Quasi-experimentelle Untersuchungsmethoden

Hans Werner Bierhoff und Georg Rudinger

Psychologische und sozialwissenschaftliche Fragestellungen lassen sich nur zum Teil in Experimenten beantworten, für deren Auswertung statistische Modelle in großer Zahl entwickelt worden sind (z.B. Cochran & Cox, 1957; Winer, Brown & Michels, 1991). Wenn von Experimenten die Rede ist, setzt man im allgemeinen voraus, daß eine Zufallszuteilung (Randomisierung) der Untersuchungseinheiten (UEn) auf die Versuchsbedingungen durchgeführt wird; ansonsten spricht man von Quasi-Experimenten. Experimente sind darüber hinaus durch eine unvermittelt gesetzte Intervention („*treatment*") charakterisiert, die als „kausal" betrachtet wird für Veränderungen von Merkmalen der UEn, die diesem *treatment* unterzogen wurden. Experimente zeichnen sich also durch Kenntnisse bzgl. der Selektionskriterien für die UEn, bzgl. des Zeitpunktes der Intervention und bzgl. der Situation, wie sie sich ohne Intervention darstellen würde, aus (vgl. Bredenkamp, in diesem Band).

1 Randomisierung und Quasi-Experimente

Feldexperimente oder Sozial-Experimente zeigen in der Regel weniger Standardisierung, weniger Kontrolle und Isolation und länger andauernde, komplexere Interventionen. Entscheidend aber ist, daß bei solchen Untersuchungen die UEn nur selten zufällig den Interventionsbedingungen zugeteilt werden, weil sehr oft eine Selbstselektion vorliegt oder eine Zuteilung nach Verdienst oder Bedürfnis. Einschränkungen der zufälligen Zuteilung können sich aus verschiedenen Gründen ergeben:

(1) Nicht jede interessierende unabhängige Variable (UV) erlaubt eine Zufallszuteilung. Das gilt z.B. für häufig verwendete Merkmale wie Geschlecht und Alter.

(2) Eine Zufallszuteilung ist in manchen Fällen aus ethischen Erwägungen nicht vertretbar. Das gilt z.B. bei der Testung eines neuen vielversprechenden Medikaments, das nach moralischen Kriterien nicht willkürlich einer Zufallsauswahl der Patienten vorenthalten werden kann.

(3) Angewandte Fragestellungen lassen es oft als unpraktisch erscheinen, Experimente durchzuführen. Das gilt z.B. für die Prüfung der Frage, ob eine bestimmte Lehrmethode anderen Lehrmethoden im schulischen Unterricht überlegen ist.

(4) Die Übertragbarkeit experimenteller Befunde auf Alltagssituationen läßt sich am besten in Quasi-Experimenten testen, die größere Repräsentativität der UEn und der Untersuchungskontexte im Hinblick auf die intendierten Anwendungen bieten.

Solange man den Anspruch nicht aufgibt, die Auswirkungen von Interventionen systematisch abschätzen zu wollen, müssen auch für nicht-experimentelle Situationen Designs entwickelt werden, die eindeutige Schlußfolgerungen erlauben. Das zunehmende Interesse an quasi-experimentellen Designs hat nun aber nicht zur Folge,

https://doi.org/10.25521/HQM04

daß der Wert von kontrollierten Experimenten in Frage gestellt wird; wo immer es möglich und zulässig ist, sollte auch bei Interventionen im Felde, bei der Überprüfung sozialer Maßnahmen u.ä. Zufallszuteilung vorgenommen werden (Campbell & Boruch, 1975). Unbeschadet dessen hat man sich aber u.U. mit weiteren als den hinreichend bekannten Bedrohungen der internen Validität auseinanderzusetzen, die alle eine Verwischung der Grenzen zwischen Versuchs- und Kontrollgruppe durch Affizierung der Kontrollgruppe betreffen: Diffusion des *treatments*, kompensatorische Rivalität, kompensatorische Angleichung, Demoralisierung (Cook & Campbell, 1979). Vielleicht intensiver noch als beim Laborexperiment müssen deswegen im Feld Zufallszuteilung, Selektivität und „*treatment implementation*" kontrolliert werden (*monitoring*). Zufallszuteilung im Felde hat also mit ethischen, theoretischen, praktisch-prozeduralen und sozialpsychologischen Problemen zu kämpfen (Cook & Shadish, 1994). Außerdem bezieht sie sich dort sehr oft nicht auf Personen als kleinste UE, sondern auf Schulen, Klassen, Städte, Kurse, Kohorten etc. Die Personen sind in diesen Aggregaten von vorneherein enthalten. Mit der hierarchischen linearen Modellierung (Bryk & Raudenbush, 1992) lassen sich diese Mehrebenenprobleme, die typisch für Felduntersuchungen sind, einer Lösung zuführen.

2 Typologie quasi-experimenteller Pläne

Aus der Erkenntnis heraus, daß Zufallszuteilung auf Experimentalbedingungen nicht ausschließlich über den Wert einer Studie bestimmt, wurden in den letzten 25 Jahren quasi-experimentelle Versuchspläne entwickelt, die sich – Fife-Schaw (1995) folgend – auf drei Grundtypen reduzieren lassen:

(1) *Nichtäquivalente Kontrollgruppenpläne*: Diese Pläne implizieren die Existenz einer Versuchsgruppe und einer Kontrollgruppe, denen die UEn nicht durch Zufallszuteilung zugeordnet werden. Der Vortest gibt Aufschluß darüber, ob schon vor der Intervention Unterschiede in der abhängigen Variablen (AV) zwischen den Gruppen vorgelegen haben. In der Abbildung 1 sind die Meßzeitpunkte mit O bezeichnet,

$$
\begin{array}{ccc}
O_1 & X & O_2 \\
\multicolumn{3}{c}{\text{— — — — — — — — —}} \\
O_1 & & O_2
\end{array}
$$

ABBILDUNG 1. Nichtäquivalenter Kontrollgruppenplan.

während *treatment* bzw. Intervention durch X symbolisiert wird. In der ersten Zeile wird die Versuchsgruppe dargestellt, in der zweiten Zeile die Kontrollgruppe. Die Linie zwischen beiden Gruppen deutet an, daß keine Zufallszuteilung der UEn auf die Bedingungen stattgefunden hat, so daß von Anfang an systematische Unterschiede zwischen den Versuchs- und Kontrollgruppen bestehen können. Die Indizierung mit 1 (O_1) bezieht sich auf den Vortest, die mit 2 (O_2) auf den Nachtest.

(2) *Zeitreihen-Versuchspläne*: Sie beruhen auf nur einer Versuchsgruppe (es fehlt also die Kontrollgruppe), die vor und nach der Intervention mehrfach beobachtet wird, so daß sich Hinweise auf die zeitliche Entwicklung der AV ergeben. In Abbildung 1 würde also die zweite Zeile (Kontrollgruppe) entfallen und anstelle von O_1 bzw. O_2 würden mehrere Vortest- bzw. Nachtestwerte über die Zeit erfaßt.

(3) *Zeitreihen kombiniert mit nichtäquivalentem Kontrollgruppenplan*: Dieses multiple Zeitreihen-Design entspricht dem Plan in Abbildung 1 unter Hinzufügung mehrerer Vortest- und Nachtestwerte in beiden Gruppen. Der Vorteil dieses relativ aufwendigen Ansatzes besteht darin, daß die meisten Bedrohungen der Validität quasi-experimenteller Pläne ausgeschlossen werden können (s. unten).

Zusammenfassend kann festgestellt werden: Ein quasi-experimenteller Versuchsplan beruht darauf, daß eine Zufallszuteilung der UEn auf die Bedingungen des Versuchs (z.B. Versuchsgruppe und Kontrollgruppe) nicht gegeben ist bzw. daß überhaupt nur eine Gruppe (u.U. zu mehreren Meßzeitpunkten) untersucht wird.

3 Ausgewählte quasi-experimentelle Pläne

In diesem Beitrag ist nicht der Raum, die in diesen drei Kategorien subsumierte Vielfalt der quasi-experimentellen Ansätze und Auswertungsstrategien angemessen zu berücksichtigen. Dies gilt insbesondere für Zeitreihen-Designs. Sie stellen den Prototyp eines Designs dar, welches spezifische „Kausalhypothesen" bzgl. Niveau, Verlauf, Varianz der Zeitreihe in Folge der Intervention zu formulieren und zu testen gestattet (zur Auswertung vgl. Harrop & Velicer, 1985, 1990; Krauth, in diesem Band). Einen Sonderfall bilden die querschnittlichen Zeitreihen-Analysen von Simonton (1977) und Swaminathan und Algina (1977). Diese Analysen kommen mit relativ wenigen Vor- und Nachtest-Messungen aus, dies aber nur auf Kosten sehr restriktiver Annahmen über die zeitlichen Abhängigkeiten der Beobachtungen.

Wir gehen jedoch ausführlich auf den wohl am häufigsten benutzten nichtäquivalenten Kontrollgruppenplan ein, der zwar konzeptuelle Schwächen zeigt, aber relativ einfach zu realisieren ist. Ebenfalls behandeln wir die konzeptuell überlegene Regressions-Diskontinuitäts-Analyse (RDA), die als Spezialfall eines nichtäquivalenten Kontrollgruppenplans betrachtet werden kann, bei dem sich die Vortestwerte von Versuchs- und Kontrollgruppe durch bewußte Zuteilung überhaupt nicht oder nur geringfügig überschneiden.

3.1 Nichtäquivalente Kontrollgruppenpläne

Fehlende Zufallszuteilung auf die Versuchsbedingungen legt die Vermutung nahe, daß sich die Gruppen von vorneherein in relevanten Merkmalen systematisch unterscheiden. Daher muß das Auswertungsverfahren die Möglichkeit von systematischen Vortestunterschieden zwischen den Gruppen berücksichtigen. Wir haben dabei das generelle Problem, daß die statistische Kontrolle eigentlich nur zweite Wahl gegenüber dem direkten Ausschluß alternativer Erklärungen via Design ist.

Die Auswertung eines nichtäquivalenten Kontrollgruppenplans läßt sich entscheidend durch die Einbeziehung von Kovariaten (als die insbesondere die Vortests dienen können) verbessern. Kovariaten sind jedoch nur dann nützlich, wenn sie pas-

send gewählt sind, d.h. wenn sie in einem theoretisch spezifizierten, zumindest aber „plausiblen" Verhältnis zur AV stehen. Geeignet gewählte Kovariaten können die *power* des varianzanalytischen Tests wesentlich erhöhen. Es gibt allerdings mehrere gravierende Einwände gegen die Kovarianzanalyse (vgl. Elashoff, 1969). Beispielsweise führt Unreliabilität der Kovariaten zu verzerrten Schätzungen des *Treatment-Effektes* (vgl. schon Lord, 1960). Diesem Problem kann bei hinreichend großen Stichproben u.U. durch Rekurs auf Strukturgleichungsmodelle mit latenten Variablen begegnet werden (vgl. Rietz, Rudinger & Andres, in diesem Band). Diese Modelle geben aber keinen Hinweis auf die Auswirkungen von Selektionsprozessen und ersetzen auch nicht automatisch eine präzise Theorie über das erwartete Ergebnismuster.

Schon Linn und Werts (1977) haben ein allgemeines pfadanalytisches Modell für nichtäquivalente Kontrollgruppenpläne dargestellt und drei Spezialfälle unterschieden: (1) Aufteilung der UEn auf der Basis von Vortestwerten (siehe 3.2), (2) Aufteilung der UEn auf der Basis von Selbstselektion und (3) Aufteilung der UEn aufgrund von Gruppenzugehörigkeiten (z.B. Geschlecht). Während im erstgenannten Fall die Berechung einer Kovarianzanalyse empfohlen wird, bietet sich im zweiten und dritten Fall die Einbeziehung von Paralleltests für die Durchführung des Vortests an und eine Auswertung, die auf kanonische Korrelationsanalysen bzw. auf eine Analyse der Kovarianzstrukturen hinausläuft (vgl. Bierhoff & Rudinger, 1980).

Eine Alternative zur Kovarianzanalyse besteht darin, die Vortestwerte (im Nachhinein) zur Parallelisierung der Versuchsgruppen heranzuziehen (Rubin, 1974); dabei wird versucht, die Zusammensetzung von Versuchs- und Kontrollgruppe anzugleichen, indem eine Anpassung der Gruppen aufgrund ihrer Vortestwerte erfolgt. Der Erfolg der Parallelisierung hängt davon ab, wie ähnlich sich die beiden parallelisierten UEn sind. Was nun die Vor- und Nachteile von Kovariaten und Parallelisierung angeht, so wird folgende Empfehlung gegeben (Rubin, 1974): Wenn die Vortest-Nachtest-Korrelationen unter 0.40 liegen, sollte Parallelisierung bevorzugt werden, wenn sie über 0.60 liegen dagegen die Kovarianzanalyse.

Die Auswertung eines nichtäquivalenten Kontrollgruppenplans kann jedoch nicht nach Maßgabe solcher Regeln schematisch erfolgen. Eine wichtige Überlegung bezieht sich auch darauf, ob die Mitglieder der Versuchs- und der Kontrollgruppe Extremgruppen darstellen. Wenn die Versuchsgruppe z.B. die Schüler mit schlechten Leseleistungen enthält und die Kontrollgruppe die mit guten Leseleistungen, dann liegt eine Bildung von Extremgruppen vor, die es wahrscheinlich macht, daß im Nachtest der Leseleistung Regressionseffekte zur Mitte auftreten. In einem solchen Fall wäre die Berechnung einer Kovarianzanalyse oder eines verwandten Verfahrens aus dem Bereich von Pfadanalysen angemessen. Hingegen sind solche Regressionseffekte weniger zu erwarten, wenn die beiden Gruppen z.B. Schüler von unterschiedlichen Schulen oder aus unterschiedlichen Klassen sind. In diesem Fall bietet es sich an, einfache Differenzwerte zwischen Vor- und Nachtest zu benutzen. Entgegen vielfachen Warnungen vor solchen Veränderungswerten, wenn Individualaussagen beabsichtigt sind, bewähren sie sich für Gruppenaussagen als natürliche Metrik der Veränderung. Die Reliabilität der Differenzwerte ist für Gruppenaussagen in der Regel hinreichend (Gottman & Rushe, 1993; Rogosa, 1988).

Die Interpretierbarkeit auch der sophistiziertesten Auswertung eines nichtäquivalenten Kontrollgruppenplans hängt aber nicht zuletzt von dem spezifischen Er-

gebnismuster ab, das unter Einbeziehung von Vor- und Nachtests in der Versuchs- und Kontrollgruppe beobachtet wird (vgl. Rudinger & Bierhoff, 1980). Dies ist ein theoretisch unbefriedigender Zustand. Wenn sich z.B. eine disordinale Interaktion ergibt (etwa dann, wenn die Versuchsgruppe vor der Intervention niedriger und danach höher als die Kontrollgruppe liegt), ist es eher möglich, weiterreichende kausale Interpretationen zu rechtfertigen, als wenn Experimental- und Kontrollgruppe, die sich schon im Vortest unterscheiden, im Nachtest nur noch weiter auseinanderrücken. Wenn das zuletzt genannte Ergebnismuster beobachtet wird, liegt die rivalisierende Hypothese nahe, daß die Gruppen in Versuchs- und Kontrollbedingung unterschiedlichen Populationen entstammen (Selektionseffekt) und/oder Reifungsprozesse in unterschiedlicher Weise stattgefunden haben.

Dieser „klassische" nichtäquivalente Kontrollgruppenplan kann im Sinne eindeutiger Interpretierbarkeit durch genaue Analyse der Selektionsprozesse, durch Einführung multipler Kontrollgruppen, durch Einführung multipler Vortests vor der Intervention, durch Parallelisieren zur Vermeidung von Gruppendifferenzen (sofern sich kein Regressionseffekt einstellt) verbessert werden (Cook & Shadish, 1994).

3.2 Regressions-Diskontinuitäts-Analyse (RDA)

In vielen Bereichen wird versucht, bestimmte Problemgruppen gezielt anzusprechen bzw. zu fördern. Die Strategie besteht oft darin, Personen, die unterhalb oder oberhalb eines bestimmten Wertes auf der Auswahl- oder Eignungsdimension liegen, in das Programm einzubeziehen. So wäre es etwa vorstellbar, daß nur die schlechten Schüler an der Einrichtung eines Förderunterrichts teilnehmen. Wie kann in solchen Fällen überprüft werden, ob die Fördermaßnahme erfolgreich war?

Zunächst erscheint eine solche Prüfung sehr problematisch, da die Vortestwerte zum Kriterium für die Zuteilung der Personen zu der Versuchsgruppe und zu der Kontrollgruppe gemacht werden, so daß in einer Gruppe die Personen versammelt sind, die hohe Werte aufweisen, und in der anderen Gruppe die Personen, die niedrige Werte erreichen. Wie wir sehen werden, ist die Interpretation der Ergebnisse aus einem solchen Versuch auch nicht unproblematisch. Die RDA aber bietet in vielen Fällen die Möglichkeit, fundierte Aussagen über die Auswirkungen des *treatments* bzw. der Fördermaßnahme zu treffen (Thistlethwaite & Campbell, 1960).

Die RDA läßt sich gut durch ein Beispiel veranschaulichen, das aus dem schulischen Bereich entnommen ist (vgl. Abadzi, 1984): Über eine Gruppe von Schülern stehen beispielsweise Vorinformationen über die Leseleistung zur Verfügung, die auf einer 20-stufigen Skala erfaßt wurde. Nun soll – aufgrund administrativer Beschlüsse – ein Förderprogramm für die Schüler durchgeführt werden, die einen Leistungswert > 10 aufweisen. Für die weitere Planung solcher Fördermaßnahmen interessiert, ob die Maßnahme erfolgreich gewesen ist. Dazu wird nach Beendigung der Fördermaßnahme ein Test der Leseleistung durchgeführt. Dieser Nachtest könnte ein Ergebnismuster aufweisen, welches dem idealtypischen Muster in Abbildung 2 entspricht. Ein Hinweis auf einen *Treatment-Effekt* ist dann gegeben, wenn sich an der Trennungslinie zwischen Versuchs- und Kontrollgruppe eine abrupte Diskontinuität zeigt (Abbildung 2A). Wenn sich die Steigungen in beiden Gruppen unterscheiden (Abbildung 2B), wäre das kein eindeutiger Hinweis auf einen *Treatment-Effekt.*

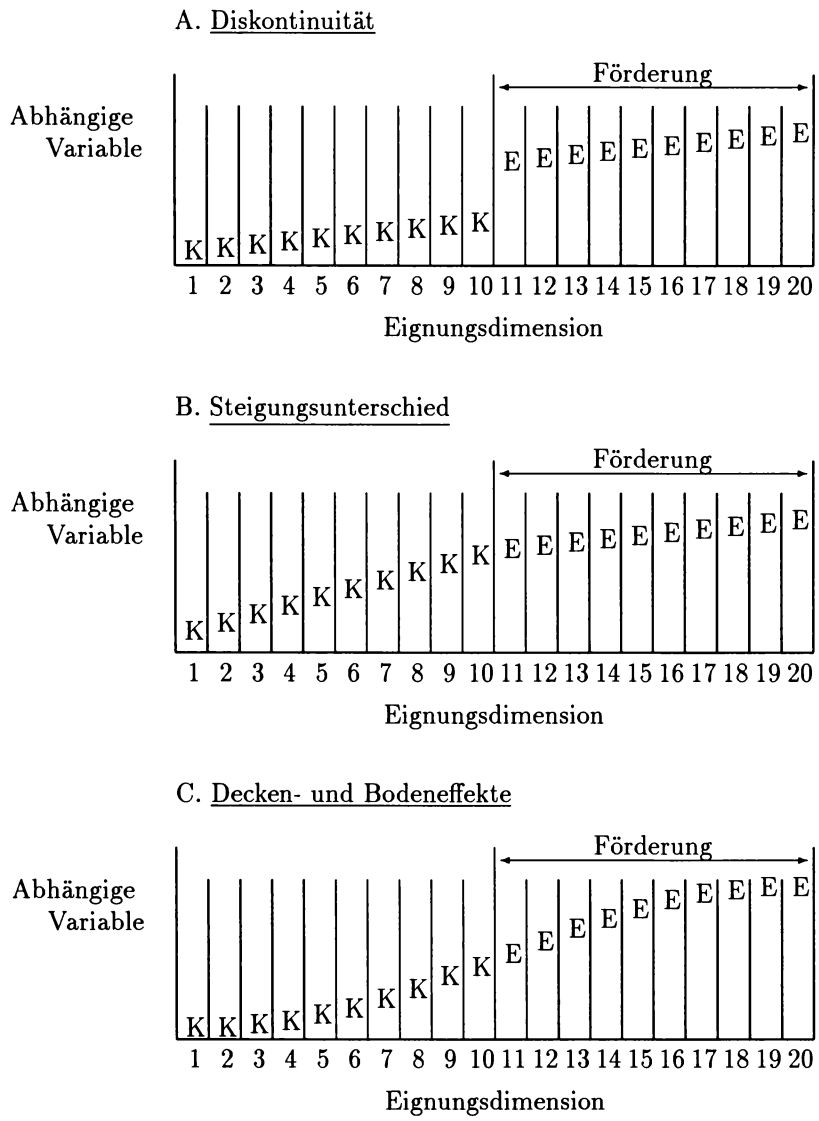

ABBILDUNG 2. **Hypothetische Ergebnismuster (A, B, C) nach einer Fördermaßnahme für ausgewählte Personen. *E* bezeichnet die Regressionslinie in der Versuchsgruppe (>10) und *K* die in der Kontrollgruppe (≤ 10 auf der Eignungsdimension).**

Eine Möglichkeit der Auswertung für diesen Versuchsplan besteht in der Durchführung einer Kovarianzanalyse mit dem Vortest der Leseleistung als Kovariate, wenn die entsprechenden Voraussetzungen erfüllt sind (Linn & Werts, 1977). Neben weiteren Auswertungsalternativen (Stanley, 1991; Trochim, Cappelleri & Reichardt, 1991; Visser & de Leeuw, 1984) ist jedoch die multiple Regressionsanalyse vorzuziehen. Dieses Verfahren hat den Vorteil, daß die unstandardisierten Regressions-

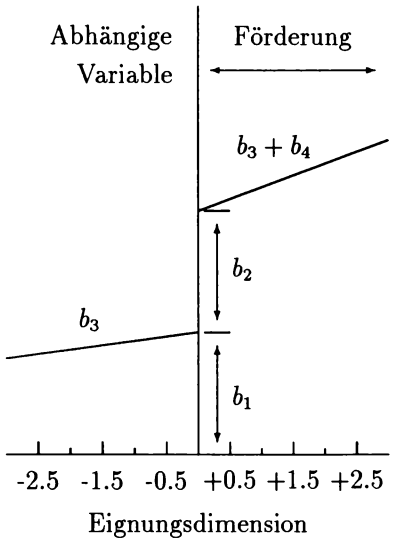

ABBILDUNG 3. Graphische Darstellung des regressionsanalytischen Modells der RDA. Die E-Werte wurden hier so umkodiert, daß in der Bezugskategorie negative und in der Vergleichskategorie positive Werte auftreten. Der Nullpunkt auf der E-Achse entspricht der Trennungslinie zwischen Versuchs- und Kontrollgruppe. Die b-Koeffizienten des regressionsanalytischen Modells beziehen sich auf die Schnittpunkte der Regressionslinie mit der Trennungslinie bzw. auf die Steigungen der Regressionslinien.

gewichte (b) im Hinblick auf Steigungen und Schnittpunkte der Regressionsgeraden von Versuchs- und Kontrollgruppe interpretierbar sind. Das Regressionsmodell für die RDA enthält zwei Prädiktorvariablen und deren Interaktion:

$$Y' = b_1 + b_2T + b_3E + b_4TE.$$

Y' bezeichnet die abhängige bzw. durch die Regressionsgleichung vorhergesagte Variable, E die Vortestwerte auf der Eignungsdimension und T eine Dummy-Variable, die sich auf das Vorhandensein oder Fehlen des *treatments* bzw. der Fördermaßnahme bezieht. In dieser Dummy-Variablen stellt die Null-Kategorie (0) die Bezugskategorie (keine Förderung) dar, die Eins-Kategorie (1) die Vergleichskategorie (Förderung).

Im Anwendungsbeispiel sind in der Bezugskategorie niedrige E-Werte und in der Vergleichskategorie hohe Werte zu finden. Abbildung 3 illustriert das regressionsanalytische Modell der RDA. Ein simuliertes Beispiel mit additivem Zufallsfehler ist in Abbildung 4 dargestellt. Für das Regressionsmodell ergibt sich folgende Gleichung:

$$Y' = 16.13 + 3.88T + 2.15E - 0.20TE.$$

t-Tests zeigen, daß die Koeffizienten b_2 und b_3 signifikant werden ($p < 0.01$). Da b_2 den *Treatment-Effekt* abbildet, ergibt sich die Schlußfolgerung auf einen signifikanten

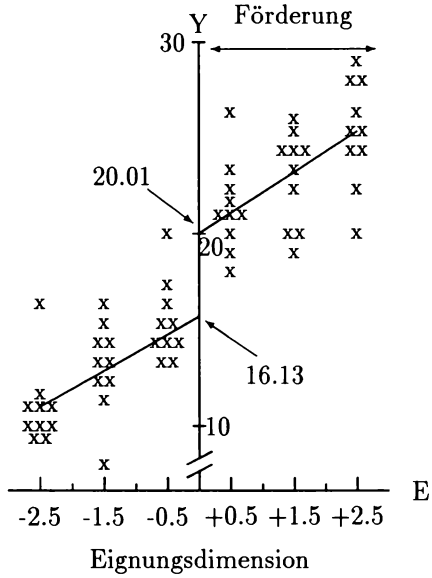

ABBILDUNG 4. **Simuliertes Ergebnismuster der Regressions-Diskontinuitäts-Analyse. Die Regressionsgleichung in der Bezugskategorie lautet** $Y_0' = b_1 + b_3E = 16.13 + 2.15E$, **in der Vergleichskategorie** $Y_1' = (b_1 + b_2) + (b_3 + b_4)E = 20.01 + 1.95E$.

Effekt. Der Koeffizient b_3 spiegelt den Zusammenhang zwischen Vor- und Nachtest wieder, der in Bezugs- und Vergleichskategorie ähnlich ausfällt.

Die regressionsanalytische Auswertung verdeutlicht die spezifische Hypothese, die in der RDA geprüft wird, nämlich daß die Beobachtungen an einer bestimmten Stelle eines Kontinuums von einem vorherigen Muster abweichen (vgl. Zeitreihen-Designs). Die RDA ermöglicht eine ähnliche Schätzung eines *Treatment-Effektes* wie in einem Experiment mit Zufallszuteilung (Rubin, 1977). Umso erstaunlicher ist es, daß so wenige Anwendungen der RDA zu finden sind.

Vergleicht man die regressionsanalytische Auswertung mit einer kovarianzanalytischen, so ist festzustellen, daß die Voraussetzung der Homogenität der Regressionsgeraden (die Parallelität der Regressionsgeraden innerhalb der Gruppen), die für die Kovarianzanalyse gemeinhin gemacht wird, für die regressionsanalytische Auswertung nicht notwendig ist. Eine andere Voraussetzung, die jedoch gleichermaßen für beide Auswertungsstrategien gültig ist, bezieht sich auf die Fehlerbehaftetheit der Vortestwerte. Die Ergebnisse der RDA sind nur schwer interpretierbar („Bedrohung" der internen Validität), wenn – etwa aufgrund eines Mildeeffekts – Personen gefördert werden, die aufgrund ihrer Vortestwerte keinen Anspruch hierauf haben. Die Trennungslinie muß bei der Zuteilung der Fördermaßnahme strikt eingehalten werden; dies ist natürlich bei geringer Reliabilität der Zuteilungsvariablen nur schwer möglich. Zur Erhöhung der Präzision der Zuteilung empfehlen sich multiple Messungen des Kriteriums. Eine weitere Voraussetzung ist, daß in den Gruppen ein hinrei-

chend großer Wertebereich in der Eignungsdimension, d.h. im Zuteilungskriterium, repräsentiert ist, damit die Regressionslinien zuverlässig schätzbar sind.

Ein besonderes Problem für die Durchführung der RDA stellt eine Nichtlinearität des Zusammenhangs zwischen Vor- und Nachtestwerten dar, wie sie in Abbildung 2C veranschaulicht wird. Ein solches Ergebnismuster kann durch Decken- und Bodeneffekte zustandekommen. Wenn bei einem nichtlinearen Zusammenhang, der auf Decken- und Bodeneffekte zurückgeht, lineare Regressionen in den beiden Gruppen berechnet werden, besteht die Gefahr, daß fälschlicherweise ein *Treatment-Effekt* erschlossen wird. Als Vorsichtsmaßnahme kann empfohlen werden, die Eignungsdimension in drei Bereiche einzuteilen. Während der untere Bereich der Kontrollgruppe und der obere Bereich der Versuchsgruppe zugeordnet wird (oder umgekehrt), soll im Mittelbereich eine Zufallszuteilung auf die Versuchs- und die Kontrollgruppe vorgenommen werden. Hierdurch ergibt sich eine Überlappung der Vortestwerte in beiden Gruppen, was eine Prüfung der Nichtlinearität des Zusammenhangs ermöglicht.

4 Validität quasi-experimenteller Versuchspläne

Abschließend sollen vier Bereiche der Validität von empirischen Untersuchungen, die von Cook und Campbell (1979) angesprochen werden, gegenübergestellt und im Hinblick auf nichtäquivalente Kontrollgruppenpläne bzw. Regressions-Diskontinuitäts-Analysen diskutiert werden: die interne Validität, die externe Validität, die Konstruktvalidität und die Validität der statistischen Schlußfolgerungen.

Interne Validität bezieht sich auf das Problem, ob eine verläßliche Beziehung zwischen UVn und AVn erschlossen werden kann. An dieser Stelle soll nicht die wissenschaftslogische Diskussion über die Beziehung zwischen interner Validität und kausalen Schlüssen aufgegriffen werden. Ein Beispiel für eine Bedrohung der internen Validität ist durch zwischenzeitliches Geschehen gegeben, das sich zwischen den Zeitpunkten Intervention und Messung der AV auswirkt und zu einer plausiblen Alternativerklärung der Ergebnisse beitragen kann. Eine solche Bedrohung ist aber bei Versuchsplänen mit Kontrollgruppen relativ unwahrscheinlich, wenn auch in nichtexperimentellen Kontexten nicht gänzlich auszuschließen.

Mit einer Bedrohung durch Regressionseffekte ist ernsthaft zu rechnen, wenn die Untersuchungsgruppen sich in ihren Vortestwerten unterscheiden und/oder aufgrund von fehlerbehafteten Vortests im Nachtest eine statistische Regression zur Mitte wirksam wird. Solche Regressionseffekte können *Treatment-Effekte* vortäuschen, und zwar auch dann, wenn das *treatment* keinen Einfluß ausübt. Wenn etwa ein Förderprogramm für Personen mit schwacher Leseleistung durchgeführt wird und die Leseleistung fehlerhaft gemessen wurde, dann ist zu erwarten, daß die „schwache" Gruppe im Nachtest allein aufgrund der Regression zur Mitte verbesserte Werte zeigt, während die „gute" Gruppe aufgrund desselben Phänomens verschlechterte Werte erreicht. Dieses Problem kann offensichtlich die Validität der RDA belasten.

Durch Parallelisierung kann ebenfalls ein Regressionseffekt zur Mitte ausgelöst werden, der einen *Treatment-Effekt* vortäuscht. Wenn die parallelisierten Gruppen aus Populationen mit unterschiedlichen Mittelwerten stammen, unterscheiden sich die beobachteten Mittelwerte im Vortest aufgrund des Parallelisierens zwar nicht, aber im Nachtest tendieren die beiden Stichproben jeweils in Richtung auf ihren Po-

pulationsmittelwert, so daß dieser Regressionseffekt Unterschiede im Nachtest hervorrufen kann. Dieses Problem ist insbesondere bei nichtäquivalenten Kontrollgruppenplänen zu beachten.

Die externe Validität bezieht sich zunächst auf die Generalisierbarkeit der Ergebnisse auf andere Personen als die, die in der untersuchten Stichprobe enthalten waren. Sie bezieht sich aber auch auf die Frage, für welche *settings* die gefundenen Ergebnisse gültig und ob sie zeitinvariant sind. Man kann annehmen, daß die RDA im Hinblick auf einige Facetten der externen Validität gut abschneidet, weil ihr Daten zugrundeliegen, die in natürlichen Umgebungen erhoben werden können.

Die Konstruktvalidität bezieht sich auf die Frage, inwieweit wesentliche Aspekte des interessierenden Merkmals in der Operationalisierung nicht berücksichtigt wurden und inwieweit Aspekte ins Spiel kommen, die für diese Merkmale irrelevant sind. Diese Probleme können sowohl bei der Operationalisierung der UV als auch bei der Erfassung der AV auftreten. Das Konzept des multiplen Operationalismus zielt darauf ab, Probleme bei der Konstruktvalidierung durch konzeptuell unterschiedliche Operationalisierungen der AV und der UV zu vermeiden. Die Idee eines multiplen Operationalismus, der anstelle der exakten Replikation den Gedanken der konzeptuellen Replikation in den Vordergrund stellt (s. Hendrick & Jones, 1972), gewinnt ihre Attraktivität nicht zuletzt aus der Überlegung, daß sich die verfahrenstypischen Fehler nicht oder nur geringfügig überschneiden, wenn unterschiedliche Techniken zur Herstellung der UV als auch zur Messung der abhängigen Merkmale verwendet werden. Ein Vorteil von Quasi-Experimenten, der insbesondere im Kontext des multiplen Operationalismus sichtbar wird, besteht darin, daß eine Vielzahl von AVn herangezogen werden kann, um eine spezifische Hypothese zu prüfen.

Was die Validität der statistischen Schlußfolgerungen angeht, so liegt eine Bedrohung dieser Form der Validität dann vor, wenn sich falsche Schlüsse über die Kovariation zwischen UVn und AVn aufgrund der statistischen Auswertung ergeben. Diese Bedrohungen können bei nichtäquivalenten Kontrollgruppenplänen beispielsweise dann auftreten, wenn ein statistisches Verfahren gewählt wird, welches die eventuellen Vortestdifferenzen nicht hinreichend berücksichtigt. Das Problem der Entscheidung für das jeweils geeignete Verfahren, Kovarianzanalyse mit oder ohne Reliabilitätskorrektur, mehrfaktorielle Varianzanalyse mit wiederholten Messungen, Strukturgleichungsmodelle oder schlichte Differenzwerte, muß in diesem Kontext gesehen werden. Quasi-experimentelle Pläne beinhalten Auswertungs- und Interpretationsprobleme, die sie auf den ersten Blick als unterlegen gegenüber experimentellen Plänen erscheinen lassen. Systematische Vergleiche analoger experimenteller und quasi-experimenteller Pläne (Rubin, 1974, 1977) zeigen jedoch, daß von einer generellen Unterlegenheit nicht grundsätzlich gesprochen werden kann.

Diese Betrachtung der Validitätsprobleme war natürlich selektiv und umfaßt nicht alle Bedrohungen der Validität, die unter den vier Rubriken abzuhandeln sind. Sie soll den Leser aber für die Frage sensitivieren, inwieweit die Interpretation der Ergebnisse im Einzelfall zu weit geht – sei es im Hinblick auf kausale Schlüsse oder im Hinblick auf die Generalisierbarkeit der Ergebnisse. Quasi-experimentelle Versuchspläne und deren Auswertungsstrategien folgen im Grunde einer anderen Theorie der Kausalität als die Experimente (vgl. Mackie, 1974). Mit Quasi-Experimenten steht eine Methodologie zur Verfügung, die berücksichtigt, daß Kausalbeziehungen

in einen je spezifischen und komplexen Kontext eingebettet sind. „Kausale Generalisierung" kann durch konzeptuelle Replikation und nachfolgende Meta-Analysen erreicht werden (Cook, 1991).

5 Weiterführende Literatur

Eine umfassende Darstellung quasi-experimenteller Ansätze findet sich bei Cook und Campbell (1979), die auch ausführlich auf die verschiedenen Aspekte der Validität eingehen. Einen kürzeren Überblick liefern schon Campbell und Stanley (1966). Ebenso sei auf Bierhoff und Rudinger (1980) sowie Rudinger und Bierhoff (1980) verwiesen. Neuere Darstellungen finden sich bei Cook und Shadish (1994), Cook, Campbell und Perrachio (1990), Sechrest, Perrin und Bunker (1990) und Trochim (1986).

Literaturverzeichnis

Abadzi, H. (1984). Ability grouping effects on academic achievement and self-esteem in a southwestern school district. *Journal of Educational Research, 77*, 287–292.

Bierhoff, H. W. & Rudinger, G. (1980). Probleme der Versuchsplanung im quasi-experimentellen Bereich. In K. D. Hartmann & K. F. Köppler (Hrsg.), *Fortschritte der Marktpsychologie*, Band 2 (S. 115–134). Frankfurt: Fachbuchhandlung für Psychologie.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park: Sage.

Campbell, D. T. & Boruch, R. F. (1975). Making the case for randomized assignment to treatment by considering the alternatives: Six ways in which quasi-experimental evaluations tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experience: Some critical issues in assessing social programs* (pp. 195–296). New York: Academic Press.

Campbell, D. T. & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cochran, W. G. & Cox, G. M. (1957). *Experimental design.* New York: Wiley.

Cook, T. D. (1991). Meta-analysis: Its potential for causal description and causal explanation within program evaluation. In G. Albrecht, H.-U. Otto, S. Karstedt-Henke, & K. Bollert (Eds.), *Social prevention and the social sciences: Theoretical controversies, research problems and evaluation strategies* (pp. 245–285). Berlin: de Gruyter.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation.* Boston: Houghton Mifflin.

Cook, T. D., Campbell, D. T. & Perrachio, L. (1990). Quasiexperimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*, 2nd ed. (pp. 491–576). Palo Alto: Consulting Psychology Press.

Cook, T. D. & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*, 545–580.

Elashoff, J. D. (1969). Analysis of covariance. A delicate instrument. *American Educational Research Journal, 6*, 383–401.

Fife-Schaw, C. (1995). Quasi-experimental designs. In G. M. Breakwell, S. Hammond & C. Fife-Schaw (Eds.), *Research methods in psychology* (pp. 85–98). London: Sage.

Gottman, J. M. & Rushe, R. H. (1993). The analysis of change: Issues, fallacies, and new ideas. *Journal of Consulting and Clinical Psychology, 61*, 907–910.

Harrop, J. W. & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time series. *Multivariate Behavioral Research, 20*, 27–44.

Harrop, J. W. & Velicer, W. F. (1990). Computer programs for interrupted time series analysis: A qualitative evaluation. *Multivariate Behavioral Research, 25*, 219–231.

Hendrick, C. & Jones, R. A. (1972). *The nature of theory and research in social psychology.* New York: Academic Press.

Linn, R. L. & Werts, C. E. (1977). Analysis implications of the choice of a structural model in the nonequivalent control group design. *Psychological Bulletin, 84*, 229–234.

Lord, F. M. (1960). Large-scale covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 307–321.

Mackie, J. L. (1974). *The cement of the universe.* Oxford: Oxford University Press.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–210). New York: Springer.

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics, 2*, 1–26.

Rudinger, G. & Bierhoff, H. W. (1980). Quasi-experimentelle Versuchspläne für die Markt- und Kommunikationspsychologie. In K. D. Hartmann & K. F. Köppler (Hrsg.), *Fortschritte der Marktpsychologie*, Band 2 (S. 135–163). Frankfurt: Fachbuchhandlung für Psychologie.

Sechrest, L., Perrin, E. & Bunker, J. (Eds.) (1990). *Research methodology: Strengthening causal interpretations of nonexperimental data.* Rockville: AHCPR, PHS.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analysis. *Psychological Bulletin, 84*, 489–502.

Stanley, T. D. (1991). „Regression-discontinuity design" by any other name might be less problematic. *Evaluation Review, 15*, 605–624.

Swaminathan, H. & Algina, J. (1977). Analysis of quasi-experimental time-series designs. *Multivariate Behavioral Research, 12*, 111–131.

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*, 309–317.

Trochim, W. M. K. (Ed.) (1986). *Advances in quasi-experimental design analysis: New directions for program evaluation* (Vol. 31). San Francisco: Jossey-Bass.

Trochim, W. M. K., Cappelleri, J. C. & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review, 15*, 571–604.

Visser, R. A. & de Leeuw, J. (1984). Maximum likelihood analysis for a generalized regression-discontinuity design. *Journal of Educational Statistics, 9*, 45–60.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Autorenhinweis

Die Autoren danken Thomas Richter und Monika von Wachter für ihre konstruktive Mitarbeit.