	*ja*	*18*	*14*
	nein	*4*	*4*

Tabelle 2: *Ergebnisse getrennt nach motivierten und nicht-motivierten Patienten: Therapierte Patienten haben in beiden Gruppen höhere Heilungsquoten als nicht-therapierte Patienten.*

Das Ergebnis ist verblüffend. Sowohl bei den motivierten als auch bei den nicht-motivierten Patienten ist die Therapiebedingung die erfolgreichere Bedingung. Wohlgemerkt, es handelt sich um die gleichen Daten wie in Tabelle 1, nur aufgeteilt in zwei Teilgruppen. Fügt man die Daten wieder zu einer Tabelle zusammen, so ist diese mit Tabelle 1 identisch. Bei den motivierten Patienten beträgt die Heilungsquote in der Therapiebedingung 12 von insgesamt 18, also 66.7%, während nicht-therapierte Patienten nur eine Heilungsquote von 26/42, also 61.9% aufweisen. Bei den nicht-motivierten Patienten ergibt sich ein ähnliches Bild: Die Heilungsquote unter der Therapiebedingung beträgt 18/32, also 56.3%, während nicht-therapierte Patienten nur eine Heilungsquote von 4/8, also 50% aufweisen. In beiden Teilgruppen ist die Therapiebedingung die erfolgreichere Bedingung, nur in der Gesamtgruppe aller Patienten nicht. Oder pointiert ausgedrückt: Die Therapie wirkt bei motivierten und bei nicht-motivierten Patienten, aber nicht bei Patienten allgemein. Wie ist dieses paradoxe Phänomen zu erklären?

Ein frühes Beispiel für solch paradoxe Phänomene stammt von H. E. Simpson (1951), man spricht seitdem vom *Simpson-Paradoxon*. Bevor wir formal dieses Phänomen zu fassen versuchen, soll zunächst inhaltlich überlegt werden, wie es zu einer Situation wie der oben beschriebenen kommen

kann. In der im Beispiel beschriebenen Untersuchung spielt die Therapiemotivation der Patienten eine zentrale Rolle. Zwischen ihr und der Zuordnung der Patienten zur Therapie gab es einen Zusammenhang: Motivierte Patienten wurden seltener therapiert (18 von 60, also 30%) als nicht-motivierte Patienten (32 von 40, also 80%), entsprechend überproportional häufig finden sich motivierte Patienten in der Kontrollgruppe. Außerdem hängt die Heilungsquote unter beiden Experimentalbedingungen von der Motivation ab. Tabelle 2 zeigt genau dieses Phänomen: Je nach Ausprägung der *Therapiemotivation* sind die Heilungsquoten unterschiedlich. Auf diese Weise kommen zwei 'Effekte' der Therapiemotivation zusammen. Zum einen gibt es einen Zusammenhang mit der Treatmentvariable, zum

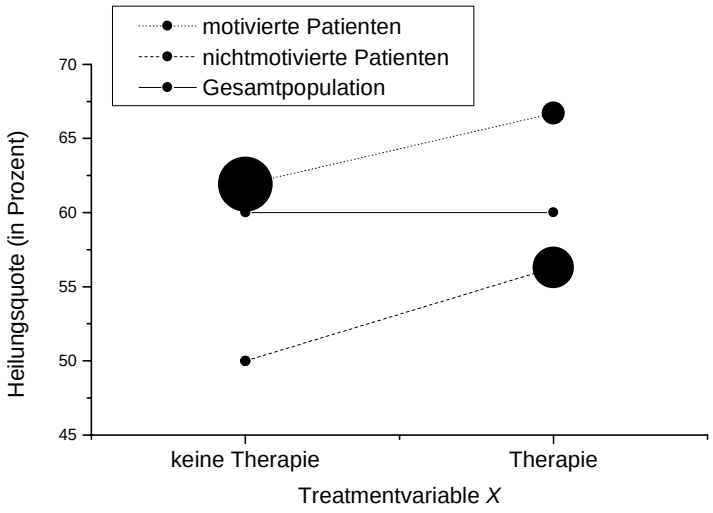

Abbildung 1: Veranschaulichung des Zusammenhangs zwischen der Heilungsquote, der Behandlungsform (Therapie vs. keine Therapie) und der Therapiemotivation der Patienten aus Beispiel 1.

anderen ändert sich der Zusammenhang zwischen Treatment- und Responsevariable in Abhängigkeit von der Therapiemotivation (die motivierten Patienten haben unter beiden Behandlungsbedingungen höhere Heilungsquoten als die nicht-motivierten Patienten). Betrachtet man aber die Ergebnisse beider Patientengruppen zusammengefasst, so führt dies in diesem Beispiel dazu, dass auf den ersten Blick kein Zusammenhang zwischen *Therapie* und *Heilung* sichtbar ist.

Abbildung 1 veranschaulicht noch einmal den zugrundeliegenden Mechanismus. Die gestrichelte Linie gibt die Heilungsquoten in Abhängigkeit von der Behandlungsbedingung für die Gruppe der nicht-motivierten Patienten wieder, die gepunktete Linie die entsprechenden Ergebnisse der motivierten Patienten. Wie man deutlich sehen kann, unterscheiden sich sowohl in der Therapie- als auch in der Kontrollgruppe die Heilungsquoten der motivierten Patienten von denen der nichtmotivierten Patienten; insofern verändert also die Variable *Therapiemotivation* den Zusammenhang zwischen der Responsevariablen *Heilungsquote* und der Treatmentvariable *Therapie*. Die durchgezogene Linie zeigt die Heilungsquoten für die beiden Behandlungsbedingungen in der Gesamtpopulation. Die unterschiedlich großen Kreise veranschaulichen den Zusammenhang zwischen der Therapiemotivation und der Zuordnung zur Therapie- bzw. Kontrollgruppe. Die Größe der Kreise ist dabei proportional zur Anzahl der Personen des jeweiligen Typs gewählt (also z. B. proportional zur Anzahl der nichtbehandelten, motivierten Patienten). Hier kann man erkennen, dass motivierte Patienten eher keine Therapie erhielten, während der größere Teil der nicht-motivierten Patienten in der Therapiebedingung zu finden ist. Daraus resultieren deutlich unterschiedliche Zusammensetzungen der Therapie-

und der Kontrollgruppe. Bei der Zusammenfassung der Ergebnisse, um den Zusammenhang zwischen Heilung und Behandlung in der Gesamtpopulation zu beschreiben, erhält dadurch in der Therapiegruppe die Heilungsquote der nicht-motivierten Patienten ein größeres Gewicht, in der Kontrollgruppe der unbehandelten Patienten „dominiert“ demgegenüber die Heilungsquote der motivierten Patienten. Im Endeffekt resultieren daraus gleiche Heilungsquoten der Therapie- und der Kontrollgruppe in der Gesamtpopulation.

Damit wird deutlich, was unter einer *Konfundierung* zu verstehen ist: Betrachtet man den Zusammenhang zweier Variablen (hier *Therapie* und *Heilung*), so bedeutet eine Konfundierung, dass eine dritte Variable (hier *Therapiemotivation*) einerseits mit der Treatmentvariable zusammenhängt, und andererseits der Zusammenhang von Treatment- und Responsevariable von dieser dritten Variable abhängt. Die Konsequenz ist, dass die Wirksamkeitsbeurteilung eines Treatments allein aufgrund des Prima Facie Effektes zu ganz anderen Ergebnissen kommt, als wenn die konfundierende Variable (im Beispiel die Therapiemotivation) mit berücksichtigt wird. Daher ist die Berücksichtigung konfundierender Variablen enorm wichtig, wenn man kausale Aussagen über die Wirksamkeit eines Treatments machen möchte.

Man mag zu diesem Beispiel einwenden, dass es doch recht unrealistisch sei. In der Praxis werden vermutlich eher motivierte Patienten systematisch in der Therapiebedingung und unmotivierte Patienten systematisch häufiger in der Kontrollgruppe zu finden sein. Insbesondere außerhalb experimenteller Untersuchungen, etwa in therapeutischen Praxen, werden vermutlich motivierte Patienten eher einen Therapieplatz bekommen als nicht-motivierte Patienten. Der Leser möge sich überlegen, was dies für die kausale Interpretierbarkeit von empirisch ermittelten Heilungsquoten bedeutet. So kann ein vorhandener Prima Facie Effekt, der scheinbar die Wirksamkeit einer Therapie zeigt, lediglich aufgrund einer Konfundierung zustande kommen. Diese Überlegungen weisen eindringlich darauf hin, dass das Aufspüren eventuell vorhandener Konfundierung eine herausragende Bedeutung bekommt, wenn empirische Befunde kausal interpretiert werden. Allerdings sollen Konzepte wie *kausale Interpretierbarkeit* oder *kausale Effekte* in diesem Beitrag nicht vertiefend diskutiert werden (vergleiche hierzu z. B. die Arbeiten von Nachtigall et. al., 1999, 2000; Pearl, 2000; Rubin, 1974; Steyer et al., 2000 a), sondern wir beschränken uns auf das Konzept der Konfundierung.

Die weiteren Abschnitte dieses Beitrags gliedern sich wie folgt: Zunächst soll das inhaltliche Beispiel aus der Einführung formal erfasst werden. Dazu ist es notwendig, genau anzugeben, was als *Zusammenhang* zweier Merkmale verstanden werden soll. Anschließend wird der Begriff der Konfundierung genauer präzisiert. Sodann werden Möglichkeiten des Testens von Konfundierung aufgezeigt und Bedingungen diskutiert, die Konfundierung ausschließen. Daran schließt sich ein Abschnitt an, in dem der Bezug des hier verwendeten Konfundierungsbegriffs zu anderen verwandten methodischen Konzepten aufgezeigt wird. Im letzten Abschnitt werden die wesentlichen Ergebnisse zusammengefasst und diskutiert.

2 Zusammenhänge von Variablen

Der zentrale Begriff in unseren bisherigen Überlegungen lautet *Zusammenhang*. In der empirischen Forschung sind wir an möglichst allgemeinen Zusammenhängen interessiert. Auch bei dem inhaltlichen Therapiebeispiel (Beispiel 1) ging es um Zusammenhänge zwischen unabhängiger und abhängiger Variable. Von Konfundierung wurde gesprochen, wenn die unabhängige Variable (im Beispiel die *Teilnahme an einer Therapie*) einerseits von einer dritten Variable (im Beispiel die *Motivation* der Patienten) "abhängt" und wenn andererseits der "Zusammenhang" von UV und AV durch diese dritte Variable "beeinflusst" wird, diese also einen "Effekt" darauf hat. An diesen sprachlichen Formulierungen wird

deutlich, dass es in der Umgangssprache eine Menge mehr oder minder synonyme Formulierungen für das Sprachfeld der Zusammenhänge und Abhängigkeiten von Variablen gibt. Aber was genau ist damit gemeint?

In den nachfolgenden Unterabschnitten werden wir verschiedene Möglichkeiten vorstellen, wie der Begriff „Zusammenhang“ präzisiert werden kann. Wir behandeln dabei die Korrelation, die stochastische Abhängigkeit und die regressive Abhängigkeit. Diese verschiedenen Arten der Abhängigkeit werden dabei anhand von Beispielen illustriert. Da sowohl die stochastische Abhängigkeit als auch die regressive Abhängigkeit in der Konfundierungsanalyse eine wichtige Rolle spielen, werden diese beiden Formen des Zusammenhangs mit Hilfe des einführenden Therapiebeispiels veranschaulicht: Wie wir später sehen werden, spielt der Begriff der stochastischen Abhängigkeit eine Rolle, wenn es um die Beschreibung des Zusammenhangs zwischen der unabhängigen Variablen *Teilnahme an einer Therapie* und der dritten Variablen *Motivation* geht, während die sogenannte *regressive Abhängigkeit* bei der Beschreibung des Zusammenhangs zwischen der abhängigen Variablen *Heilung* einerseits und der unabhängigen sowie der dritten Variablen andererseits ins Spiel kommt.

2.1 Korrelation und stochastische Abhängigkeit

Oft wird der Zusammenhang zweier Variablen vereinfacht mit der Korrelation dieser Variablen gleichgesetzt. Allerdings ist ein Korrelationskoeffizient lediglich eine spezielle Kennzahl für spezielle Zusammenhänge. So sind Situationen denkbar, in denen Variablen unkorreliert sind, aber dennoch ein enger Zusammenhang zwischen ihnen besteht. Dies wird im nachfolgenden Beispiel gezeigt.

Beispiel 2: Sie wetten mit jemandem auf das nächste Spielresultat einer Fußballmannschaft. Sieg und Niederlage mögen gleich wahrscheinlich sein. Im Falle eines Sieges gewinnen Sie eine Flasche Wein ($X=1$), im Falle einer Niederlage verlieren Sie eine Flasche Wein ($X=-1$), bei einem Unentschieden gibt es keinen Gewinn oder Verlust ($X=0$). Die Zufallsvariable X beschreibt also Ihren Wettgewinn bzw. Verlust. Nun beschließen Sie und Ihr Wettgegner, dass, egal wer gewinnt, der Wein in jedem Fall gemeinsam getrunken und gerecht aufgeteilt wird, wobei jeder genau eine halbe Flasche trinken soll ($Y=0.5$). Nur bei einem Unentschieden gibt es keinen Wein ($Y=0$). Die Zufallsvariable Y beschreibt die letztlich resultierende 'Ausschüttung' der Wette. Zwischen X und Y besteht dann ein Zusammenhang. Aufgrund des Wertes von X läßt sich sogar exakt vorhersagen, welchen Wert Y hat. Ist $X=0$, so ist $Y=0$. Ist $X=1$ oder $X=-1$, so ist $Y=0.5$. Es gibt einen deterministischen Zusammenhang zwischen dem Ausgang der Wette und der anschließenden Trinkmenge.

Aber die Variablen X und Y sind unkorreliert, wie man beim Nachrechnen feststellen kann[1]: Benützen wir die Abkürzung $p=P(X=-1)=P(X=1)$. Es ist $Kor(X,Y)=Cov(X,Y)/Std(X)Std(Y)$ und es gilt $Cov(X, Y) = E(XY) - E(X)\,E(Y)$. Nun ist $E(XY)= -1\cdot0.5\cdot p+0\cdot0\cdot(1-2p)+1\cdot0.5\cdot p = 0$ und $E(X)=1\cdot p+0\cdot(1-2p)-1\cdot p=0$, also gilt $Kor(X,Y)=0$.

[1] *Eine explizite Berechnung einer Korrelation findet man z. B. in Nachtigall & Wirtz (1998), S.79 u. 80.*

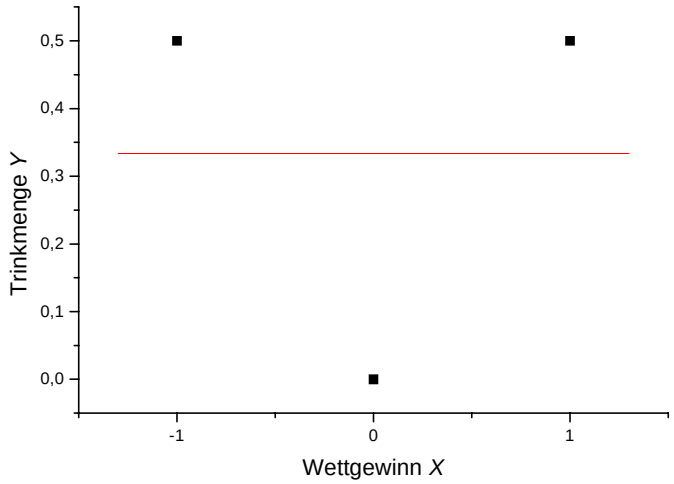

Abbildung 2: Der Zusammenhang zwischen der Trinkmenge Y und dem Wettgewinn X in Beispiel 2. Die Punkte zeigen die jeweilige Trinkmenge an; die Gerade veranschaulicht den linearen Zusammenhang zwischen Trinkmenge und Wettgewinn.

Das Beispiel zeigt, dass es irreführend sein kann, sich bei der Beschreibung von Zusammenhängen allein auf die Betrachtung von Korrelationen zu beschränken. Eine Korrelation ist ein Kennwert, der die Enge des *linearen* Zusammenhanges zweier Variablen angibt. Im Wettbeispiel handelt es sich aber um einen speziell gewählten nicht-linearen Zusammenhang (also ein Zusammenhang, den man nicht durch eine Gerade beschreiben kann, siehe Abbildung 2), bei dem die Korrelation Null war. Aus dem Umstand, dass eine Korrelation zwischen zwei Variablen gleich Null ist, kann man demnach nicht darauf schließen, dass zwischen den Variablen kein Zusammenhang besteht. Um Zusammenhänge zu erfassen und zu beschreiben, werden also allgemeinere Begriffe als die Korrelation benötigt.

Der allgemeinste Begriff eines Zusammenhangs von Zufallsvariablen ist der Begriff der *stochastischen Abhängigkeit* bzw. *Unabhängigkeit*. Beschränken wir uns der Einfachheit halber auf diskrete Variablen X und W, deren Werte jeweils positive Wahrscheinlichkeit haben. Von *stochastischer Unabhängigkeit* von X und W wird dann gesprochen, wenn das Eintreten eines Ereignisses $W=w$ nichts an der Wahrscheinlichkeit eines Ereignisses $X=x$ ändert, wenn also $P(X=x \mid W=w) = P(X=x)$ für alle Werte von W und X gilt. *Stochastische Abhängigkeit* herrscht also dann, wenn es Werte x von X und w von W gibt, so dass $P(X=x \mid W=w) \neq P(X=x)$. Erinnern wir uns an Beispiel 1, wo es um die Wirksamkeit einer Therapie ging: X ist die Treatmentvariable *Teilnahme an einer Therapie* und W die Drittvariable *Therapiemotivation.* Betrachten wir die 100 untersuchten Personen als gesamte Population und wählen eine dieser Personen rein zufällig aus, dann können die relativen Häufigkeiten aus Tabelle 1 und 2 als Wahrscheinlichkeiten interpretiert werden. Wie man an diesen relativen Häufigkeiten sehen kann (vgl. die Zahlenwerte in Tabelle 2), ist

$$P(X=Therapie \mid W=motiviert) = 18/60 = 0.3, \quad (1)$$

während $P(X=Therapie)=0.5$ ist. Da damit $P(X=Therapie \mid W=motiviert) \neq P(X=Therapie)=0.5$ gilt, sind *Therapie* und *Motivation* also stochastisch abhängig. Als Häufigkeiten ausgedrückt: Von den 60 motivierten Patienten wurden nur 18, also 30%, therapiert, insgesamt werden aber 50% aller Patienten therapiert. Allerdings drückt stochastische Abhängigkeit zunächst keinerlei Richtung aus. Die Tatsache, dass die bedingte Wahrscheinlichkeit für die Teilnahme an der Therapie sich in (1) von der unbedingten Wahrscheinlichkeit unterscheidet, muss nicht bedeuten, dass die Motivation in irgendeinem kausalen Sinn die Teilnahme an der Therapie bestimmt. Umgekehrt unterscheidet sich

len Sinn die Teilnahme an der Therapie bestimmt. Umgekehrt unterscheidet sich nämlich auch die bedingte Wahrscheinlichkeit für motivierte Patienten unter der Therapiebedingung von der unbedingten Wahrscheinlichkeit für motivierte Patienten, wie man in Gleichung (2) sieht. Es gibt insgesamt 50 therapierte Patienten, von denen 18 motiviert waren (siehe Tabelle 2), es ist also

$$P(W=motiviert \mid X=Therapie) = 18/50 = 0.36 \neq 0.6 = P(W=motiviert). \quad (2)$$

Man könnte daher auch mutmaßen, dass die Therapie kausalen Einfluß auf die Motivation hat. Das kann in vielen praktischen Fällen zutreffen. Die Erfahrung erster erfolgreicher Veränderungsschritte in einer Therapie dürfte die Therapiemotivation verstärken. Aber im Beispiel wurde die Motivation *vor* Beginn der Therapie erhoben. Somit scheidet eine solche kausale Interpretation von vornherein aus.

Grundsätzlich ist also festzuhalten: Stochastische Abhängigkeit von X und W bedeutet, dass in Abhängigkeit der Werte der einen Variable es zu *irgendwelchen* Änderungen der Verteilung der anderen Variable kommt. Allerdings kann aufgrund einer stochastischen Abhängigkeit allein nichts über eine Gerichtetheit des Zusammenhanges, schon gar nicht über kausale Einflussrichtungen ausgesagt werden. Man kann nur sagen, dass X und W stochastisch abhängig sind, nicht, dass X von W abhängt. Sind X und W stochastisch abhängig, dann auch umgekehrt W und X. Technisch ausgedrückt: Stochastische Abhängigkeit ist eine symmetrische Relation.

Das Konzept der stochastische Abhängigkeit ist die allgemeinste Möglichkeit, um auszudrücken, dass es einen Zusammenhang zwischen Zufallsvariablen gibt. Um zu beschreiben, wie dieser Zusammenhang konkret aussieht, betrachten wir die bedingten Wahrscheinlichkeiten. In Beispiel 1 interessieren uns z.B. die bedingten Wahrscheinlichkeiten für die verschiedenen Werte von X, wenn wir nur die motivierten Patienten betrachten. Aus den Daten in Tabelle 2 kann man ablesen, dass $P(X=Therapie \mid W=motiviert)=18/60=0.3$ und $P(X=keine\ Therapie \mid W=motiviert)=42/60=0.7$ ist. Abbildung 3 zeigt die unbedingten und die unter den verschiedenen Stufen von W bedingten Wahrscheinlichkeiten für die beiden Treatmentgruppen.

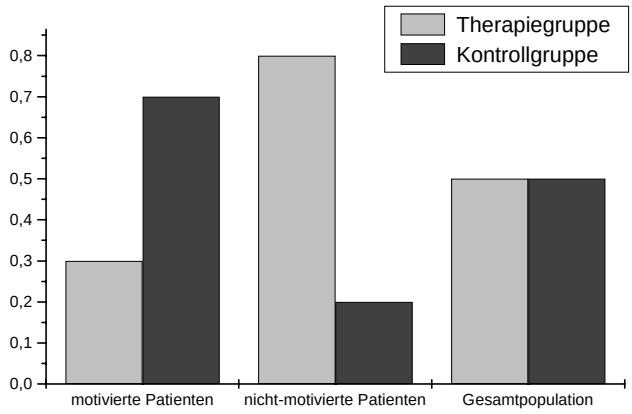

Abbildung 3: Unbedingte und unter den verschiedenen Stufen von W bedingte Wahrscheinlichkeiten für die Werte der Treatmentvariable X.

Damit ist der erste Teil dessen, was eine Konfundierung ausmacht, geklärt: Die Treatmentvariable X und eine Drittvariable W müssen stochastisch abhängig sein. Das reicht allein jedoch noch nicht für eine Konfundierung aus. Zusätzlich zu dieser stochastischen Abhängigkeit muss die Variable W auch den Zusammenhang von Treatment- und Responsevariable "beeinflussen". Dies wird im Folgenden formalisiert.

2.2 Bedingte Erwartung und Regression

Eine stochastische Abhängigkeit ist die allgemeinste Form eines Zusammenhanges von Zufallsvariablen X und W. Sie besagt, dass sich bedingte Wahrscheinlichkeiten *irgendwie* von unbedingten Wahrscheinlichkeiten unterscheiden. Für die Praxis ist das oft zu allgemein. Betrachten wir wiederum das Beispiel einer Therapiestudie. Dort wird üblicherweise nicht eine dichotome Responsevariable (Heilung: ja, nein) erhoben, sondern es wird versucht, das Ausmaß an Therapieerfolg als Responsevariable Y auf einer Intervallskala zu messen. Als Ergebnis einer solchen Untersuchung ist man dann am mittleren Therapieerfolg interessiert. Dazu vergleicht man die Mittelwerte von Y bei therapierten und nicht-therapierten Patienten. Damit kommen wir zur dritten Möglichkeit, was unter einem Zusammenhang von Zufallsvariablen zu verstehen ist; wir betrachten nun die sogenannte *regressive Abhängigkeit* (siehe z. B. Steyer, 1999): Bei diesem Ansatz beschreibt man die Beziehung zwischen zwei Zufallsvariablen, indem man angibt, wie groß die sogenannten *bedingten Erwartungswerte* der einen Zufallsvariable sind in Abhängigkeit vom jeweiligen Wert der anderen Variable. Eine in dieser Weise beschriebene Beziehung bezeichnet man als *bedingte Erwartung* bzw. synonym als *Regression*. Um das Vorgehen zu verdeutlichen, betrachten wir die folgende Situation: Es sei X wieder eine Treatmentvariable (mit den Werten *Therapie* bzw. *keine Therapie*) und Y eine numerische Responsevariable. Dann ist der bedingte Erwartungswert $E(Y \mid X = \textit{Therapie})$ gerade der (theoretische) Mittelwert der Responsevariable Y der therapierten Patienten und $E(Y \mid X = \textit{keine Therapie})$ ist entsprechend der (theoretische) Mittelwert von Y im Falle der nicht-therapierten Patienten, also der Kontrollgruppenbedingung. Er berechnet sich durch

$$E(Y \mid X = x) = \sum_{y} y \cdot P(Y = y \mid X = x) \qquad (3)$$

Die Formel bedeutet folgendes: Als Beispiel betrachten wir nur therapierte Patienten, also $X = \textit{Therapie}$. Zur Bestimmung des bedingten Erwartungswertes $E(Y \mid X = \textit{Therapie})$ werden die Werte der Responsevariablen Y aufsummiert und dabei mit der Wahrscheinlichkeit, mit der diese Werte in der Gruppe der therapierten Patienten auftreten, gewichtet. Ist $X = \textit{Therapie}$, dann sind diese Wahrscheinlichkeiten gerade die bedingten Wahrscheinlichkeiten $P(Y = y \mid X = \textit{Therapie})$. Auf diese Weise erhält man den theoretischen Mittelwert von Y unter der Bedingung $X = \textit{Therapie}$. Analog verfährt man zur Bestimmung von $E(Y \mid X = \textit{keine Therapie})$, dem theoretischen Mittelwert von Y für die nicht-therapierten Patienten. Diese theoretischen Mittelwerte können in empirischen Untersuchungen aufgrund von Stichprobenmittelwerten geschätzt werden. Ihre Differenz ist dasjenige, was den Forscher interessiert. Die (möglichen) Unterschiede werden bei der statistischen Auswertung eines Experimentes üblicherweise mit t-Test oder Varianzanalyse getestet. Alle bedingten Erwartungswerte gemeinsam, im Beispiel also $E(Y \mid X = \textit{Therapie})$ und $E(Y \mid X = \textit{keine Therapie})$, machen die sogenannte *Regression* aus; sie wird zusammenfassend mit $E(Y \mid X)$ bezeichnet.

Der Begriff der bedingten Erwartung bzw. Regression stellt also eine nützliche Formalisierung dessen dar, was bei einem Großteil der empirischen Sozialforschung Gegenstand ist: Das Zurückführen theoretischer Mittelwerte einer Variablen Y auf verschiedene Ausprägungen einer Variablen X. Unterscheiden sich die Erwartungswerte der Responsevariablen Y in Abhängigkeit von der Treatmentvariablen X, dann ist Y *regressiv abhängig* von X. Unterscheiden sich die Erwartungswerte nicht, dann ist Y regressiv unabhängig von X.

Betrachten wir erneut Beispiel 1. Eigentlich kann dort kein bedingter Erwartungswert berechnet werden, denn die Responsevariable Y nimmt nur die "Werte" Heilung und Nicht-Heilung an. Wie soll darüber ein Mittelwert gebildet werden? Dies ist allerdings mit einem bestimmten Trick möglich und sinnvoll: Nämlich dann, wenn man Y bei Heilung mit 1 und bei Nicht-Heilung mit 0 kodiert. In diesem Falle ist nämlich

$$E(Y \mid X = x) = 1 \cdot P(Y = 1 \mid X = x) + 0 \cdot P(Y = 0 \mid X = x) = P(Y = 1 \mid X = x),$$

es sind also die bedingten Erwartungswerte $E(Y \mid X=x)$ und die bedingte Wahrscheinlichkeit $P(Y=1 \mid X=x)$ identisch. $E(Y \mid X=x)$ ist dann gerade die Heilungsquote unter der Bedingung $X=x$. Die regressive Abhängigkeit der Responsevariable von der Treatmentvariable ist hier äquivalent mit der stochastischen Abhängigkeit der beiden Zufallsvariablen. Dies gilt jedoch nur in diesem Spezialfall, wo Y lediglich die Werte 0 und 1 annimmt. Abbildung 1 zeigt die Regression von Y auf X für unser Therapiebeispiel: Sie entspricht der durchgezogenen Linie.

Generell ist die stochastische Abhängigkeit der allgemeinste Begriff für Zusammenhänge von Variablen, regressive Abhängigkeit ist eine spezielle Form von stochastischer Abhängigkeit und eine von Null verschiedene Korrelation ist wiederum ein Spezialfall regressiver Abhängigkeit. Beispiel 2 war ein Beispiel für zwei Variablen, die zwar unkorreliert, aber trotzdem regressiv und damit auch stochastisch abhängig sind. Umgekehrt gilt: Die stochastische Unabhängigkeit von zwei Zufallsvariablen X und Y impliziert deren regressive Unabhängigkeit und diese wiederum impliziert deren Unkorreliertheit.

Festzuhalten ist, dass auch eine regressive Abhängigkeit von Y und X zunächst keinerlei kausale Beeinflussung von Y durch X bedeuten muss, sondern dass es sich lediglich um eine formale und präzise Beschreibung dessen handelt, was unter Zusammenhängen von Variablen verstanden wird.

2.3 Zum Umgang mit bedingten Erwartungswerten

Im vorangegangenen Abschnitt haben wir die Beschreibung des Zusammenhangs mit Hilfe einer Regression für zwei Zufallsvariablen Y (*Heilung*) und X (*Teilnahme an einer Therapie*) kennen gelernt. Dies ist eine noch sehr einfache Situation, und in vielen empirischen Anwendungen interessiert man sich für komplexere Abhängigkeitsbeziehungen zwischen drei und auch mehr Zufallsvariablen. In Tabelle 2 unseres Eingangsbeispiels geht es z. B. darum zu beschreiben, wie Y (*Heilung*) sowohl von der Behandlung X (*Therapie* vs. *keine Therapie*) als auch von der Motivation W (*motiviert* vs. *nicht-motiviert*) abhängt. Der Regressionsansatz kann auch auf solche Situationen ausgedehnt werden: Man schreibt dann $E(Y \mid W, X)$. Diese Schreibweise zeigt an, dass nun untersucht bzw. angegeben werden soll, wie sich die bedingten Erwartungswerte von Y in Abhängigkeit von X und W verändern. In unserem Beispiel lassen sich vier solche bedingten Erwartungswerte bestimmen. Aus Tabelle 2 können wir entnehmen, dass

$$E(Y \mid \textit{motiviert, Therapie}) = P(Y=1 \mid \textit{motiviert, Therapie}) = 12/18 = 0.67,$$

man beachte, dass $Y=1$ gerade für Heilung steht. Entsprechend sind

$$E(Y \mid \textit{nicht-motiviert, Therapie}) = 18/32 = 0.56.$$
$$E(Y \mid \textit{motiviert, keine Therapie}) = 26/42 = 0.62$$
$$E(Y \mid \textit{nicht-motiviert, keine Therapie}) = 4/8 = 0.5.$$

Alle vier bedingten Erwartungswerte gemeinsam ergeben die Regression $E(Y \mid X,W)$: Wir kennen damit die Heilungsquoten in den Subpopulationen der motivierten und der nicht-motivierten Patienten jeweils sowohl für die Therapie- als auch für die Kontrollbedingung. Kann man daraus die Heilungsquote für die Therapie- bzw. die Kontrollbedingung in der gesamten Population berechnen? Formal betrachtet fragen wir damit jetzt nach der Regression $E(Y \mid X)$, die aus den beiden bedingten Erwartungswerten $E(Y \mid X = \textit{Therapie})$ und $E(Y \mid X = \textit{keine Therapie})$ besteht. Die Antwort auf obige Frage lautet "ja". Um z. B. den bedingten Erwartungswert $E(Y \mid X = \textit{Therapie})$ zu berechnen, nehmen wir die

beiden bedingten Erwartungswerte $E(Y \mid motiviert, Therapie)$·und $E(Y \mid nicht\text{-}motiviert, Therapie)$·aus den Teilpopulationen und gewichten diese mit "ihren" Wahrscheinlichkeiten, also mit der Wahrscheinlichkeit, dass unter der Therapiebedingung Patienten motiviert sind bzw. nicht motiviert sind.

$$\begin{aligned} E(Y \mid Therapie) &= E(Y \mid motiviert, Therapie)\cdot P(motiviert \mid Therapie) \\ &\quad + E(Y \mid nicht\text{-}motiviert, Therapie)\cdot P(nicht\text{-}motiviert \mid Therapie) \\ &= 0.67 \cdot 18/50 + 0.56 \cdot 32/50 = 0.6 \end{aligned}$$

Der bedingte Erwartungswert von Y unter einer Bedingung $X=x$ kann also aus den entsprechenden bedingten Erwartungswerten in Subpopulationen berechnet werden. Schreibt man das in Form einer Gleichung, so lautet diese

$$E(Y \mid X = x) = \sum_{w} E(Y \mid W = w, X = x) \cdot P(W = w \mid X = x) \qquad (4)$$

Diese Gleichung ist, wie wir im nächsten Abschnitt sehen werden, von besonderer Wichtigkeit für die Konfundierungsanalyse. Man kann zeigen, dass sie immer richtig ist[2].

3 Konfundierung einer Regression

Nun sind wir in der Lage, das Phänomen der Konfundierung genau zu erfassen. Haben wir festgelegt, was wir unter einer Konfundierung verstehen, dann können wir in den nächsten Schritten zum einen nach Wegen suchen, wie wir prüfen können, ob eine Konfundierung vorliegt. Zum anderen können wir möglicherweise angeben, unter welchen Bedingungen sichergestellt werden kann, dass *keine* Konfundierung vorliegt.

Rekapitulieren wir zunächst noch einmal das Phänomen aus dem Eingangsbeispiel. Wir haben uns dort für den Zusammenhang zwischen dem Heilungserfolg Y und der Teilnahme an einer Therapie X interessiert. Wurde dieser Zusammenhang in der Gesamtpopulation analysiert (siehe Tabelle 1), dann ergaben sich gleiche Heilungsquoten sowohl für die Gruppe der Therapieteilnehmer als(W, X) auch für die Gruppe der unbehandelten Patienten. Untersuchte man den Zusammenhang dagegen getrennt in den Subpopulationen der motivierten und der nicht-motivierten Patienten, zeigte sich *in beiden Fällen* die Therapiebedingung als die überlegene. Die Analyseergebnisse sind gerade deshalb überraschend, weil sich das Ergebnis für die Gesamtpopulation noch nicht einmal ergibt, wenn man die Teilergebnisse aus den Subpopulationen mittelt und dabei berücksichtigt, dass es mehr motivierte als nicht-motivierte Patienten gibt. Dieses Phänomen, das wir als Konfundierung bezeichnet haben, wird nun in der nachfolgenden Definition formal beschrieben: Ausgangspunkt ist dabei die Beschreibung des Zusammenhangs zwischen einer Responsevariable Y und einer Treatmentvariable X in der Gesamtpopulation mit Hilfe der Regression $E(Y \mid X)$, die auch als *Treatmentregression* bezeichnet wird, da es um die Beschreibung der Abhängigkeit von einem Treatment (im Beispiel: *Therapie* bzw. *keine Therapie*) geht. Es wird dann genau festgelegt, unter welchen Bedingungen man davon spricht, dass diese Regression bezüglich einer dritten Variablen konfundiert ist.

[2] *Die Gültigkeit dieser Gleichung ergibt sich aus den allgemeinen Regeln zum Rechnen mit bedingten Erwartungswert (vgl. z.B. Schmitz, 1983, S. 146, Steyer & Eid, 1993, S.357). Folgt man der Notation aus Steyer & Eid (1993), so ergibt sich Gleichung (4) als Spezialfall von Rechenregel v. aus Box G.1. Das dortige* X *entspricht hier dem Variablenvektor und* $f(W,X)$ *entspricht* X.

Bevor wir zu der Definition kommen, soll aber zunächst noch festgelegt werden, was für "dritte Variablen" *W* wir betrachten wollen. Wir bezeichnen solche Variablen als *potenzielle Störvariablen.* In der Konfundierungsanalyse sind potenzielle Störvariablen solche Variablen, die Eigenschaften von Personen beschreiben. Ein Beispiel ist das Geschlecht einer Person, eine andere Störvariable ihre soziale Schichtzugehörigkeit, in obigem Beispiel ist es die *Therapiemotivation.* Potenzielle Störvariablen teilen die Population in Subpopulationen (z.B. motivierte und nicht-motivierte Patienten). Formal lässt sich das wie folgt ausdrücken: Wenn in einer empirischen Untersuchung die Variable *U* angibt, welche Person (technisch: welche *Unit*) untersucht wird, dann sollen unter potenziellen Störvariablen *W* beliebige Funktionen von *U* verstanden werden[3]. Man spricht dabei von einer *potenziellen* Störvariablen, da man noch nicht weiß, ob eine solche Variable *W* tatsächlich eine Konfundierung „bewirkt". Nun ist zu klären, wann eine solche potenzielle Störvariable zu einer konfundierenden Variable wird.

Definition 1: *Sei $E(Y|X)$ eine Regression und W eine potenzielle Störvariable. Dann heißt $E(Y|X)$ unkonfundiert hinsichtlich der potenziellen Störvariable W, wenn für alle Werte x von X*

$$E(Y|X=x) = \sum_{w} E(Y|W=w, X=x)\cdot P(W=w) \qquad (5)$$

erfüllt ist. Die Regression $E(Y|X)$ und die potenzielle Störvariable W heißen konfundiert, wenn Gleichung (5) nicht gilt.

Diese Formel bedarf der Erläuterung. Sie besagt, dass die bedingten Erwartungswerte $E(Y|X=x)$ eine Art "mittlerer Wert" der entsprechenden bedingten Erwartungswerte in den Subpopulationen sein müssen. "Gemittelt" wird dabei mit den Wahrscheinlichkeiten der Subpopulationen $W=w$. Betrachten wir wieder das Eingangsbeispiel. Die Regression $E(Y|X)$ beschreibt dort, wie die Heilungsquote in der gesamten Population von der Behandlung *X* (*Therapie* vs. *keine Therapie*) abhängt, und besteht aus den beiden bedingten Erwartungswerten $E(Y|X=Therapie)$ und $E(Y|X=keine\ Therapie)$, den Heilungsquoten für die Therapie- und Kontrollbedingung. Beide bedingten Erwartungswerte betragen jeweils 0.6. In den Subpopulationen der motivierten bzw. nicht-motivierten Patienten waren die Heilungsquoten unter der Therapiebedingung aber jeweils höher als unter der Kontrollbedingung. Es liegt eine Konfundierung vor. "Mitteln" wir nämlich gemäß Gleichung (5) die bedingten Erwartungswerte aus den Subpopulationen, so erhalten wir

$$E(Y|motiviert, Therapie)\cdot P(W=motiviert) + E(Y|nicht\text{-}motiviert, Therapie)\cdot P(W=nicht\text{-}motiviert)$$
$$= 0.67 \quad 0.6 \; + 0.56 \cdot 0.4 = 0.623 \text{ und}$$
$$E(Y|motiviert, k.\ Therapie)\cdot P(W=motiviert) + E(Y|nicht\text{-}motiviert, k.\ Ther.)\cdot P(W=nicht\text{-}motiviert)$$
$$= 0.56 \quad 0.6 \; + 0.50 \cdot 0.4 = 0.54.$$

Diese Werte stimmen *nicht* mit den Werten von $E(Y|X)$ überein, denn in der Gesamtpopulation gilt ja $E(Y|Therapie) = 0.6 = E(Y|keine\ Therapie)$. Es liegt also mit der Therapiemotivation *W* eine Konfundierung der Treatmentregression $E(Y|X)$ vor. Dadurch ist es möglich, dass der Prima Facie Effekt Null ist, obwohl in den Subpopulationen die Heilungsquoten bei Therapie höher sind als unter der Kontrollgruppenbedingung.

Mit Definition 1 ist festgelegt, was eine konfundierende Variable sein soll. Wenn es keine solche konfundierende Variable gibt, spricht man von (allgemeiner) Unkonfundiertheit:

[3] *Die Variable U ist selbst ebenfalls eine potenzielle Störvariable.*

Definition 2: *Eine Regression $E(Y \mid X)$ heißt unkonfundiert, wenn sie unkonfundiert ist hinsichtlich jeder potenziellen Störvariable W.*

Die Unkonfundiertheit einer Regression bedeutet also, dass es keine konfundierende Variable *W* gibt. Ist eine Regression unkonfundiert, dann ist sichergestellt, dass so verblüffende Phänomene wie das Simpson Paradoxon nicht auftreten können. Denn im Falle der Unkonfundiertheit ist der Effekt eines Treatments in der Gesamtpopulation immer eine Art (gemäß Gleichung (5) berechneter) "Mittelwert" der Effekte in den Subpopulationen. Zeigt sich ein Treatment in allen Subpopulationen als überlegen, so kann dieser Effekt nicht in der Gesamtpopulation verschwinden, wie es im Beispiel 1 gerade der Fall war. Gilt die Unkonfundiertheit, dann können Zusammenhänge einer Treatmentregression kausal interpretiert werden. Gibt es aber eine Konfundierung, dann kann ein Prima Facie Effekt verfälscht sein, kausale Interpretationen sollten nicht gemacht werden. Die Aufgabe des nächsten Abschnittes ist es daher, genauer zu beleuchten, wann eine Konfundierung vorliegt.

3.1 Bedingungen für Konfundierung

Wir haben anhand des Eingangsbeispiels versucht, uns inhaltlich klar zu machen, was eine Konfundierung ist. Betrachtet man den Zusammenhang zweier Variablen (dort *Therapie* und *Heilung*), so bedeutet eine Konfundierung, dass eine dritte Variable (dort *Therapiemotivation*) einerseits mit der Treatmentvariable zusammenhängt, und andererseits der Zusammenhang von Treatment- und Responsevariable von dieser dritten Variable abhängt. In Abschnitt 3 wurde dann formal der Konfundierungsbegriff eingeführt. Im Falle einer Konfundierung einer Treatmentregression gibt es eine potenzielle Störvariable *W*, für die Gleichung (5) nicht gilt. Wie passt das zusammen? Zunächst ist keineswegs deutlich, dass der formale Begriff eine Formalisierung und Präzisierung unserer inhaltlichen Fragestellung bedeutet.

Das dem doch so ist, wird deutlich, wenn wir eine andere, äquivalente Bedingung für die Konfundierung einer Treatmentregression betrachten:

Konfundierungssatz: Eine Treatmentregression $E(Y \mid X)$ ist genau dann konfundiert, wenn es eine Treatmentbedingung x und einen Wert w einer Störvariable W gibt, so dass

i.: *die Ereignisse $X=x$ und $W=w$ stochastisch abhängig sind*

und

ii.: $E(Y \mid X=x) \neq E(Y \mid X=x, W=w)$ *gilt.*

Der Beweis findet sich im Anhang.

Gemäß dem Konfundierungssatz entspricht und präzisiert also eine Konfundierung unsere ursprüngliche inhaltliche Deutung des Phänomens: Konfundierung bedeutet, dass es eine Störvariable gibt, von der die Treatmentvariable *X* stochastisch abhängig ist (Bedingung i.), zusätzlich muss die Störvariable den Zusammenhang von Treatment- und Responsevariable "verändern" (Bedingung ii.). Mit "verändern" ist in diesem Kontext allerdings nicht gemeint, dass der Effekt der Therapie auf die Heilungsquote sich in Abhängigkeit von der Therapiemotivation ändert, also beispielsweise für die motivierten Patienten die Heilungsquote der Therapiebedingung deutlich größer ist als die der Kontrollbedingung, während sich für die nichtmotivierten Patienten die Heilungsquoten der beiden Behandlungsbedingungen nicht unterscheiden; dies bezeichnet man in der Statistik als *Interaktion* bzw. *Moderation*. Viel-

mehr meinen wir im Rahmen der Konfundierungsanalyse mit "verändern" lediglich, dass innerhalb einer Treatmentbedingung x die Werte der Regression nicht alle konstant gleich $E(Y \mid X=x)$ sind, sondern sich in Abhängigkeit von w unterscheiden, also „verändern“. Dies ist eine andere Form, wie die Beziehung zwischen einer Treatment- und einer Responsevariable durch eine Störvariable verändert werden kann und sollte nicht mit einer Interaktion verwechselt werden! Auf die Unterschiede zwischen Konfundierung und Interaktion werden wir noch genauer in Abschnitt 3.4 eingehen.

Im Eingangsbeispiel waren beide im Konfundierungsatz genannten Bedingungen erfüllt: Es wurden einerseits vermehrt nicht-motivierte Patienten therapiert, also $P(X=\textit{Therapie} \mid W=\textit{nicht-motiviert}) \neq P(X=\textit{Therapie})$, die Ereignisse $X=\textit{Therapie}$ und $W=\textit{nicht-motiviert}$ sind stochastisch abhängig. Andererseits unterscheidet sich die Heilungsquote bei nicht-motivierten Patienten von der Heilungsquote insgesamt, also $E(Y \mid W=\textit{nicht-motiviert}, X=\textit{Therapie}) \neq E(Y \mid X=\textit{Therapie})$.

Es gibt noch weitere äquivalente Bedingungen für die Konfundierung einer Treatmentregression (vgl. Steyer et al. 2000 b). Für unseren Zweck einer Einführung reicht aber der Konfundierungssatz in der hier wiedergegebenen Fassung völlig aus.

3.2 Das Testen von Konfundierung

Eine Konfundierung hat große praktische Bedeutung für die Interpretierbarkeit von Effekten. Vor diesem Hintergrund ist es entscheidend, dass Konfundierung kein rein theoretisches Konstrukt bleibt, sondern dass anhand von Daten praktisch darüber entschieden werden kann, ob eine Konfundierung vorliegt. Wie läßt sich das Vorhandensein oder nicht Vorhandensein von Konfundierung empirisch überprüfen? Der nächstliegende Weg wäre es, mit Hilfe der Definition eine Überprüfung zu versuchen. Anhand eines Datensatzes wäre zu testen, ob $E(Y \mid X=x)$ und $\Sigma_w E(Y \mid W=w,\ X=x)\ P(W=w)$ übereinstimmen (vgl. Gleichung (5)). Es können die bedingten Erwartungswerte $E(Y \mid X=x)$ und $E(Y \mid W=w,\ X=x)$ sowie die Wahrscheinlichkeiten $P(W=w)$ aus Stichprobendaten geschätzt werden. Jedoch gibt es kein statistisches Standardverfahren, um den Test selbst durchzuführen. Der Grund ist folgender: In der Regel können Testprobleme, bei denen geschätzte Größen multiplikativ eingehen (hier z. B. der Ausdruck $E(Y \mid W=w, X=x) \cdot P(W=w)$) nur mit sogenannten asymptotischen Tests getestet werden, was große Stichproben erfordert und Ungenauigkeiten beim Signifikanzniveau bedeutet (vgl. von Davier, 2000). Daher ist es einfacher, die im Konfundierungssatz aufgeführten Bedingungen zur empirischen Überprüfung heranzuziehen. Demnach ist folgendes zu tun. Zunächst ist ein Kandidat für eine Konfundierung, nämlich eine potenzielle Störvariable W zu suchen. Dann ist Bedingung i. zu überprüfen: Es ist zu testen, ob die Treatmentvariable X und W stochastisch abhängig sind. Dies kann z.B. mit einem χ^2 -Test gemacht werden. Liegt stochastische Abhängigkeit vor, dann ist Bedingung ii. zu überprüfen, also die Gleichheit von Mittelwerten. Dies kann z. B. im Rahmen des Allgemeinen Linearen Modells durchgeführt werden. Zu einer ausführlicheren Darstellung von Testmöglichkeiten vergleiche Nachtigall & Steyer (in Vorbereitung). An dieser Stelle beschränken wir uns auf die Darstellung der grundsätzlichen Ideen zur Testung von Konfundierung.

3.3 Wann kann eine Konfundierung ausgeschlossen werden?

Wenn Effekte kausal interpretiert werden sollen, dann ist eine vorliegende Konfundierung ein großes Problem. Der vorangegangene Abschnitt zeigt Wege auf, wie überprüft werden kann, ob tatsächlich eine Konfundierung vorliegt. Doch kann man für diese Vorgehen natürlich nur diejenigen potenziellen Störvariablen berücksichtigen, die tatsächlich erhoben wurden. Angenommen, im Eingangsbeispiel

wäre nur die Treatmentvariable *Therapie* und die Responsevariable *Heilung* gemessen worden. Es kann dann nicht ausgeschlossen werden, dass es eine nicht-beobachtete konfundierende Variable wie die Therapiemotivation gibt. Eine letztendliche Überprüfung der Unkonfundiertheit ist auf diesem Wege grundsätzlich nicht möglich: Ob eine potenzielle Störvariable eine konfundierende Variable ist, kann nur geprüft werden, wenn diese Variable auch tatsächlich mit erhoben wurde. Daher ist es sinnvoll, sich *vor* der Durchführung einer Untersuchung Gedanken über mögliche konfundierende Variablen zu machen und diese mit zu erheben.

Anderseits kann man in einer Untersuchung jedoch aus praktischen Gründen nicht beliebig viele Variablen erheben. In der Therapiestudie kann man sicherlich eine große Anzahl von potenziellen Störvariablen zusammenstellen, aber man wird grundsätzlich nicht jede potenzielle Störvariable erfassen können, da hierfür jede beliebige Eigenschaft einer Person in Frage kommt. Aus diesem Grund wäre es wünschenswert, eine Vorgehensweise zu finden, welche eine Konfundierung gänzlich ausschließt. Eine solche Vorgehensweise ist durch die Technik der *randomisierten Zuweisung zum Treatment* gegeben.

Weist man die Personen in einem Experiment rein zufällig den verschiedenen Treatmentbedingungen zu, dann ist die Personenvariable U stochastisch unabhängig von X. Man kann nun zeigen, dass mit U auch jede Funktion W von U, also jede potenzielle Störvariable, stochastisch unabhängig von der Treatmentvariable ist. Jede mögliche Konfundierung ist damit ausgeschlossen. Bedingung i. aus dem Konfundierungssatz ist niemals erfüllt.

Eine Randomisierung in einem kontrollierten Experiment ermöglicht es, eine systematische Variation der Responsevariable allein auf die Variation der Treatmentvariable zurückzuführen (Huber, 1987, S.99). Eine Konfundierung kann ausgeschlossen werden. Bei der Interpretation wird von einem kausalen Einfluß der unabhängigen Variable auf die abhängige Variable gesprochen. Die Variable X wird als Ursache für Y angesehen. Allerdings besteht die Möglichkeit zur Randomisierung auch nur bei Untersuchungen, bei denen die Zuweisung zum Treatment X unter der Kontrolle des Versuchsleiters steht (also bei kontrollierten "echten" Experimenten, z. B. nicht bei Feldexperimenten). In anderen Untersuchungen, wo verschiedene Stufen einer Treatmentvariable lediglich beobachtet werden, ist dies nicht möglich. Hier muss daher mit der in Abschnitt 3.2 gezeigten Strategie versucht werden, eventuell vorhandene Konfundierungen zu ermitteln. Es besteht dann die Möglichkeit, diese Konfundierung durch geeignete Berechnungen "zum Verschwinden" zu bringen. Diese Verfahren der Adjustierung werden an anderer Stelle eingehend besprochen (vgl. Wüthrich-Martone et al., 1999; Nachtigall et al., 2000).

3.4 Konfundierung, Interaktion, Moderation und Co.

Der Einfluss einer Drittvariable auf den Zusammenhang zwischen einer Treatment- und einer Responsevariablen ist in den empirischen Wissenschaften ein wichtiges Thema. Dabei muss es sich keineswegs immer um Konfundierung handeln. Viel häufiger findet man in der Literatur die Konzepte "Interaktion" und "Moderation". Beide Begriffe bedeuten das gleiche, Interaktion wird dabei meist im Zusammenhang mit Varianzanalysen gebraucht, Moderation im Zusammenhang mit Regressionsanalysen. Im Falle einer linearen Regression bedeutet eine Moderation unterschiedliche *Steigungen* von $E(Y \mid X, W=w)$ in den verschiedenen Subpopulationen w. Steyer (1999) spricht in diesem Fall von bedingt reg-linearer Abhängigkeit, die Variable W wird von ihm auch als *Modifikator* bezeichnet. Zur Illustration dient das folgende Beispiel:

Beispiel 3: In Abbildung 4 ist eine Moderation bzw. Interaktion grafisch dargestellt. Es geht wiederum um den Zusammenhang zwischen Heilungserfolg Y (*geheilt* versus *nicht geheilt*) und Behandlungs-

form *X* (*Therapie* versus *keine Therapie*). Dargestellt sind abermals die Ergebnisse für zwei verschiedene Subpopulationen, nämlich einmal für die motivierten Patienten, zum anderen für die nicht-motivierten Patienten. Die Moderation bzw. Interaktion ist daran erkennbar, dass die Linien, die den regressiven Zusammenhang zwischen dem Heilungserfolg *Y* und der Behandlungsform *X* beschreiben, nicht parallel verlaufen, mithin der Effekt der Behandlungsform in den Subpopulationen unterschiedlich ist. Die Variable *Motivation* moderiert den Zusammenhang von *Y* und *X*. In diesem Beispiel sind wie in unserem Eingangsbeispiel die Zahlen so gewählt, dass sich bei Zusammenfassung der Ergebnisse in der Gesamtpopulation kein Unterschied zwischen Therapie- und Kontrollgruppe bezüglich der Heilungsquoten ergibt. Die Gleichheit der Heilungsquoten in der Gesamtpopulation (durchgezogene Linie in Abbildung 4) kommt aber auf eine andere Art und Weise zustande als im Eingangsbeispiel:

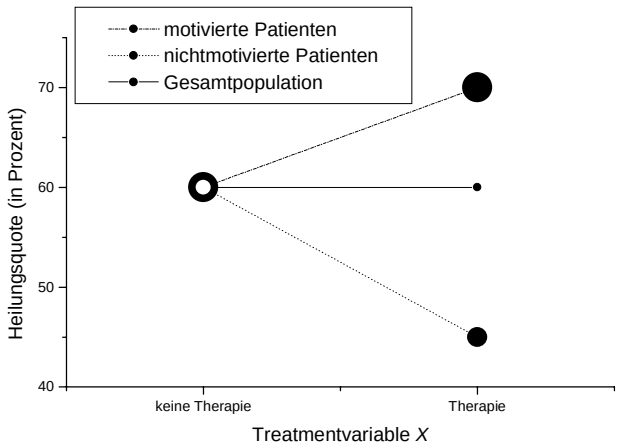

Abbildung 4: Veranschaulichung einer Interaktion. Unterschiedliche Effekte in verschiedenen Subpopulationen zeigen sich dadurch, dass die Linien im Diagramm nicht parallel sind.. Es liegt trotzdem keine Konfundierung vor, da Motivation und Treatment stochastisch unabhängig sind.

Die gegenläufigen Effekte in beiden Subpopulationen heben sich bei diesem Beispiel einer Moderation wechselseitig auf. Während bei motivierten Patienten die Heilungsquote unter der Therapiebedingung deutlich größer ist als unter der Kontrollbedingung, kehren sich die Verhältnisse bei nicht-motivierten Patienten um. Das Beispiel wurde jedoch so konstruiert, dass *keine* Konfundierung der Regression $E(Y \mid X)$ durch die Störvariable *W* (*Therapiemotivation*) vorliegt. Anders als in Beispiel 1, wo nicht-motivierte Patienten vermehrt Therapie machten und motivierte vermehrt in die Kontrollgruppe kamen, sind in diesem Beispiel die Motivation und die Treatmentzuweisung stochastisch unabhängig. In der Abbildung kommt dies dadurch zum Ausdruck, dass die Kreise der motivierten Patienten für beide Behandlungsbedingungen gleich groß sind (Die Größe der Kreise ist dabei wie in Abbildung 1 proportional zur Anzahl der Personen dieses Typs gewählt). Es resultieren gleichartige Zusammensetzungen der Therapie- und Kontrollgruppe, und in der Gesamtpopulation heben sich die gegenläufigen Effekte aus beiden Subpopulationen gegenseitig auf. Bei einer Konfundierung ist dagegen die stochastische Abhängigkeit zwischen der Störvariablen *W* und der Treatmentzuweisung *X* ein entscheidendes Merkmal; die Effekte der Behandlung können dabei in den durch die Werte von *W* gebildeten Subpopulation durchaus gleich sein. Wie das Eingangsbeispiel zeigte, ist es sogar möglich, dass eine Konfundierung auftritt, ohne dass eine Interaktion bzw. Moderation zwischen *W* und *X* vorliegt: In Abbildung 1 verlaufen die Linien, die den Zusammenhang zwischen Heilungsquote und Behandlungsform veranschaulichen, nahezu parallel, mithin liegt keine Interaktion zwischen *W* und *X* in ihrer

Wirkung auf die Heilungsquote Y vor. Die Begriffe „Interaktion" bzw. „Moderation“ und „Konfundierung“ beschreiben damit unterschiedliche Phänomene. In unseren Zahlenbeispielen ist zwar der Effekt in der Gesamtpopulation in beiden Fällen der gleiche, dass nämlich kein Unterschied zwischen Therapie- und Kontrollbedingung erkennbar ist; dieses Gesamtergebnis kommt aber aufgrund zweier verschiedener Wirkmechanismen zustande. Im Falle einer Konfundierung ist von einer kausalen Interpretation dringend abzuraten. Im Fall einer Interaktion bzw. Moderation kann eine kausale Interpretation durchaus sinnvoll sein, sofern keine zusätzliche Konfundierung vorliegt.

Was hat Moderation bzw. Interaktion mit Konfundierung zu tun? Im allgemeinen gibt es keinen Zusammenhang der beiden Begriffe. Eine Regression kann durch eine Variable W konfundiert sein, ohne dass W ein Moderator ist. In Beispiel 1 waren die Steigungen der Regression in den beiden Subpopulationen annähernd gleich. Umgekehrt war Beispiel 3 gerade so konstruiert, dass zwar eine Moderation, aber keine Konfundierung vorlag. Zwar verlangt eine Konfundierung eine Änderung der Regression $E(Y \mid X)$ in Abhängigkeit von einer Drittvariablen W, diese muss aber nicht die Steigung der Regression betreffen.

4 Diskussion

In den vorangegangenen Abschnitten haben wir anhand einfacher Beispiele den Konfundierungsbegriff eingeführt. Um noch einmal zusammenzufassen: Von einer *Konfundierung* spricht man dann, wenn es eine Störvariable W gibt, die einerseits mit der Treatmentzuweisung X zusammenhängt und andererseits den regressiven Zusammenhang zwischen der Outcome-Variablen Y und dem Treatment X verändert. Unter diesen Bedingungen tritt das im Eingangsbeispiel 1 illustrierte Phänomen auf: Der Treatmenteffekt in der Gesamtpopulation entspricht nicht mehr dem Mittel der Effekte in den durch W gebildeten Subpopulationen und vermittelt damit einen falschen Eindruck hinsichtlich der Wirksamkeit der betreffenden Behandlung. Im Eingangsbeispiel sind die Zahlenwerte dabei so gewählt, dass in beiden Subpopulationen, den motivierten und den nicht-motivierten Patienten, die Heilungsquoten der Patienten der Therapiegruppe jeweils größer waren als die der unbehandelt gebliebenen Patienten. Dennoch ergaben sich in der Gesamtpopulation gleiche Heilungsquoten für die Therapie- und die Kontrollgruppe. Aus den beiden genannten Bedingungen ergeben sich unmittelbar Prüfmöglichkeiten, ob für eine bestimmte Störvariable W eine Konfundierung der Treatmentregression $E(Y \mid X)$ vorliegt (siehe Abschnitt 3.2), bzw. unter welchen Bedingungen eine Konfundierung generell ausgeschlossen werden kann (siehe Abschnitt 3.3). Das Verhältnis von Konfundierung zu ähnlichen Begriffen wie Interaktion und Moderation wurde in Abschnitt 3.4 beleuchtet.

Das Problem der Konfundierung ist auch in anderen Disziplinen wie der Epidemiologie seit längerem bekannt und wurde ausführlich diskutiert (vgl. Breslow & Day, 1980, p. 95; Mc Neil, 1996, p. 13). Allerdings widersetzte es sich lange einer formalen Erfassung. Bei unserer Einführung in die Konfundierungsanalyse handelt es sich um eine spezielle Formalisierung von Konfundierung. Auf der Basis einer eigenen kausaltheoretischen Ausgangsposition wurde das Konzept der Konfundierung ebenfalls von Pearl (1998) formalisiert und Kriterien für Konfundierung daraus abgeleitet. Bisher steht ein Vergleich der kausaltheoretischen Ansätze von Pearl und Steyer noch aus. Eine abschließende Bewertung der verschiedenen Formalisierungen von Konfundierung muss noch vorgenommen werden, soll aber nicht Gegenstand dieser Einführung sein.

5 Anhang: Beweis des Konfundierungssatzes

Zunächst formulieren wir den Konfundierungssatz durch logische Verneinung äquivalent um: Er lautet dann: Eine Treatmentregression ist genau dann unkonfundiert, wenn für alle Werte x von X und für alle Werte w jeder potenziellen Störvariablen W zumindest eine der Bedingungen

$$\text{i': } P(X=x \mid W=w)=P(X=x) \text{ oder ii': } E(Y \mid X=x) = E(Y \mid W=w, X=x)$$

erfüllt ist[4]. Dies wird im Folgenden bewiesen.

Die Regression sei unkonfundiert und es seien x, W und w gegeben. Zu zeigen ist, dass Bedingung i' oder Bedingung ii' erfüllt ist. Wir betrachten die potenzielle Störvariable $I_{W=w}$, also die Indikatorvariable für $W=w$. Nach Voraussetzung gilt hinsichtlich $I_{W=w}$,

$$E(Y \mid X = x) = E(Y \mid I_{W=w} = 1, X = x)P(I_{W=w} = 1) + E(Y \mid I_{W=w)} = 0, X = x)P(I_{W=w} = 0) \quad (6)$$

und gemäß (4) gilt

$$E(Y \mid X = x) = E(Y \mid I_{W=w} = 1, X = x)P(I_{W=1} = 1 \mid X = x) + E(Y \mid I_{W=w)} = 0, X = x)P(I_{W=w} = 0 \mid X = x) \quad (7)$$

Zur Abkürzung und Vereinfachung schreiben wir: $p_1 := P(I_{W=w}=1)$, $p_2 := P(I_{W=w}=1 \mid X=x)$, $a := E(Y \mid I_{W=w}=1, X=x)$, $b := E(Y \mid I_{W=w}=0, X=x)$, $c := E(Y \mid X=x)$. Dann lauten die Gleichungen (6) und (7)

$c = ap_1 + b(1-p_1)$

$c = ap_2 + b(1-p_2)$.

Subtrahieren wir die zweite Gleichung von der ersten, so erhalten wir

$$0 = a(p_1-p_2) + b(1-p_1-1+p_2)$$

$\Leftrightarrow b(p_1-p_2) = a(p_1-p_2)$

Daraus folgt, dass entweder $p_1-p_2=0$ oder $a=b$ sein muss, und damit $P(W=w \mid X=x) = P(W=w)$ oder $E(Y \mid X=x) = E(Y \mid I_{W=w}=1, X=x) = E(Y \mid W=w, X=x)$, also gilt Bedingung i. oder Bedingung ii..

Nun zur Gegenrichtung des Beweises: Es soll für alle x und alle potenziellen Störvariablen W mit Werten w Bedingung i'. oder Bedingung ii'. gelten. Wir betrachten beliebige aber feste x und W. Zu zeigen ist, dass

$$E(Y \mid X=x) = \sum_{w} E(Y \mid W = w, X = x) \cdot P(W = w)$$

gilt. Dazu fassen wir diejenigen Werte w zu Teilmengen der Wertemenge von W zusammen, die jeweils nur Bedingung i'. oder Bedingung ii'. genügen.

$$A := \{w: P(X=x \mid W=w)=P(X=x)\},\ B := \{w: E(Y \mid X=x) = E(Y \mid W=w, X=x) \wedge w \notin A\}$$

[4] *Dabei geht die Tatsache ein, dass* $P(X=x \mid W=w)=P(X=x) \Leftrightarrow P(W=w \mid X=x)=P(W=w)$. *Dies gilt, wenn die Wahrscheinlichkeit aller Ereignisse* $X=x$ *und* $W=w$ *positiv ist, was wir zu Beginn dieses Artikels vorausgesetzt haben.*

Auf diese Weise erhalten wir disjunkte Mengen, die nach Voraussetzung den gesamten Wertebereich von W umfassen, es gilt also $P(I_A=1)+P(I_B=1)=1$.[5]

Wir betrachten

$$\sum_w E(Y|W=w, X=x)\cdot P(W=w)$$

$$= \sum_{w\in A} E(Y|W=w, X=x)\cdot P(W=w) + \sum_{w\in B} E(Y|W=w, X=x)\cdot P(W=w)$$

$$= \sum_{w\in A} E(Y|W=w, X=x)\cdot P(W=w|X=x) + \sum_{w\in B} \mathrm{E}(\mathrm{Y}|\mathrm{X}=\mathrm{x})\cdot P(W=w) \qquad (8)$$

Weiter ist

$$\sum_{w\in B} P(W=w) = P(I_B=1).$$

Nun ist $I_A=1$ stochastisch unabhängig von $X=x$, denn

$$P(I_A=1|X=x) = P(\bigcup_{w\in A}\{W=w\}|X=x) = \sum_{w\in A} P(W=w|X=x) = \sum_{w\in A} P(W=w) = P(I_A=1).$$

Da $B=A^c$, so ist $I_B=1$ stochastisch unabhängig von $X=x$ und es gilt

$$P(I_B=1)=P(I_B=1\,|\,X=x) = \sum_{w\in B} P(W=w|X=x).$$

Damit kann (8) umgeformt werden zu

$$\sum_{w\in A} E(Y|W=w, X=x)\cdot P(W=w|X=x) + \sum_{w\in B} E(Y|X=x)\cdot P(W=w|X=x)$$

$$= \sum_w E(Y|W=w, X=x)\cdot P(W=w|X=x)$$

$$= E(Y\,|\,X=x).$$

Die letzte Gleichheit ist gerade die allgemeingültige Gleichung (4). Damit ist der Beweis vollständig.

6 Literatur

Breslow, N. & Day, N. (1980*): The analysis of case-control studies. Statistical analysis in cancer research*. Lyon: International Agency for research on cancer.

Huber, O. (1987): *Das psychologische Experiment: Eine Einführung*. Bern: Huber.

Mc Neal, D. (1996): *Epidemiological research methods*. New York: Wiley.

Nachtigall, C. & Wirtz, M. (1998):Wahrscheinlichkeitsrechnung und Inferenzstatistik. Weinheim: Juventa.

Nachtigall, C., Wüthrich-Martone, O. & Steyer, R. (1999). Was wirkt? Kausale Effekte in der Psychotherapieforschung. In Krampen, G., Zeyer H., Schönpflug, W. & Richardt, G. (Eds.) *Beiträge zur Angewandten Psychologie* (pp. 101-104). Bonn: Deutscher Psychologen Verlag.

[5] *Da wir uns auf diskrete Zufallsvariablen beschränken, sind I_A und I_B messbar.*

Nachtigall, C., Steyer, R. ,& Wüthrich-Martone, O. (2000): Causal effects in empirical research. In: May, M. & Ostermeyer, U. *Perspectives on Causality* (in press).

Nachtigall, C. & Steyer, R. (in Vorbereitung): Das Testen von Konfundierung.

Pearl, J. (1998): Why there is no statistical test for confounding, why many think there is, and why they are almost right. *UCLA*, Cognitive Systems Laboratory, R-256.

Pearl, J. (2000). *Causality - Models, Reasoning and Inference*. Cambridge University Press.

Rubin, D. B. (1974): Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688-701.

Schmitz, N. (1983): *Wahrscheinlichkeitstheorie. Teil 1: Zufallsexperimente. Skripten zur mathematischen Statistik*. Universität Münster.

Simpson, E.H. (1951): The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*, Series B, 13, 238-241.

Steyer, R. & Eid, M. (1993): *Messen und Testen*. Berlin: Springer.

Steyer, R. (1999): Einführung in die Regressionstheorie. Skript am Lehrstuhl für Methodenlehre und Evaluationsforschung der Universität Jena. http://www.uni-jena.de/svw/metheval/publikationen/regskript.html.

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C. & Buhl, T. (2000 a). Causal Regression Models I: Individual and Average Causal Effects. *Methods of Psychological Research-Online.*

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000 b). Causal Regression Models II: Unconfoundedness and Causal Unbiasedness. Submitted to *Methods of Psychological Research-Online*.

von Davier, A. A. (2000)*: Tests of Unconfoundedness in Regression Models with Normally Distributed Variables.* Dissertation am Fachbereich Mathematik der Universität Magdeburg.

Wüthrich-Martone, O., Nachtigall, C. & Steyer, R. (1999). Was wirklich wirkt: Kausale Analyse am Beispiel des Vergleichs verschiedener Therapien. In Krampen, G., Zeyer H., Schönpflug, W. & Richardt, G. (Eds.) *Beiträge zur Angewandten Psychologie* (pp. 104-108). Bonn: Deutscher Psychologen Verlag.