4.1 Planung und Durchführung organisationspsychologischer Untersuchungen

Klaus Moser

1. Einleitung

Gegenstand der Organisationspsychologie sind Beschreibung, Erklärung, Prognose und Veränderung des Erlebens und Verhaltens von Menschen in Organisationen. Analyseebenen sind dabei Tätigkeiten, Individuen, Gruppen oder Organisationen. Die vielfachen Ziele der Organisationspsychologie werden mit Hilfe von Methoden zu erreichen versucht. Methoden können zur Erreichung von fünf Teilzielen beitragen. Zunächst beruhen Beschreibung, Erklärung, Prognose und Veränderung auf der Diagnose von Sachverhalten. Für solche Diagnosen können *Methoden zur Datengewinnung* eingesetzt werden. Insbesondere zur Überprüfung von Erklärungen werden oft eigenständige Untersuchungen vorgenommen, die durch *Methoden der Untersuchungsplanung* gesteuert werden. In vielen Fällen mag man sich aber noch gar nicht über die genaue Fragestellung im klaren sein; hier können *Methoden zur Untersuchungsvorbereitung* helfen. Kommt man schließlich zum Ergebnis, daß eine Untersuchung stattfinden soll, dann werden meistens nicht nur diagnostische, sondern auch *Interventionsmethoden*, also Methoden zur Gestaltung von Sachverhalten, eingesetzt. Schließlich stellt sich in verschiedenen Phasen einer Untersuchung die Frage, wie die Methoden bzw. die mit ihnen gewonnenen Ergebnisse zu bewerten sind, *wozu Methoden zur Bewertung von Untersuchungen, vor allem zur statistischen Analyse und Evaluation* verwendet werden können. In diesem Kapitel wird es vor allem um die ersten drei Methodenvarianten gehen. Die Interventionsmethoden werden in weiteren Lehrbuchkapiteln beschrieben, während auf die Darstellung von statistischen Methoden aus Platzgründen verzichtet wird (vgl. als anwendungsorientierte Einführungen Bortz, 1989, sowie Backhaus, Erichson, Plinke, Schuchard-Ficher & Weiber, 1987). Im folgenden wird nun ganz allgemein von einer *Untersuchung* gesprochen, gleich ob es um Beschreibung, Erklärung, Prognose oder Intervention als Ziel organisationspsychologischen Handelns geht.

2. Untersuchungsvoraussetzungen

Bevor eine Untersuchung konkret geplant werden kann, ist zunächst das Umfeld, in dem die Untersuchung stattfinden soll, zu analysieren. Organisationspsychologen werden, wenn sie Untersuchungen in Organisationen vornehmen, Teil des Handlungssystems «Organisation» und müssen sich damit mit Durchsetzungsstrategien, Koalitionsbildungen usw. wohl oder übel befassen, wobei diese allerdings nicht nur Hindernis sein müssen, sondern auch zur Durchsetzung der eigenen Ziele ausgenützt werden können.

Probleme des *Projektmanagements* ergeben sich entweder zu Beginn oder sogar schon in der Vorbereitung einer Untersuchung. So stellt sich gleich zu Beginn eines Projekts bzw. einer Untersuchung die Frage, *wie* konkret die Ausarbeitungen aus der Phase vor dem «offiziellen Projektbeginn» bereits sind. Da eine detaillierte Planung oft schon das halbe Projekt ist, gilt es hier schwierige und riskante Kosten-/Nutzenabwägungen zu treffen, bevor z. B. Projektanträge gestellt werden (Wottawa & Thierau, 1990). Vor allem in der betriebswirtschaftlichen Literatur werden Methoden zur Zeit- und Kostenabschätzung dargestellt. Von den Methoden

der Zeitabschätzung seien die Balkenplan- und die Netzplantechnik erwähnt (Darstellung in Wottawa & Thierau, 1990).
Am Beginn einer Untersuchung steht zudem die Ist-Analyse oder Bestandsaufnahme, also die Erkundung der Ausgangssituation und der Umstände, unter denen Untersuchungen stattfinden. Beispiele sind die Ermittlung eventuell zugänglicher Stichprobengrößen, das Vorliegen anderer (möglicherweise ähnlicher und deshalb Reaktivität erzeugender) Untersuchungen, eine eventuelle Sekundäranalyse bereits vorliegender Daten oder die Nutzbarmachung von existierenden Methoden oder know how. Zur Ist-Analyse sollen hier nicht Varianten von Diagnosen oder eventuelle Vortests zählen; sie werden in Abschnitt 4 behandelt. In den folgenden Teilabschnitten wird nun auf zwei weitere Untersuchungsvoraussetzungen näher eingegangen, die *Klärung von Bedürfnissen und Interessen* und die *Präzisierung von Untersuchungszielen und Kriterien.*

2.1 Die Klärung von Bedürfnissen und Interessen

Jeglichem organisationspsychologischen Handeln kann (auch) ein Bedürfnis oder ein Interesse unterstellt werden. Vor und während Untersuchungen sind das Interesse eines eventuellen Auftraggebers, das eigene Erkenntnisinteresse und die möglichen Interessenkonflikte zu klären. Die Explizierung der tatsächlichen und wahrgenommenen humanitären, wirtschaftlichen und wissenschaftlichen Ziele einer Untersuchung kann Konflikte verhindern und die Entstehung von Gerüchten vermeiden helfen. Hingegen mag die Ignorierung der Interessen der Auftraggeber (oder gar eventueller Koalitionen) vor allem im Falle eines eher «wissenschaftlichen» Interesses des Untersuchungsleiters zwar naheliegen, kann aber zur Konterkarierung des eigentlich geplanten Vorgehens führen. Denn Untersuchungen haben aus Organisationssicht nicht selten mikropolitische Funktionen wie z. B. Verantwortungsdelegation, Verantwortungsdiffusion, Durchsetzungshilfe oder Rechtfertigung für ohnehin geplante Veränderungen. Daher empfiehlt es sich, das organisationale Umfeld genauer zu studieren (vgl. weiterführend Kapitel 4.2). Ein mögliches Vorgehen lautet, Entscheidungsträgern die *möglichen* Konsequenzen einer Untersuchung zu beschreiben und die Erwünschtheit bzw. Akzeptabilität von Konzequenzen zu erfragen. Warnsignale eines zu erwartenden mikropolitischen Mißbrauchs sind unter anderem,

- daß nur *ein* Ergebnis akzeptiert werden wird;
- daß selbst widersprüchliche Ergebnisse identische (eben ohnehin geplante oder erwünschte) Maßnahmen zur Konsequenz haben oder
- daß *keines* der möglichen Ergebnisse irgendwelche Konsequenzen hat (was den Gedanken nahelegt, daß ein innerorganisationales Problem in einer «wissenschaftlichen Untersuchung» vergraben werden soll).

Fehlende Klärungen mikropolitischer Interessen oder Randbedingungen können organisationspsychologische Untersuchungen nahezu sinnlos machen. Auf Untersuchungen, die aufgrund der Randbedingungen nicht solide durchführbar sind oder deren Ergebnisse mit hoher Wahrscheinlichkeit verzerrt kommuniziert oder mißbraucht werden, sollte am besten ganz verzichtet werden. Positiv formuliert sind als Randbedingungen der Durchführbarkeit einer Untersuchung folgende Voraussetzungen bei den Beteiligten zu nennen (vgl. Wottawa & Thierau, 1990):

- Bereitschaft, etwas zu verändern
- Erwartung eines Erfolgs der Untersuchung
- ersichtlicher Nutzen der Untersuchung für die Beteiligten
- Akzeptanz der Fakten/Ergebnisse aus der Untersuchung.

Diese vier Bedingungen sind nur in Ausnahmefällen gänzlich ohne Probleme herstellbar. Zudem haben die Beteiligten öfter schwer in Übereinstimmung bringbare Interessen (vgl. Kapitel 4.2). Information der

Untersuchungspersonen über Ziel und Vorgehen, Transparenz von Methoden oder Partizipation an einzelnen Entwicklungsschritten sind nur einige der relevanten Faktoren, deren Erhöhung die Akzeptanz einer Untersuchung aber im allgemeinen verbessern wird.
Die Klärung von Bedürfnissen und Interessen geht fließend in die nächste Phase der Untersuchungsvoraussetzungen, der Zieldefinition bzw. -konkretisierung, über. Gerade Bedürfnisse, Interessen und abzusehende Widerstände haben nicht selten schon zur Neuformulierung oder gar Neuorientierung von Untersuchungszielen geführt.

2.2 Ziele und Kriterien

Wenn Untersuchungsideen bereits allgemein festgelegt worden sind, so können die Ziele doch zunächst vage sein. Nehmen wir als Beispiel das Thema «Analyse der Arbeitszufriedenheit in einer Organisation», dann ist damit das Ziel nur ungefähr bekannt: Besteht das Problem darin, daß man zu wenig über das Phänomen weiß, daß nach einem Ansatzpunkt zur Veränderung gesucht oder eine Bestandsaufnahme geplant wird? Eine Problemlösung setzt demnach die Konkretisierung des Ziels voraus. Entsprechend sind dann auch die Kriterien zu konkretisieren, an denen die Zielerreichung bzw. Problemlösung erkennbar ist.
Konkretisierung des Ziels: Ziele oder Kriterien werden oftmals zunächst nur sehr allgemein und vage formuliert. Hier bestehen die Aufgaben darin, das Ziel konkreter zu formulieren, mögliche verschiedene Kriterien zu explizieren und eventuell auch bereits eine Gewichtung der verschiedenen (Teil-) Ziele vorzunehmen. *Quellen* für diese Konkretisierung sind insbesondere Experteninterviews, Diskussionen mit den beteiligten Gruppen und die Analyse bereits existierender Untersuchungsansätze (z.B. Literaturstudium). *Hilfsmittel bzw. Methoden* zur Konkretisierung sind z.B. Interviewleitfäden, Moderationstechniken (z.B. Schnelle, 1982) und Kreativitätstechniken (z.B. Hoffmann, 1987).

Zur Konkretisierung des Ziels, aus der sich zumindest teilweise auch der später zu verwendende Untersuchungsplan ergibt, soll hier auch die *Präzisierung von Hypothesen* bzw. Annahmen über den Untersuchungssachverhalt gezählt werden. So können einer Untersuchung zur Arbeitszufriedenheit folgende Arten von Fragen zugrundeliegen (vgl. auch Bortz, 1984):

- *beschreibend und hypothesenerkundend,* z.B.: Welche Varianten der Arbeitszufriedenheit existieren? (wobei die Betonung auf Detail- und Einzelbeschreibungen liegt)
- *populationsbeschreibend,* z. B.: Wieviel Prozent der Belegschaft sind gänzlich arbeitszufrieden? (Betonung auf Repräsentativität)
- *unspezifische Hypothesen überprüfend,* z. B.: Hängen unterschiedliche Formen der Arbeitszufriedenheit mit der schulischen Ausbildung zusammen? (Schwerpunkt auf Korrelationsanalysen)
- *spezifische Hypothesen überprüfend,* z. B.: Verursacht geringe Arbeitszufriedenheit die Erhöhung der Fluktuation? (Betonung von Kausalzusammenhängen)

Vorteile einer möglichst konkreten Formulierung von Hypothesen wurden bereits in Abschnitt 2.1 angesprochen. Diese Ausführungen sollen nun ergänzt werden. Zum einen wird durch eine frühzeitige Formulierung von Hypothesen vermieden, daß zu untersuchende Sachverhalte nur unvollständig ermittelt werden (z.B.: man «vergißt» die Erhebung erst nachträglich als interessant erkannter Variablen). Zudem steuern Hypothesen die Datenanalyse; wird stattdessen einfach «herumprobiert», dann drohen zufällige Effekte für ernst genommen zu werden. Diese Gefahr ist deshalb in der (Organisations-) Psychologie besonders groß, weil nahezu beliebige Zusammenhänge im Nachhinein doch recht plausibel erscheinen.
Die Konkretisierung des Ziels und des Hypothesentyps hat selbstverständlich für die nachfolgenden Untersuchungsphasen entscheidende Bedeutung. Einige der im folgenden zu beschreibenden methodischen Fragen ergeben sich auch deshalb, weil aus

den verschiedensten Erwägungen heraus das Ziel nur unvollständig geklärt werden kann. Hinzu kommt, daß Ziele in Kriterien umgesetzt und handhabbar, sprich: operationalisiert werden müssen.

Kriterien und deren Operationalisierung: In vielen Fällen werden Organisationspsychologen als Experten zu Problemfällen herangezogen werden, wobei die Problemerkennung nicht zu den Aufgaben zählt. Dies kann bedeuten, daß nicht ein Kriterium, sondern die Untersuchung an sich im Vordergrund steht. So werden beispielsweise Maßnahmen der Organisationsentwicklung oft ohne spezifisches Kriterium unternommen, von Fall zu Fall bestehen Interventionen auch darin, die Kriterien erst von den Untersuchungspersonen erarbeiten zu lassen (vgl. auch Kapitel 4.2). In der überwiegenden Mehrzahl von Untersuchungen dürften aber Ziele auch in zu operationalisierende *Kriterien* umzusetzen sein. Solide Operationalisierungen setzen entweder die Verwendung von Standardverfahren voraus oder sollten zumindest von einer sorgfältigen definitorischen Phase begleitet sein (siehe auch Abschnitt 4.2). Kriterien für Kriterien sind unter anderem:

- Sparsamkeit (z.B.: Beurteilungen sollten mit möglichst wenigen globalen Skalen möglich sein).
- Umfang (Beurteilungsskalen sollten eine Beschreibung vieler Personen erlauben).
- Differenzierungsgrad (Beurteilungsskalen sollten individuellen Besonderheiten im Verhalten gerecht werden; so bestehen Arbeitsergebnisse von Informatikern aus Programmen, von Konstrukteuren aus Zeichnungen und von Vertriebsingenieuren aus Umsatzzahlen).

Die drei Kriterien Sparsamkeit, Umfang und Differenzierungsgrad stehen in einem Spannungsverhältnis zueinander. Dies wird sich auch bei den Methoden der Datengewinnung zeigen. So stellen qualitative Methoden die *Differenzierung* in den Mittelpunkt, quantitative Methoden die *Sparsamkeit* und den *Umfang.*

Die Entscheidung für bestimmte Kriterien von Kriterien muß aber nicht willkürlich sein. Zur Gewinnung eines ersten Überblicks, aber auch zur verständlichen Kommunikation, ist es oft zweckmäßig, die ersten beiden Kriterien zu betonen. Geht es hingegen darum, konkrete Maßnahmen zu entwickeln und einzuleiten, sind differenzierte Analysen kaum Verzichtbar.

Fragt man nach den Kriterien, mit denen die Zielerreichung überprüft werden kann, dann werden sie, wie bereits gesagt, zu Beginn einer Untersuchung nur vage benannt. So kann es Aufgabe eines Organisationspsychologen sein, die Personalarbeit eines Unternehmens zu verbessern. Was aber konkret «verbessern» heißt, ist von ihm selbst zu erarbeiten. Als Kriterien einer entsprechenden Untersuchung kommen prinzipiell in Frage: mehr Bewerber, qualifiziertere Bewerber, weniger Absagen auf Stellenangebote, leistungsfähigere Mitarbeiter, geringere Fluktuation während des ersten Jahres, Reduzierung des Absentismus usw.. Die naheliegende Konsequenz lautet nun, das am einfachsten erhebbare Kriterium heranzuziehen (z.B. Zahl der Bewerber). Auf dieses Kriterium hin werden dann Maßnahmen entwickelt (z.B. ansprechendere Stellenanzeigen). Fraglich ist, ob es sich hierbei um eine *Verbesserung* handelt. Möglicherweise ist es so, daß Bewerber erst dann für ein Unternehmen akzeptabel sind, wenn sie über ein bestimmtes Ausmaß an Qualifikation verfügen. Durch die hier genannte Intervention kann aber die Zahl *qualifizierter* Bewerber nicht beeinflußt werden, und die Konsequenz wäre, daß die Organisation lediglich mehr Absagen an Bewerber formulieren muß, die sich ansonsten zuvor gar nicht erst beworben hätten. Konsequenz aus diesen Überlegungen ist, daß in solchen weniger klaren Fällen möglichst *mehrere* Kriterien erfaßt werden.

Ein weiterer Grund für die Registrierung mehrerer Kriterien sind mögliche (in der Regel unerwünschte) Nebenfolgen von Untersuchungen. Ein Beispiel ist die Einführung einer realistischen Tätigkeitsinformation (vgl. Kapitel 9), die zwar zu einer erwünschten Verringerung der Fluktuation, aber auch zu einer weniger erwünschten Erhöhung der Ablehnung von Stellenangebo-

ten führt. In solchen Fällen kann erst dann eine Ermittlung der Nettowirkung einer Maßnahme erfolgen, wenn die Teilwirkungen bekannt sind.

Die meisten Kriterien stellen eine nur unvollkommene Annäherung an das eigentliche Ziel dar. Ein letztendliches oder «ultimatives» Kriterium ist schwierig anzugeben und wird oft auch nur allgemein als «Erfolg» oder «Überleben» der Organisation formuliert. Zudem sind ultimative Kriterien oft nicht zugänglich oder nur sehr selten beobachtbar (z. B. der Erfolg von Rettungspersonal bei einer Katastrophe). Die behelfsmäßige Verwendung spezifischer Kriterien ist ebenfalls fraglich, da sie in sich heterogen sind (vgl. Kapitel 9). In vielen Fällen empfiehlt sich die Kombination bzw. Aggregation von Kriterien. Häufig gewählte Vorgehensweisen zur Kombination von Kriterien sind

- Gleich-Gewichtung der Einzelkriterien und lineare Kombination;
- Transformation der erkennbaren Wirkungen in eine gemeinsame Maßeinheit (z.B. Geld; Verhaltensvarianz);
- Schätzung der Kriteriengewichte und anschließende Addition über Expertenurteile oder Regressionsanalysen.

Die Aggregation von Kriterien kann schließlich auch intuitiv bzw. global erfolgen, nämlich durch Vornahme einer Globalskalierung. Beispiele sind die Kuninskala für die Zufriedenheit oder die Prozentrangskala (Schuler, 1972) für die Leistung auf individueller Ebene.

In diesem Teilabschnitt wurde für den reflektierten Einsatz von Kriterien plädiert. Dies steht im Gegensatz zu einem oft pragmatischen Vorgehen gerade in der Organisationspsychologie. Das hieraus sich ergebende Problem zeigt sich dann, wenn nach den *Ursachen* von Kriterienausprägungen gefragt wird. Das allzu oberflächliche Akzeptieren von vorgegebenen oder einfach ermittelbaren Kriterien kann extreme Konsequenzen für Organisation und Individuum haben. Dies zeigen exemplarisch Argumente von Staw (1984) zur Abwesenheit vom Arbeitsplatz in der Organisation als relevantes Maß zur Bewertung der Erreichung organisationaler Ziele (vgl. Informationsbox 1).

Informationsbox 1

Zur Relativität der Bedeutung von Kriterien

Abwesenheit vom Arbeitsplatz in der Organisation (Absentismus) gilt in der Regel als für eine Organisation ungünstiges oder dysfunktionales Verhalten. Beispielsweise gilt sie als Beleg für eine mangelnde Identifikation mit der Organisation. Dem stehen jedoch folgende Überlegungen entgegen (Staw, 1984):

- Absentismus kann der Erhaltung der Arbeitskraft dienen (vgl. die Begrenzung der Arbeitszeit für Kraftfahrer).
- Absentismus kann der Erholung von ungünstigen Arbeitsbedingungen dienen.
- Absentismus kann der Tätigkeit direkt dienlich sein (Klientenbesuche, Weiterbildung).
- Absentismus wird als Kriterium zunehmend durch (auch zeitlich) flexibel erreichbare Arbeitsziele ersetzt.
- Absentismus wird bei einer Zunahme von Heimarbeit am Computer obsolet werden.

3. Untersuchungsplanung

In der Phase der konkreten Untersuchungsplanung wird eine Fixierung von Zielen und Mitteln vorgenommen. Die Phase der Untersuchungsplanung soll im folgenden all das umfassen, was in der Methodenliteratur auch als Frage der «Versuchsplanung» verhandelt wird. Die Untersuchungsplanung betrifft die Aspekte, die bei der beabsichtigten Aussage über das Ergebnis einer Untersuchung zu berücksichtigen sind. Methoden der Untersuchungsplanung dienen dazu, diese Aussagen auf eine möglichst solide Basis zu stellen. Grundannahme aller Untersuchungsplanungs- oder Designfragen ist, daß eine Untersuchung prinzipiell kritisierbar ist und eventuell auch tatsächlich kritisiert wird. Solche kritischen Fragen lauten etwa: Mißt dieser Test tatsächlich Intelligenz, ist dieses Training tatsächlich wirksam, belegt diese Untersuchung tatsächlich den Zusammenhang zwischen Zielsetzung und hoher

Leistung? Untersuchungsplanung soll entweder möglicher Kritik zuvorkommen oder eine Untersuchung der Kritik aussetzbar und daraufhin prüfbar machen. Im folgenden wird zunächst erläutert, warum diese Anforderung am besten im *Experiment* erfüllt wird, und dann werden Probleme von und Alternativen zu Experimenten behandelt.

3.1 Experimente in der Organisationspsychologie

Als Königsweg unter den Untersuchungsmethoden in den Wissenschaften allgemein gilt das Experiment. Zunächst soll verdeutlicht werden, warum Experimente in den Wissenschaften so bedeutsam sind. Dann wird auf die Rolle und die Kritik an Experimenten in der Organisationspsychologie eingegangen. Bereits an dieser Stelle sei darauf hingewiesen, daß Experimente in der Organisationspsychologie mittlerweile eher die Ausnahme sind (vgl. Kapitel 4.2). Daß im folgenden zunächst dennoch von Experimenten als (vermeintlichem) «Ideal» ausgegangen wird, hat primär didaktische Zwecke.

Wissenschaft hat die Aufgabe, Sachverhalte, Ereignisse, Beziehungen *usw. zu erklären.* Hierzu werden Theorien bzw. einzelne Hypothesen aufgestellt und auf ihre Korrektheit überprüft. Mit Erklärung ist hierbei eine Kausalerklärung gemeint: es wird auf eine Ursache verwiesen. Wenn nun eine solche Kausalerklärung gegeben wird, dann begegnet man zwei Problemen: (1) Die Erklärung kann richtig sein, aber eine vorgenommene Überprüfung besagt, daß die Verursachung nicht beobachtet werden konnte. (2) Die Erklärung kann falsch sein, aber die Überprüfung tendiert (dennoch) zu einer Bestätigung. Wie kann nun eine fälschliche Bestätigung einer Erklärung vermieden werden? Die Antwort lautet: Durch ein Experiment, in dem alle Bedingungen konstant gehalten werden. Dann wird die (vermutete) Ursache hergestellt und der (vermutete) Effekt beobachtet.

Das *Ideal eines Experiments* geht davon aus, daß *genau eine* Variable manipuliert wird (die sogenannte «unabhängige Variable» (UV)) und ihre Wirkung auf eine zweite Variable (die sogenannte «abhängige Variable» (AV)) gemessen wird. Wesentliche Merkmale eines Experimentes sind demnach, daß

- alle Randbedingungen kontrolliert sind,
- eine unabhängige Variable gezielt manipuliert wird, und
- Wirkungen bei einer abhängigen Variablen erfaßt werden.

Zusammenfassend kann so argumentiert werden: Wenn nur die unabhängige Variable variiert wurde und diese sich (daher) zeitlich vor der abhängigen Variable verändert, dann kann die Beziehung zwischen UV und AV so interpretiert werden, daß die Veränderung der UV die Änderung der AV bewirkt.

Weiter oben wurde darauf hingewiesen, daß ***Kau***salhypothesen nicht die einzige Art von ***Hypothe***sen sind, die organisationspsychologischen Untersuchungen zugrundeliegen; man denke insbesondere an Zusammenhangshypothesen oder Fragen zur ***Intensität*** der Wirkung einer Maßnahme. Nicht zuletzt deshalb, weil auch im Falle korrelativer Analysen und technologischer Interventionen von unabhängigen und abhängigen Variablen gesprochen wird, existieren bis heute Auseinandersetzungen um die Frage, mit welchen Untersuchungsmethoden Hypothesen zu evaluieren sind.

Bei der Planung und Durchführung von Experimenten stößt man nun auf zwei Probleme: UV und AV sind nahezu nie ohne Ungenauigkeit herzustellen bzw. zu registrieren, und die Wirksamkeit zusätzlich bzw. alternativ einflußnehmender Variablen (sogenannte «Drittvariablen») ist nie auszuschließen. Zudem fehlt es oft an Möglichkeiten, klassische experimentelle Kontrolltechniken einzusetzen. Im folgenden sollen die methodischen Möglichkeiten beschrieben werden, wie diese Probleme angegangen werden können. (Weitere Probleme, speziell im organisationalen Kontext, sowie alternative Konsequenzen werden im nächsten Kapitel genannt.) Diese Probleme sind vom Grundsatz her nicht sehr typisch für *organisations*psychologische Experimente, wohl aber in ihren konkreten Konsequenzen: Ex-

Informationsbox 2

Mangelnde Generalisierbarkeit von Laborexperimenten?

Untersuchungen werden oft an Universitäten mit Studierenden im Anfangssemester durchgeführt. Heißt dies nun, daß die Ergebnisse auch für andere Personen in anderen Situationen zu anderen Zeitpunkten gelten? Eine klare Antwort wäre: Da der untersuchte Zusammenhang als allgemeingültig angenommen wurde und bisher (noch?) kein Beleg dagegen vorliegt, bleibt sie allgemeingültig. Eine andere klare Antwort wäre: Echte Organisationsmitglieder reagieren bestimmt anders; solche Untersuchungen sind wertlos. Beide Antworten sind nicht ganz von der Hand zu weisen. Eine klare Empfehlung lautet deshalb, die Randbedingungen von Interventionen in ihren wesentlichen Aspekten immer mit zu dokumentieren. Es fragt sich dann eben nur, was diese wesentlichen Aspekte sind.
Darüber hinaus kann aber auch auf zwei bemerkenswerte Sammlungen von Untersuchungen verwiesen werden, in der die Effekte von Untersuchungen unter «natürlichen» vs. «künstlichen» Bedingungen verglichen werden. Locke (1986) kam zu folgendem Ergebnis: In den Fällen, in denen Ergebnisse in verschiedenen Laboruntersuchungen konsistent waren, waren sie es auch im Feld; waren die Ergebnisse im Labor uneinheitlich, so waren sie es ebenso außerhalb desselben. Gordon, Slade und Schmitt (1986) fanden hingegen in solchen Untersuchungen Unterschiede zwischen Labor und Feld, in denen die Ergebnisse statistisch ausgewertet wurden, nicht aber in den Untersuchungen ohne statistische Tests (die die Mehrzahl bildeten).

perimente sind die Ausnahme in der Organisationspsychologie (Fromkin & Streufert, 1976). Zunächst werden einige Einwände gegen organisationspsychologische Experimente behandelt werden. In den nachfolgenden Abschnitten werden dann alternative Methoden der Datengewinnung untersucht.

Ein Hauptargument gegen das Experiment als Methode lautet, daß die so gewonnenen Ergebnisse nicht auf die (organisationale) Realität übertragbar oder «generalisierbar» sind. Die Langlebigkeit dieses Arguments hat trotz verbreiteter Kritik am Begriff der «Generalisierbarkeit» (z.B. Gadenne, 1984) kaum an Vitalität seit den 70er Jahren eingebüßt. Zudem werden Experimente oft synonym als Laborexperimente bezeichnet. Das führt zu teilweise schwerwiegenden Mißverständnissen, weshalb es wichtig ist, darauf hinzuweisen, daß Experimentieren eine Untersuchungsstrategie ist, ein Labor aber ein Setting (eine Untersuchungsumgebung). Tatsächlich reduziert sich Generalisierbarkeit auf die Frage, ob es *relevante* Differenzen zwischen Forschungssetting und spezifischem organisationalem Setting gibt (Fromkin & Streufert, 1976; vgl. auch Informationsbox 2). Mögliche *relevante* organisationale Merkmale könnten sein (Bouchard, 1976; Fromkin & Streufert, 1976):

Intensität der UV: In Organisationen können Maßnahmen sehr viel intensiver variiert sein bzw. werden. Man denke etwa an den Streß bei bevorstehenden Kündigungen oder Auslandsaufenthalten.

Variation der Phänomene: Bestimmte Variationen von organisationalen Phänomenen scheinen nicht in einem Labor herstellbar zu sein, so z. B. Gruppengrößen (Bouchard, 1976), formale Hierarchien oder Rollendefinitionen.

Häufigkeit und Dauer: Bestimmte Maßnahmen sind erst nach ihrem gehäuften Auftreten oder nach längerer Zeit wirksam. Beispiele sind etwa familiäre Konflikte nach der Einführung von Schichtarbeit.

Zeitstrukturen: Bestimmte organisationspsychologische Phänomene haben ihren Lebenszyklus, der nicht Verkürzbar ist, z. B. der Verlauf von Forschungs- und Entwicklungsprojekten.

Natürliche Einheiten: Organisationen haben eine Tendenz zur Selbstorganisation bzw. natürlichen Entwicklung und Adaptation an die Umwelt; hier ist z. B. an die Tendenz von Organisationen, die Zusammensetzung der Persönlichkeiten ihrer Mitglieder homogener werden zu lassen, zu denken (Moser, 1991).

Effekte der Umgebung/ des «Settings»: Die bloße Definition einer Umgebung als «Labor» bzw. «echtes Unternehmen» kann schon entscheidend sein (vgl. Kapitel 4.2).

Repräsentativität der Operationalisierungen:

In einigen Fällen mögen begründete Zweifel bestehen, ob die Operationalisierung im Labor derjenigen in Organisationen entspricht. So wurden die meisten Untersuchungen zu Urteilsfehlern anhand von studentischen Beurteilungen von Professoren durchgeführt, was sicherlich andere Konsequenzen für die Beurteilten hat als innerbetriebliche Leistungsbeurteilungen.

Interdependenzen: Personen in Organisationen haben meistens mehrere Aufgaben oder Ansprechpartner, es existieren wechselseitige Abhängigkeiten und Widersprüche. In einem Laborexperiment ist in der Regel das eigene Verhalten allenfalls vor den Versuchsleitern zu rechtfertigen, nicht aber vor Kollegen, der Familie oder dem Betriebsrat.

Ein grundlegender Einwand gegen Experimente in der *Organisationspsychologie* lautet, daß wesentliche Eigenschaften von Organisationen nicht in Experimenten abgebildet werden und oft auch gar nicht abbildbar sind. Die prinzipielle Kritisierbarkeit von Experimenten (aber auch allen anderen Untersuchungsformen) geht aber auch auf den Umstand zurück, daß «Organisation» selbst ein vager Begriff ist. So werden von verschiedenen Autoren jeweils eine Vielzahl als zentral bezeichnete definitorische Merkmale angeführt, so z.B. Differenzierung von Positionen, Arbeitsteilung, Bedeutsamkeit der bearbeiteten Aufgaben u.v.m. (Weick, 1967). Wer nun der Auffassung ist, daß das eine oder andere Merkmal «zentral» oder «relevant» für eine Organisation ist, der wird dessen Fehlen bzw. andere Ausprägung in einer Untersuchung als Defizienznachweis interpretieren. Da es aber keine klaren definierenden Kriterien für «Organisationen» gibt, kann ein Experiment (aber auch fast jegliche Untersuchung) stets als irrelevant für den organisationalen Kontext bezeichnet werden. Hiergegen kann aber vorgebracht werden:

- Die isolierte Betrachtung weniger Merkmale ist ja gerade die Grundidee experimentellen Arbeitens.
- Eine 1:1-Übersetzung der Variablen, wie sie in der Realität gefunden werden, in ein Experiment wird nicht beabsichtigt und verbietet sich oft auch: Wer beispielsweise die Wirkung von Zielsetzung untersuchen will, will nicht lediglich eine Aussage über *eine* Organisation zu *einem* bestimmten Zeitpunkt, sondern vielmehr generelle Aussagen über ein (relativ abstraktes) Konzept machen. In diesem Sinne erfordern Experimente durchaus Kreativität.
- Schließlich sind auch andere Untersuchungsansätze nicht-experimenteller Art nicht weniger hiervon betroffen, solange überhaupt irgendeine Verallgemeinerung über den spezifischen Einzelfall hinaus beabsichtigt wird.

Nehmen wir nun an, daß Experimente sowohl möglich als auch sinnvoll durchführbar sind. In den nächsten Abschnitten sollen die erwähnten zwei generellen Probleme behandelt werden, der Einfluß von Meßfehlern und Drittvariablen. Die Unterscheidung zwischen diesen beiden Einflußfaktoren ist aber eher pragmatischer Natur. Drittvariablen, die (noch) nicht bekannt sind, werden den Meßfehlern zugeschlagen. Meßfehler, die (teilweise) aufgeklärt wurden, können als Drittvariablen analysiert werden.

3.2 Der Umgang mit Meßfehlern

In der Organisationspsychologie sind sowohl unabhängige als auch abhängige Variablen oft nur mit einer gewissen Toleranz bzw. einem Meßfehler zu beschreiben. Wer die Hypothese überprüfen will, daß die Bereicherung der Arbeitstätigkeit zu einer Erhöhung der Arbeitszufriedenheit führt, der wird das Problem haben, daß sowohl die Bereicherung der Arbeitstätigkeit als auch die Arbeitszufriedenheit nicht exakt gemessen werden können. Maßnahmen zum Umgang mit Meßfehlern können sowohl mit Hilfe von Methoden der Untersuchungsplanung als auch durch die Verwendung entsprechender Methoden der Datengewinnung ergriffen werden. Letztere werden in Abschnitt 4 besprochen, so z.B. die Verwendung standardisierter Meß- bzw. Beobachtungsverfahren oder das Training der Beobachter bzw. Interviewer. Von den Mög-

lichkeiten des Umgangs mit Meßfehlern ist auf der Seite der Untersuchungsplanung die *Reduzierung von Meßfehlern* die wichtigste Maßnahme. Unter der Annahme, daß es sich um unsystematische Fehler handelt, kann aber auch versucht werden, den Fehler *zu kontrollieren* bzw. zu unterdrücken. Eine besondere Rolle kommt hier der *Meßwiederholung* zu. Wiederholte Messung bzw. Beobachtung ergibt nach anschließender Bildung des arithmetischen Mittels ein verläßlicheres Beobachtungsdatum. Zudem ermöglicht die Berechnung der Streuung der beobachteten Werte eine Schätzung der Ungenauigkeit der Beobachtungen.

An dieser Stelle sei betont, daß für Experimente weder Untersuchungsgruppen noch Kontrollgruppen benötigt werden; im Idealfall genügt es, ein Experiment an genau einer Person durchzuführen. Die im folgenden zu besprechenden Probleme werden aber oft als so selbstverständlich existent akzeptiert, daß Experimente in der Organisationspsychologie tatsächlich die Ausnahme bilden.

Da die wiederholte Beobachtung des Erlebens und Verhaltens der gleichen Person zur Konsequenz haben kann, daß sie beispielsweise Lerngewinne hat oder ermüdet, werden in der Regel in der Psychologie Ver*suchsgruppen* der Manipulation der unabhängigen Variable ausgesetzt; danach werden dann die Beobachtungswerte pro Gruppe *aggregiert.* Da in der Organisationspsychologie die Untersuchung von Gruppen die Regel ist, wird dieser Zweck der Bildung von Untersuchungsgruppen oft aus den Augen verloren (vgl. Informationsbox 3). Weitere Beispiele für die Schätzung und Reduktion des Meßfehlers durch Aggregation sind die Verwendung mehrerer Fragen zum gleichen Themenkomplex, der Einsatz mehrerer Beobachter oder Interviewer und die Zusammenfassung verschiedener objektiver Daten.

Der überzeugende Nachweis der Wirkung der Variation einer unabhängigen Variable oder auch des Vorliegens eines Zusammenhangs zwischen Variablen wird durch den Meßfehler mitbestimmt. Je größer der Meßfehler ist, desto stärker müssen die beobachtbaren Wirkungen sein, um nicht als Zufall interpretiert werden zu müssen. Die Verfahren zur Analyse von unabhängigen und abhängigen Variablen können selbst wiederum methodischen Prozeduren unterzogen werden, es können Methodenstudien zur Meßfehlerbestimmung durchgeführt werden (vgl. Abschnitt 4.2).

Informationsbox 3

Einige Mißverständnisse über die Bildung von Untersuchungsgruppen

Eine oft geäußerte Behauptung lautet, daß die in psychologischen Experimenten untersuchten Gruppen repräsentativ für die Personen sein sollten, über die Aussagen intendiert sind. Da dann feststellbar ist, daß diese Gruppen z.B. häufig aus Studierenden bestehen, aber Schlußfolgerungen auf Menschen in (Wirtschafts-)Organisationen gezogen werden, werden solche Untersuchungen als wenig aussagekräftig bezeichnet. Die Antwort hierauf lautet:

1) Die Personen sollen solche Merkmale haben, die für die untersuchte Fragestellung *relevant* sind. Wenn es für die untersuchte Fragestellung bedeutungslos ist, ob es sich bei den Personen um Organisationsmitglieder handelt, dann sind Studierende ebenfalls taugliche Untersuchungspersonen und zudem Untersuchungen an ihnen eher ethisch rechtfertigbar (Schuler, 1980).

2) Für den Fall der Überprüfung von *Kausal*hypothesen können die Gruppen gar nicht repräsentativ sein i. S. einer Stichprobe relativ zu einer Population, da solche Hypothesen sich auf raum-zeitlich unbestimmte Populationen beziehen. Tatsächlich dient eben die Gruppenbildung in einem Experiment lediglich der Aggregation von Meßwerten.

3.3 Drittvariablen und deren Kontrollierbarkeit

Im vorangehenden Abschnitt wurden zusätzlich wirksame Einflüsse auf UV und AV als *unsystematische* Meßfehler behandelt. Zusätzliche Einflüsse können aber auch in zahlreichen Fällen *systematisch* sein. Variablen, die für systematische Einflüsse verant-

wortlich sind, werden als Drittvariablen bezeichnet. Bei organisationspsychologischen Untersuchungen kann eigentlich immer von einer Vielzahl zusätzlich wirksamer Drittvariablen ausgegangen werden.

Drittvariablen (DVn) können an verschiedenen Stellen der (vermeintlichen) Beziehung zwischen einer UV und einer AV ansetzen. So kann ein Zusammenhang zwischen einem bestimmten Führungsverhalten (UV) und der Leistung der Mitarbeiter (AV) von der Situation (DV) abhängen, in der das Verhalten gezeigt wird (vgl. Kapitel 11). Die Situation wirkt hier als Drittvariable auf die Beziehung, bzw. sie moderiert die Beziehung zwischen UV und AV Man spricht hier auch von einer *Moderatorwirkung* der DV bzw. bezeichnet die DV als *Moderatorvariable.* Über die Art der Moderatorwirkung ist bisher noch wenig ausgesagt worden. Die Beziehungen zwischen UV, AV und DV können recht unterschiedlich aussehen (z.B. Moser, 1987); wirkt die DV als Bindeglied zwischen UV und AV, dann wird von der DV auch als einem *Mediator* gesprochen.

Im folgenden wird zunächst eine Reihe möglicher Drittvariablenwirkungen angeführt, die die Aussagekraft organisationspsychologischer Untersuchungen bedrohen können. Um diese Erörterungen übersichtlicher zu gestalten, sei zunächst nochmals daran erinnert, wie das klassische Experimentaldesign aussehen könnte. Einzelnen Personen bzw. einer Gruppe von Personen wird ein Stimulus dargeboten, bzw. sie werden einem Treatment ausgesetzt, und anschließend wird die Reaktion hierauf beobachtet. Der unvorbereitete Leser wird nun möglicherweise überrascht sein, wenn ein Klassiker der Versuchsplanung zur Aussage kommt, daß dieses Design im Prinzip wertlos ist (Cook & Campbell, 1976)! Zur Verdeutlichung gehen wir von einem Beispiel aus, nämlich der Hypothese, daß der Besuch eines Führungstrainings ein effektiveres Führungsverhalten bewirkt. Das Führungstraining besteht aus Vorträgen, Rollenspielen und Gruppendiskussionen. Das effektivere Führungsverhalten wird mittels bestimmter Fragen in einem an die Teilnehmer später ausgeteilten Fragebogen erfaßt. Wie kann die Schlußfolgerung, daß das Führungstraining effektiv war, kritisiert werden? Warum wird das Design (Führungstraining, dann Befragung zu den Auswirkungen des Trainings) manchmal sogar für «wertlos» gehalten, und welche Rolle spielen dabei Drittvariablen? Tabelle 1 bietet einen Überblick zu den wichtigsten Einflußgrößen und nennt entsprechende Alternativhypothesen zur *Erklärung der vermeintlichen Wirkung* des Führungstrainings. (In einigen Fällen wird zudem angenommen, daß eine nichttrainierte Vergleichs- bzw. Kontrollgruppe herangezogen werden kann, worauf weiter unten nochmals eingegangen wird.)

Wer sich Tabelle 1 vergegenwärtigt, mag zunächst resignativ auf Untersuchungen überhaupt verzichten wollen. Dies wäre aber eine überzogene Reaktion, denn zum einen handelt es sich um *Bedrohungen,* die nicht praktisch auftreten bzw. relevant sein müssen, zum zweiten sind *nicht alle Faktoren* bei allen Arten von Untersuchungen von Interesse (beispielsweise die letzten drei), und schließlich existieren *Techniken,* um diese Bedrohungen *zu kontrollieren.* Die wichtigsten Kontrolltechniken, die im folgenden besprochen werden, sind die Einrichtung von Kontrollgruppen, die Randomisierung und die Durchführung von Messungen zu mehreren Zeitpunkten.

Kontrollgruppenbildung und Randomisierung

Experimentieren heißt gezieltes Eingreifen bzw. Vorgabe eines Treatments. Damit ergibt sich, daß sich ohne diesen Eingriff nichts getan hätte (im folgenden vernachlässigen wir zufällige bzw. spontane Veränderungen). Die übliche Vorgehensweise besteht zudem darin, daß der Eingriff zu einem «neutralen» Bezugspunkt relativiert wird, es wird eine *Kontrollbedingung* eingeführt. (Neben solchen materialen Kontrollbedingungen, die als *experimentelle* Kontrollen bezeichnet werden, existieren auch Verfahren der *statistischen* Kontrolle, die hier aber nicht besprochen werden können; vgl. Bortz, 1989.) Eine Kontrollbedingung kann dadurch hergestellt werden, daß eine *Vorhermessung*

Tabelle 1: Drittvariableneinflüsse und die Bedrohung der Gültigkeit von organisationspsychologischen Untersuchungen am Beispiel eines Führungstrainings.

Problem	beispielhafter Effekt
Zeitlich parallele Ereignisse	Nicht das Führungstraining ist für das veränderte Verhalten verantwortlich, sondern der Erholungswert einer Seminarteilnahme oder die veränderte ökonomische Situation des Unternehmens.
Reifung	Die Teilnehmer hatten wenig Erfahrung mit Führung und benötigten einfach noch etwas Zeit, um in ihre Rolle als Führungskraft hineinzuwachsen.
Messung/Testung	Bestandteil des Trainings sind Informationen darüber, wie das Meßinstrument auszufüllen ist, um einen hohen Wert zu erzielen (siehe auch Abschnitt 2).
Instrumentierung	Eigenschaften des Fragebogens sind für die vermeintliche Wirksamkeit verantwortlich; Beispielsweise können Fragen nach der selbsteingeschätzten Veränderung des Führungsverhaltens als (suggestive) Aufforderung wirken, eine solche zu beschreiben.
Statistische Regression	Möglicherweise wurden nur Personen mit (zufälligerweise) weniger Führungsqualitäten zum Training geschickt; ihr Verhalten wird sich auch ohne Training verbessern, da bei einer zweiten Messung (z.B. der Führungsqualitäten) die Extremwerte zur Mitte tendieren.
Selektionseffekt	Das Training ist nur für die spezifische Gruppe wirksam (z.B. weil sie eine Vorauslese besonders motivierter Personen darstellte), die sich von Vergleichs- bzw. Kontrollgruppen unterscheidet.
Mortalität	Personen, denen das Training wenig gebracht hat (oder für die es sogar schädlich war), schieden vorzeitig aus, und ihr Verhalten wurde nicht evaluiert.
Wechselwirkung mit Selektionseffekt	Mehrere der oben genannten Drittvariablen können gemeinsam auftreten und Wirkungen vortäuschen. Wird das Verhalten der trainierten Gruppe mit einer anderen (nicht trainierten) verglichen, dann können z.B. zeitlich parallele Ereignisse, Reifung oder Instrumenteneffekt nur bei den Trainierten aufgetreten sein.

Tabelle 1 (Fortsetzung): Drittvariableneinflüsse und die Bedrohung der Gültigkeit von organisationspsychologischen Untersuchungen am Beispiel eines Führungstrainings.

Problem	beispielhafter Effekt
Unklarheiten über die Richtung der Kausalwirkung	Da im Beispiel das Training der Befragung vorausgegangen ist, kann die Richtung der Wirkung (wenn überhaupt, dann) nur vom Training zur Befragung gehen. Problematisch ist aber der Sachverhalt dann, wenn Ursache und Wirkung zum gleichen Zeitpunkt erhoben werden. Wird beispielsweise die Teilnahme an einem Training und dessen Wirkung mit der gleichen Methode erfragt, so kann folgendes geschehen: Wer verändertes Verhalten bei sich beobachtet, gibt eine Trainingsteilnahme an, die anderen nicht; hier könnte aber die Angabe «Trainingsteilnahme» durch die Verhaltensänderung verursacht sein.
Diffusions- oder Imitationseffekte	Wenn nach dem Training das Verhalten in einer Kontrollgruppe ähnlich verbessert ist, so kann dies daran liegen, daß sich die Trainingsinhalte «herumsprechen».
Kompensatorisches Treatment	Ein (vermeintlich) wirksames Training kann durch kornpensatorische Maßnahmen ausgeglichen werden; die Kontrollgruppe erhält eine «Entschädigung».
Kornpensatorische Rivalität	Die nichttrainierte Kontrollgruppe kann versuchen nachzuweisen, daß sie das Training «nicht nötig hat», und reagiert ebenfalls mit verändertem Verhalten.
Demoralisierung	Die Wirksamkeit des Trainings besteht darin, daß die Kontrollgruppen in ihren Leistungen nachlassen/demoralisiert wurden, weil sie nicht teilnehmen «durften».
Eigenschaften der beteiligten Personen	Die trainierten Personen bringen bestimmte Eigenschaften mit, die zwar nicht in den Kontrollgruppen vorliegen, insgesamt aber untypisch für Organisationsmitglieder sind (z.B. wird das Training mit bezahlten Studierenden durchgeführt).
Eigenschaften des Settings	Die Wirksamkeit des Trainings ist auf eine bestimmte Umgebung oder eine bestimmte Organisation beschränkt.
Zeiteffekte	Das Training ist wirksam, weil es zu einem bestimmten Zeitpunkt stattfand oder weil es gerade dem Zeitgeist entsprach.

stattfindet. Unterscheiden sich dann die Beobachtungswerte vor und nach dem Treatment, so könnte auf dessen Wirkung geschlossen werden. Eine andere Möglichkeit besteht in der Heranziehung einer *Kontrollgruppe*. Wenn sich die Beobachtungswerte in Versuchs- und Kontrollgruppe unterscheiden, dann kann auf eine Treatmentwirkung geschlossen werden. Voraussetzung ist hier allerdings eine sinnvolle Vergleichbarkeit der Ausgangssitutation von Versuchs- und Kontrollgruppe. Dies sei an einem Beispiel aus der Forschung zur organisationalen Sozialisation verdeutlicht, der Hypothese, daß sich neue Organisationsmitglieder an die Normen und Werte ihrer Arbeitsgruppe angleichen. Im Extremfall wäre hier eine Gruppe von Personen keiner organisationalen Sozialisation auszusetzen - aber an welche Normen und Werte ihrer Arbeitsgruppe sollten sie sich dann anpassen? Eine solche Kontrollgruppe wäre aber möglicherweise dann sinnvoll, wenn eine Alternativerklärung geprüft wird, die ganz anders lautet, nämlich, daß sich die Werte junger Mitarbeiter aufgrund eines Alterseffekts gewandelt haben. Dieser Alterseffekt ließe sich auch in besagter Kontrollgruppe aufzeigen, womit die Hypothese eines organisational bedingten Sozialisationseffekts kritisierbar wäre.

Eine naheliegende Vermutung ist, daß sich Versuchs- und Kontrollgruppe systematisch unterscheiden, und zwar bereits *vor* dem Treatment. Für den Umgang mit diesem Effekt bieten sich verschiedene Strategien an. Eine Strategie lautet, die vermuteten Wirkungen von Drittvariablen zu zufälligen Fehlern zu machen. Man mag etwa annehmen, daß das Führungstraining (= UV) nur bei einem Teil der trainierten Personen zu einem veränderten Führungsverhalten führt. Will man nun trotzdem untersuchen, ob ein Trainingseffekt vorliegt, so müßte sichergestellt sein, daß die entsprechenden Ursachen in einer als Vergleichsbasis herangezogenen Kontrollgruppe *gleichermaßen* wirksam sind. Dies wird erreicht, indem eine zufällige Aufteilung der Untersuchungspersonen in Versuchs- und Kontrollgruppe vorgenommen wird. Solche Aufteilungen werden z.B. mit Zufallszahlentabellen vorgenommen und *Randomisierungen* genannt. Experimente, bei denen Kontrollgruppen ohne Randomisierung zugrundeliegen, werden als *Quasi-Experimente* bezeichnet (Cook & Campbell, 1976).

Gruppenbildung, Heranziehung einer Kontrollgruppe und Randomisierung werden mittlerweile so selbstverständlich genannt, daß sie als scheinbar unverzichtbare Merkmale von *Experimenten* gelten. Tatsächlich aber handelt es sich hierbei *nicht* um a priori notwendige Kontrolltechniken, und zudem sind auch hiermit nicht alle Drittvariablenprobleme gelöst. So können Diffusions- oder Imitationseffekte, kompensatorisches Treatment, kornpensatorische Rivalität und Demoralisierung auch durch Randomisierung nicht ausgeschlossen werden, und auch die letzten drei in Tabelle 1 genannten Faktoren sind davon unberücksichtigt. Schließlich sei darauf hingewiesen, daß die Frage der Beziehung zwischen Treatment bzw. Wirkung des Treatments und deren Interpretation im Abschnitt über Konstruktvalidität (Abschnitt 4.2) besprochen wird.

Wenn eine Randomisierung ausgeschlossen ist und man nicht gänzlich auf eine Kontrollgruppe verzichten will, dann kommen sogenannte nicht-äquivalente Kontrollgruppen und Meßwiederholungen in Frage (die meisten Designs mit nicht-äquivalenten Kontrollgruppen enthalten auch Meßwiederholungen). Diese beiden Designgruppen werden in den nächsten Abschnitten behandelt. Sind zudem begründete Annahmen über die Wirkung von Drittvariablen formulierbar, so können die Untersuchungspersonen zunächst hierauf untersucht und dann so in Versuchs- vs. Kontrollgruppen aufgeteilt werden, daß sich die Werte in den Gruppen paarweise gleichen. Dieses Verfahren wird als *Parallelisierung* oder auch «Matching» bezeichnet. Des weiteren können die Drittvariablen explizit als *weitere unabhängige Variablen* berücksichtigt bzw. die UV kann unterschiedlich «intensiv» gestuft werden (vgl. hierzu den nächsten Abschnitt). Eine vierte Möglichkeit besteht schließlich darin, die Drittvariablen zwar zu erheben, sie aber erst im Nachhinein in der Phase der Datenauswertung mittels statistischer Verfahren

Tabelle 2: Probleme des unbehandelten Kontrollgruppendesigns mit Vor- und Nachtest am Beispiel der Wirkung eines veränderten Arbeitszeitsystems in zwei Filialen.

Problem	beispielhafter Effekt
Wechselwirkung von Selektion und Reifung	Die zwei Filialen unterscheiden sich in ihrem Ausgangsniveau von Fehlzeiten, was mit der («gereiften») Identifikation der Mitarbeiter mit dem Unternehmen zusammenhängt. Die scheinbare zusätzliche Wirkung der Änderung des Arbeitszeitsystems ist tatsächlich auf schon zuvor bestehende Unterschiede in der Identifikationsbereitschaft mit der Filiale zurückzuführen, die wiederum bei den beiden Gruppen unterschiedlich starke Effekte auf die Fehlzeiten hat.
Instrumentierung	Die Meßinstrumente können für beide Gruppen nicht gleich geeignet sein. Im Extremfall können die Fehlzeiten bereits so niedrig sein, daß sich keine Änderung mehr zeigen kann (sogenannte Decken- bzw. Bodeneffekte).
Differentielle statistische Regression	Die Vergleichsfiliale kann zwar so ausgewählt sein, daß ihre Ausgangswerte vergleichbar gering/hoch waren, aber die Werte verändern sich dort eher aufgrund eines Regressionseffekts.
Wechselwirkung von Selektion und zeitlich parallelen Ereignissen	Zeitlich parallel zur Einführung der neuen Arbeits-Zeitregelung werden einige Mitarbeiter, die einen Großteil der Fehlzeiten in der einen Filiale verursachten, pensioniert.

(z.B. mittels Kovarianzanalyse) auszuschließen.
Im allgemeinen ist die Randomisierung sowohl der Parallelisierung als auch dem expliziten Einbezug weiterer UVs vorzuziehen, da Drittvariablenwirkungen oft nur vage bekannt sind. Gelingt es, eine relevante Variable zur Parallelisierung zu finden, dann steigt jedoch nach Stelzl (1984) die Teststärke.

Nicht-äquivalente Kontrollgruppen

In vielen Anwendungsfällen ist keine echte Randomisierung möglich (vgl. Pro und Kontra in Cook & Campbell, 1979, S. 347ff.). Dennoch möchte man gerne «Vergleichsgruppen» bzw. «nicht-äquivalente Kontrollgruppen» heranziehen. Cook und Campbell (1979) schlagen für solche Fälle insgesamt acht Designvarianten vor, von denen im folgenden drei besprochen werden.
Betrachten wir ein *Beispiel:* Die Hypothese sei, daß die Einführung von Gleitzeit zu einer Reduktion der Fehlzeiten führt. Diese Maßnahme wird in einer Bankfiliale (A) eingeführt; zum Vergleich wird eine zweite Bankfiliale (B) herangezogen, in der diese Maß- nahme ausbleibt. Nun unterscheiden sich aber z.B. Belegschaftszusammensetzung und Arbeitsbedingungen; wie kann hier also ein Vergleich noch sinnvoll sein? Ein *unbehandeltes Kontrollgruppendesign mit Vor- und Nachtest* kann eine ganze Reihe von Bedenken trotz Fehlen von Randomisierung beseitigen: In beiden Filialen wird vor und nach Einführung der neuen Arbeitszeitregelung der Fehlzeitenanteil gemessen, und wenn die Abnahme in Filiale 1 stärker ist, dann war die Regelung wirksam. Die Probleme dieses Designs sollen nun ausführlich erörtert werden, da es auch unabsichtlich entstehen kann, wenn z.B. eine zuvor geplante Randomisierung mißglückt ist. Cook und Campbell zählen Bedrohungen der Schlußfolgerungen aus solchen Untersuchungen auf, die in Tabelle 2 zusammen mit jeweils einem Beispiel wiedergegeben werden.

Zu Beginn dieses Abschnitts wurde erläutert, weshalb Kontrollbedingungen wichtig sind, um die Wirksamkeit von Maßnahmen beurteilen zu können. Verschiedenste Ursachen können aber die Einrichtung von unabhängigen Kontrollgruppen erschweren, wenn nicht gar unmöglich machen (vgl. Kapitel 4.2). Die Alternative lautet dann, mehrfache Messungen an der gleichen Person bzw. der gleichen Gruppe, sogenannte *Meß-Wiederholungen,* vorzunehmen. Als Kontrollbedingung fungiert hier also die gleiche Untersuchungseinheit. Eine Meßwiederholung liegt bereits dann vor, wenn Vortest und Nachtest vorliegen. Sogenannte Zeitserienuntersuchungen gehen aber von mehreren Messungen vor und nach einem Treatment aus. Auch bei diesen Designs gibt es bedeutsame Probleme. Zunächst entsteht verstärkt Reaktivität, die der Meßwiederholung zuzuschreiben ist (z.B. das Gefühl, persönlich kontrolliert zu werden, oder auch Lerneffekte, wenn mehrfach mittels der gleichen Fragen geprüft wird, ob sich ein tatsächlicher Lernfortschritt ergeben hat). Ein weiterer Kritikpunkt ist die Ökonomie. Demgegenüber besteht der Vorteil solcher Designs in der Registrierung von Reifungsprozessen. Da diese Art von Untersuchungsdesigns in den letzten Jahrzehnten kaum eine Rolle in der Organisationspsychologie spielte, soll hier nicht weiter darauf eingegangen werden. Werden zwei Gruppen verglichen, nachdem eine davon ein Treatment erhalten hat, und weiß man nicht, ob sie sich schon zuvor unterschieden haben, dann sind Aussagen über Differenzen nach dem Treatment zweifelhaft. In der Organisationspsychologie sind solche Vergleiche (und auch die Kritik hieran) Legion: Unterscheidet sich das Verhalten von Verkäufern und Verwaltungsbeamten, weil die einen anders sozialisiert werden, weil sie trainiert werden usw., oder unterscheiden sich diese beiden Gruppen schon vor Berufsbeginn? Die sich hier anbietende Möglichkeit ist die Prognose von Interaktionseffekten. Hierzu wird wie folgt vorgegangen: Untersuchungs- und Vergleichsgruppen werden jeweils unterteilt, und es werden dann differentielle theoretisch begründete Datenmuster prognostiziert. Die vorhergesagten Interaktionseffekte erlauben es, eine zentrale Alternativerklärung auszuschließen. Hierzu nochmals das Beispiel: Von Personen, die in Verkäuferischen Berufen tätig sind, wird erwartet, daß sie es lernen, sich selbst positiv darzustellen. Zum Vergleich habe man eine Gruppe von Personen, die in verwaltenden Berufen tätig ist. Die naheliegende Alternativinterpretation im Falle gefundener Unterschiede lautet, daß sich die Personen bereits zuvor unterscheiden: Es handelt sich demnach um einen Selektionseffekt anstatt um einen Sozialisationseffekt. Ein Sozialisationseffekt wird aber dann wieder plausibler, wenn die Dauer der Berufstätigkeit einbezogen wird: Bei den Verkäufern müßte sich ein Zusammenhang zeigen, bei den verwaltenden Berufen aber nicht. Cook und Campbell (1979) bezeichnen solche Designs als *«Nur-Nachtest Designs mit vorhergesagten Interaktionen».*

Dieses Design kann zur Validierung von Eignungsdiagnostica angewendet werden. Wenn man die Hypothese hat, daß ein Personalauswahlverfahren nur für eine bestimmte Gruppe von Arbeitsplätzen relevant ist, dann sollte sich auch *nur* für diese Arbeitsplätze eine Korrelation zwischen den Werten in den Personalauswahlverfahren und dem beruflichen Verhalten ergeben.

3.4 Verzicht auf experimentelle Variation

Ein wesentliches Merkmal des Experiments ist die gezielte Manipulation der UV. Was geschieht nun, wenn hierauf verzichtet wird? Eine Möglichkeit wäre, eine «natürliche» Variation abzuwarten und deren Wirkung zu erfassen. Solche nichtexperimentelle Untersuchungen lassen sich nach Stone (1978) in zwei Varianten einteilen, *«cross-sectional studies»* und *«correlational studies».* Der erste Untersuchungstyp zeichnet sich dadurch aus, daß mehrere Gruppen verglichen werden, die sich jeweils in einem bereits vorgegebenen Merkmal unterscheiden. In korrelativen Untersuchungen werden in der Regel interindividuelle Unterschiede analysiert. Beide Designvarianten werden auch als *ex-post-facto-Designs* bezeichnet, da keine experimentelle Kontrolle über die

unabhängigen Variablen besteht. Beide Designs haben zudem die Gemeinsamkeiten, daß sie in der Organisationspsychologie oft vorkommen und daß ihre Aussagekraft oft kritisch beurteilt wird. Hier wird die erste Designvariante als *Feldstudie* und die zweite als *korrelative Untersuchung* bezeichnet.

Ein Grundmerkmal von *Feldstudien* besteht darin, eine entsprechende Variation abzuwarten und diese dann zu evaluieren. Obwohl ein solches Design nicht gänzlich uninterpretierbar ist, ist es doch mit zahlreichen Schwächen behaftet: Man verfügt über keine oder keine äquivalente Kontrollgruppe, kennt nicht das Ausgangsniveau der Untersuchungspersonen und kann sich nicht einmal sicher sein, ob die Variation nicht von anderen Variablen überlagert wurde; zudem sind Replikationen ausgeschlossen. Damit sind noch mehr Bedrohungen der Gültigkeit getroffener Aussagen existent, als diese in Tabelle 1 angeführt wurden. Die Ergebnisse solcher Designs sind nur mit größten Vorbehalten zu interpretieren. Die beste Empfehlung lautet, daß man am besten darauf verzichtet und stattdessen experimentelle Designs wählt. Stone (1978) schlägt vor, daß möglichst viele der eventuell störend wirkenden Variablen berücksichtigt, d.h. erhoben und auf ihre Wirkung geprüft werden. Dieses Vorgehen birgt allerdings zwei Probleme in sich, zum einen gleicht diese «Kontrollstrategie» eher einem Herumprobieren, und zum anderen entsteht die Gefahr, daß zufällige Zusammenhänge «entdeckt» bzw. überinterpretiert werden.

Eine weitere Alternative zum Experiment besteht darin, natürliche Variationen (z.B. Unterschiede zwischen Personen oder Organisationen) zu erfassen und diese mit abhängigen Variablen in Beziehung zu setzen. Betrachten wir ein Beispiel für solch eine *korrelative Untersuchung:* Gegeben sei die Hypothese, daß eigene Zielsetzung zu Leistungsverbesserung führt. Nun kann so vorgegangen werden: Es werden Personen mit unterschiedlich starker Tendenz zur Zielsetzung gesucht, und es wird geprüft, ob ihr Leistungsniveau unterschiedlich hoch ist. Damit werden dann *korrelative* Analysen durchgeführt.

Auch ohne Treatments sind Aussagen über korrelative Zusammenhänge zwischen Variablen möglich. Grundüberlegung ist hier, daß von einer natürlichen Variation von Merkmalen ausgegangen werden kann und diese miteinander in Beziehung setzbar sind. Beispiele hierfür sind die unterschiedlich starke Ausprägung von Absentismus und Arbeitszufriedenheit oder von Intelligenz und Ausbildungserfolg. In der organisationspsychologischen Forschung spielen korrelative Analysen deshalb eine Rolle, weil die Daten hierfür als relativ leicht erhebbar gelten (z.B. mittels Fragebogen), und weil es verführerisch, teilweise aber auch begründet möglich ist, Kausalaussagen über diese korrelativen Beziehungen zu machen, zumal in manchen Anwendungsbereichen die Feststellung korrelativer Beziehungen ausreicht. Bei der *Planung einer kausalen Analyse korrelativer Daten* sind aber u. a. folgende Probleme zu beachten:

- Da die Richtung der Verursachung nicht aus den Daten erkennbar ist - anders als bei einem Experiment, wo die UV zeitlich vor der AV liegt - benötigt man Annahmen über die zeitliche Abfolge. (Kann man annehmen, daß Arbeitsunzufriedenheit Absentismus verursacht oder eher das Umgekehrte?)
- Des weiteren ist es im allgemeinen günstig, wenn die beteiligten Variablen durch verschiedene Erhebungsmethoden registriert werden. Zumindest eine potentielle Quelle methodischer Probleme ist die Erhebung verschiedener Aspekte oder Variablen durch die gleiche Methode der Datenerhebung. So wurden und werden sowohl Arbeitszufriedenheit als auch deren Determinanten mit den gleichen Fragebogen bzw. ähnlichen Fragetypen, Formulierungen oder gar Betonungen in den Fragen erhoben. Wie Staw (1984) bemerkt, ist ein Zusammenhang zwischen Arbeitszufriedenheit und Arbeitsplatzmerkmalen vor allem dann nicht überraschend, wenn die Arbeitsplatzmerkmale mit wertbeladenen Frageformulierungen erfaßt werden (z. B. «die Tätigkeit ist herausfordernd»).

- Es fehlt die Kontrolle von Drittvariablen, die zeitlich vor, parallel zu oder auch nach der unabhängigen Variable auf die abhängige Variable wirken. Nehmen wir das einfache Beispiel, daß nur *eine* weitere Variable vorliegt, so kann diese UV und AV beeinflussen, aber auch zwischen UV und AV als Moderator bzw. Mediator wirken.

Die Ergebnisse korrelativer Untersuchungen sind stichprobenabhängig. Wenn die Merkmale zwischen den untersuchten Personen nur gering variieren, dann wird auch der auffindbare Zusammenhang zwischen Merkmalen gering ausfallen. Ellsworth (1977) meint, daß natürliche Variationen geringer sind als experimentell erzeugbare, wie überhaupt ein noch offenes Problem korrelativer Analysen darin besteht, ob eine Vergleichbarkeit von kurzfristiger Manipulation («state») und «natürlicher» Variation («trait») gegeben ist.

Korrelative Untersuchungen gehen von der «natürlichen» bzw. bereits vor einer Untersuchung gegebenen Variation von Merkmalen aus und setzen diese mit anderen Reaktionen in Beziehung. Diese «natürliche» Variation kann aber sowohl zu homogen als auch zu heterogen ausfallen. Ist die Variation zu homogen, so werden die Zusammenhänge geringer sein und ist sie zu heterogen, so werden die Zusammenhänge überschätzt. Ursachen solcher Varianzverzerrungen sind nichtzufällige Stichprobenziehungen bzw. Rekrutierungen von Untersuchungspersonen, die zufällig, absichtlich und manchmal sogar unvermeidlich sein können. Das Problem der Varianzhomogenität wird u.a. in der Eignungsdiagnostik behandelt. So wurde beispielsweise die These aufgestellt, daß die Mitglieder in Organisationen sich über die Zeit hinweg ähnlicher werden (vgl. Moser, 1991). Werden also Zusammenhänge zwischen Persönlichkeitsmerkmalen und Leistung analysiert, so fallen diese bei Organisationsmitgliedern geringer aus als bei Bewerbern.

Varianzheterogenität liegt üblicherweise beim Vergleich von Extremgruppen vor. Beispielsweise könnte der Zusammenhang zwischen Autonomie und Identifikation mit der Organisation untersucht werden, indem Fließbandarbeiter und Manager hinsichtlich dieser zwei Variablen untersucht werden. Der Zusammenhang dürfte hier überschätzt werden, wenn z.B. auf alle Arbeiter, alle Manager oder auch alle Organisationsmitglieder generalisiert werden soll.

Wie ist nun mit solchen Problemen umzugehen? Die erste Empfehlung lautet, eine «repräsentative» Stichprobe bzw. «repräsentative» Untersuchungspersonen heranzuziehen, wenn man am *Ausmaß* des Zusammenhangs interessiert ist. Eine zweite Möglichkeit besteht darin, daß die getroffenen Aussagen entsprechend relativiert werden. So hat der Extremgruppenvergleich *demonstrative* Zwecke gehabt, kann aber nicht typische Aussagen machen; er eignet sich aber sehr wohl, um Kausalhypothesen zu prüfen. Die dritte Behandlungsweise lautet, die Varianzen bzw. Zusammenhangsmaße zu korrigieren (Formeln hierzu bei Lienert, 1989).

4. Methoden der Datengewinnung

Untersuchungen werden von Datenerhebungen begleitet, teilweise bereiten Datenerhebungen Untersuchungen vor, teilweise sind Untersuchung und Datengewinnung identisch, teilweise dienen Datenerhebungen der Überprüfung der Wirkung von Interventionen. Methoden der Datengewinnung sowie Kriterien zu deren Entwicklung und Bewertung sind Gegenstand der psychologischen Diagnostik. Die diagnostische Methodenlehre wurde vornehmlich an praktischen Fragen der Messung interindividueller Unterschiede entwickelt (z. B. Moser & Schuler, 1989). Die nachstehenden Überlegungen gelten aber auch für die Diagnose von Gruppen, Organisationen oder Arbeitsplätzen, sind also demnach für alle in diesem Lehrbuch zu findenden diagnostischen Teilkapitel gleich relevant. In diesem Abschnitt soll zunächst ein Überblick über einige der häufiger verwendeten Methoden zur Datengewinnung gegeben werden. Im Anschluß werden Kriterien zur Beurteilung

dieser Methoden angeführt. Auch an dieser Stelle sei darauf hingewiesen, daß es wiederum Methoden gibt, die der Entwicklung und Verbesserung von Methoden der Datengewinnung dienen.

4.1 Beschreibung von Methoden der Datengewinnung

Interviews

Die unmittelbarste Methode der Datengewinnung ist die Befragung bzw. das Interview. Interviewer und Interviewter stehen in relativ direktem Kontakt, und sie tauschen untereinander (= «inter») ihre Sicht (= «view») von Sachverhalten aus. Der Interviewer kann die Fragen zuvor spezifiziert haben oder sie aus dem Gesprächsverlauf entwickeln. Die Antworten des Interviewten können zunächst offengelassen werden oder bereits (zumindest beispielhaft) vorformuliert sein; diese Vorformulierungen können dann dem Befragten vorgelegt werden oder auch lediglich dem Interviewer als Auswertungshilfe dienen. Die vorstehenden Überlegungen machen bereits deutlich, daß es eine Vielzahl von Interviewvarianten gibt; zudem werden sie oft durch weitere Datenerhebungsmethoden unterstützt (z.B. Interviewleitfäden). Interviews können auch in der Form von Gruppendiskussionen, eventuell mit Unterstützung von Metaplan-Technik, stattfinden. Weitere Varianten werden in Kapitel 4.2 angeführt.

Die *Vorteile von Interviews* lassen sich wie folgt zusammenfassen (vgl. Stone, 1978): Interviews sind flexibel und ermöglichen Nachfragen. Im Vergleich zu schriftlichen Methoden der Datengewinnung ist der Rücklauf bzw. die Antwortbereitschaft höher, und die sprachlichen Anforderungen an die Untersuchungspersonen sind geringer. Zur *Gestaltung der Interviewsituation* lassen sich insbesondere folgende Empfehlungen formulieren:

- die Statusdifferenz zwischen Interviewer und Interviewtem sollte gering sein;
- der Expertenstatus von Interviewer und Interviewtem sollte geklärt sein (der Interviewer ist Experte für Interviews, aber der Interviewte ist Experte für den Gegenstand des Interviews);
- der Erhalt der Motivation des Interviewten sollte durch unterstützende Maßnahmen gewährleistet sein (eventuell Wechsel von Themen, Frageformen usw.);
- das Interview sollte ohne das Beisein unbeteiligter anderer Personen stattfinden.

Im allgemeinen empfiehlt es sich, die Interviewer zu trainieren. Inhalte solcher Trainings sollten unter anderem sein:

- Der Interviewer sollte die Funktion des Interviews als Datengewinnungstechnik kennenlernen.
- Der Interviewer sollte das Interviewen praktisch üben.
- Der Interviewer sollte die Bewertung von Antworten trainieren.
- Der Interviewer sollte, falls möglich, an der Entwicklung des Interviews beteiligt sein.

Zu den *Nachteilen von Interviews* zählen u. a. die mangelhafte Standardisierung, der Aufwand und die (unkontrollierte) Beeinflussung der Untersuchungspersonen durch den Interviewer. So ergibt der Vorteil einer möglichen Nachfrage bei unklaren Äußerungen der Befragten zugleich den Nachteil einer mangelnden Standardisierung. Auch wenn es noch eine Reihe weiterer Argumente gegen Interviews gibt (z.B. Ökonomie-Überlegungen), so lassen sich doch die Haupteinwände auf den geringen Standardisierungsgrad zurückführen. Damit sind aber sogenannten Interviewereffekten, auch «Versuchsleitereffekte» genannt, Tür und Tor geöffnet. Insbesondere sogenannte «Rosenthal-Effekte», d.h. die Beeinflussung der Untersuchungspersonen durch Hypothesen oder Wünsche des Untersuchungsleiters, wurden ausführlich untersucht (Rosenthal & Rubin, 1978; Bungard, 1984). Über die weiter oben angeführten Trainingsmaßnahmen hinaus werden als Maßnahmen zur Kontrolle von Versuchsleiter- bzw. Interviewereffekten u. a. vorgeschlagen (vgl. Bungard, 1987):

- die Durchführung von Doppelblinduntersuchungen (d.h. weder Untersuchungs-

personen noch Untersuchungsleiter kennen die Forschungshypothesen);
- der Einsatz möglichst vieler und unterschiedlicher Versuchsleiter/Interviewer;
- die Minimierung des Kontakts zwischen Interviewer und Interviewtem. Dies kann dadurch vorgenommen werden, daß der Interviewablauf möglichst hoch strukturiert wird oder daß Datengewinnungsmethoden eingesetzt werden, die den Interviewer vom Interviewten trennen. (Dies kann allerdings Akzeptanzprobleme zur Folge haben.)

Auch wenn es die bereits kurz erwähnten Möglichkeiten zur Reduzierung dieser Probleme gibt, so sollten Interviews, gerade wenn die Kontrollmöglichkeiten eher gering sind, vor allem zur Hypothesengenerierung eingesetzt werden, während die präzisere Hypothesenprüfung objektiveren Verfahren (vgl. unten) überlassen werden sollte. Auf Empfehlungen zur Formulierung einzelner Fragen sei hier verzichtet. Ausführliche Vorschläge finden sich bei Bouchard (1976).

Fragebogen

Fragebogen sind vermutlich die am häufigsten eingesetzten Methoden zur Datengewinnung in der Organisationspsychologie (Stone, Stone & Gueutal, 1990). Die Gründe hierfür sind u. a., daß sie (im Vergleich zu Interviews) weniger Kosten verursachen, die Daten auch von weniger Qualifizierten erhoben werden können, die Durchführung standardisiert, anonym und in Gruppen erfolgen kann und die Bögen auch postalisch die Untersuchungspersonen erreichen können.

Viele Probleme und entsprechende Empfehlungen zur Konstruktion von Fragebogen sind denen zum Interview ähnlich. Besonderheiten bei Fragebogen ergeben sich in der Regel dadurch, daß Fragebogen auf Papier (neuerdings auch öfter per Computer) vorgegeben werden und damit oft ohne Anwesenheit eines Interviewers, der Erläuterungen geben oder Rückfragen stellen kann. Fragebogen, die nur schriftlich ausgegeben bzw. verschickt werden, erfordern daher noch stärker die Prüfung der Fragenqualität und der Antwortalternativen als mündliche Befragungen; grundsätzlich stellen Fragebogen höhere Anforderungen an die Verständnisleistungen von Untersuchungspersonen. Dies drückt sich auch darin aus, daß der Anteil fehlender oder unvollständiger Fragebeantwortungen größer ist (vgl. auch Stone et al., 1990). Durch die räumliche und zeitliche Distanz von Untersuchungsleiter und Untersuchungsperson wird das Problem, daß die Beantwortungshäufigkeit, -geschwindigkeit und -qualität eher zu wünschen übrig läßt, eher noch verstärkt. Dies liegt gerade im organisationalen Kontext daran, daß es kaum eine Möglichkeit gibt, persönliches Vertrauen zwischen Befragtem und Fragesteller aufzubauen.

Empfehlungen zur Verbesserung des Rücklaufs von Fragebogen, von denen allerdings einige auch für andere Methoden der Datengewinnung anwendbar sind, lauten:
- Der Fragebogen sollte möglichst einfach und übersichtlich aufgebaut sein.
- Die Person/Organisation, die fragt, sollte identifiziert sein.
- Der Fragebogen sollte einen Titel tragen, der die Bedeutung der Fragen für den Befragten deutlich macht.
- Ankündigungsschreiben können verschickt werden, da sie den Rücklauf erhöhen.
- Anreize für die Beantwortung der Fragen können sein:
  - (minimale) Geschenke
  - Informationen über die Ergebnisse werden zugesagt
  - die Bedeutung des Ergebnisses für die befragte Person bzw. Organisation wird erläutert.
- Das Beilegen frankierter und adressierter Rückumschläge erhöht den Rücklauf beträchtlich.
- Begleitbriefe sollten (kurz!) die Bedeutung deutlich machen, wenn irgendwie möglich Anonymität garantieren und persönlich (möglichst von einer «Autorität») unterzeichnet sein.

Mittlerweile existieren bereits eine Vielzahl von Fragebögen. Anstatt also für jede Un-

tersuchung einen neuen Fragebogen zu entwickeln, dürfte deren (zumindest teilweise) Verwendung auch deshalb interessant sein, weil dann der Vergleich mit den Ergebnissen anderer Untersuchungen vereinfacht wird. Zudem ist zu bemerken, daß in vielen Fällen (zumindest nachträglich!) Zweifel an der Aussagekraft der Antworten auf einzelne Fragen entstehen. Sammlungen von bewährten Fragebogen finden sich z.B. in ZUMA (1983).

Beobachtungen

Interviews und Fragebogen haben das gemeinsame Merkmal, daß die Untersuchungspersonen sich *zu Fragen äußern.* Die Datengewinnung kann aber auch unter Verzicht auf direkte Fragen erfolgen bzw. diese ergänzen. Beobachtungen des Verhaltens können z.B. Interviews ergänzen (etwa Mimik und Gestik). Beispiele in der Organisationspsychologie, bei denen Beobachtungen im Mittelpunkt stehen, sind die Durchführung von Zeitstudien, das Führen von Tagebüchern (Selbstbeobachtung) oder die Beobachtung von Gruppendiskussionen in Assessment Centers. Beobachtungen können frei, anhand von Checklisten oder mit Hilfe von Beurteilungsskalen erfolgen. Ein bekanntes Verfahren zur Dokumentation von Beobachtungen des Diskussionsverhaltens stammt von Bales (vgl. Kapitel 15).

Beobachtungen werden oft recht rasch durch Interviews oder Fragebogen ersetzt, da sie zeitaufwendig sind (und dadurch den Arbeitsablauf stören können) und der Beobachter besonders gut trainiert sein muß. Andererseits ermöglichen Beobachtungen die Gewinnung von Daten, die nicht introspektiv zugänglich sind. Ein gewichtiges Problem ist jedoch, daß unterschiedliche Datenquellen wie z.B. Fragebogen vs. Beobachtung zumindest in Einzelfällen diskrepante Resultate erbrachten (vgl. Informationsbox 4).

Informationsbox 4

Womit Industrieforscher ihre Zeit verbringen

Hinrichs (1964) stellte das Stereotyp der Tätigkeitsschwerpunkte von Industrieforschern in Frage. Um zu belegen, welche Bedeutung ***kommunikative*** Aufgaben haben, verwendete er zwei Methoden. Zum einen sollten die untersuchten Forscher ***schätzen,*** mit welchen kommunikativen Aktivitäten sie wieviel an Zeit verbringen, zum anderen wurden sie gebeten, ihr Verhalten während ihrer Arbeit zu protokollieren.

Ein Ergebnis war, daß Industrieforscher mehr als 50% ihrer Arbeitszeit mit kommunikativen Tätigkeiten verbringen. Bemerkenswert ist aber auch der Vergleich der beiden Methoden: Zwar stimmten Schätzung und Direktprotokollierung ***insgesamt*** überein, dabei wurde aber die Dauer von Meetings und informellen Diskussionen geringer geschätzt als protokolliert, während es sich mit der schriftlichen Kommunikation umgekehrt verhielt. Ob sich hinter diesem Ergebnis der Wunsch von Industrieforschern verbirgt, mehr schriftlich und weniger mündlich kommunizieren zu müssen, oder ob die Zeit bei angenehmen Tätigkeiten schneller vergeht als bei unangenehmen, kann hier nur als Frage formuliert werden.

Beurteilungsskalen

Die drei bisher angeführten Methoden der Datengewinnung zeichnen sich durch eine Art der Datenregistrierung aus, die eine weitgehende direkte Protokollierung menschlichen Erlebens und Verhaltens vornimmt. Zwar kann man auch bei diesen Methoden schon von *Messen* sprechen; Messen im strengeren Sinne des Wortes heißt allerdings vor allem das regelgeleitete Zuordnen von *Zahlen* zu Objekten oder Ereignissen bzw. zu deren variablen Ausprägungen. Einfache Fälle von Messen bestehen beispielsweise darin, dem Geschlecht, der Abteilungszugehörigkeit oder der Berufsausbildung Zahlen zuzuordnen. Beurteilungsskalen beruhen auf Grundlagen, die in der sogenannten *Meßtheorie* behandelt werden. Im folgenden wird eine komprimierte Darstellung zweier häufig in der Organisationspsychologie verwendeter Skalen, die im allgemeinen die *Intensität* einer Reaktion erfassen, vorgenommen.

Die wohl am häufigsten verwendeten Skalentypen sind Rating- bzw. Likertskalen. Grundprinzip ist, daß Aussagen vorgegeben

werden und der Befragte das Ausmaß der Zustimmung bzw. Ablehnung äußert. Die Skalenausprägungen können verbal verankert bzw. erläutert sein, teilweise sind dies auch nur ein Teil der Skalenausprägungen. Ratingkalen liegen in mannigfaltigen Variationen vor; da im allgemeinen die Gleichabständigkeit zwischen den Skalenpunkten angenommen werden kann, sind die mit ihnen erhaltenen Ergebnisse zudem oft bequem mit parametrischen statistischen Verfahren auswertbar.

Vom Format her ist das *Semantische Differential* den Ratingskalen ähnlich. Auch hier werden Skalen vorgegeben, die aber nun mit Adjektiven verankert sind. Diese Adjektive sollen die untersuchten Personen verwenden, um ihren Eindruck oder ihre Vorstellung von einem Sachverhalt zu beschreiben. Die Adjektivpaare können entweder in standardisierter Form (vgl. Hofstätter, 1977) oder auf den spezifischen Untersuchungszweck hin konstruiert vorgegeben werden. Im letzteren Fall sollte aber insbesondere die tatsächliche Bipolarität der Skalenenden einer Prüfung unterzogen werden. Zur Interpretation der Ergebnisse empfiehlt sich in der Regel entweder der Vergleich mit der Einschätzung anderer Sachverhalte oder/ und die Reduktion der Daten mittels Faktorenanalyse oder multidimensionaler Skalierung.

Testverfahren

Psychologische Testverfahren dienen ähnlich wie Beurteilungsskalen der Ermittlung einer individuellen Ausprägung bzw. Reaktion. Während aber bei der Beurteilung das *Urteil* im Mittelpunkt steht, sind Testverfahren eher Verfahren, um *Reaktionen zu* provozieren. (In einer Reihe von Persönlichkeitstests werden aber auch (implizite) Selbstbeurteilungen erfragt.) Zudem werden diese Reaktionen nicht «direkt» interpretiert, sondern als Indikatoren für zugrundeliegende Dimensionen oder Konstrukte behandelt. Vor allem bei Persönlichkeitstests führt dies in der Öffentlichkeit immer wieder zu Mißverständnissen. Insofern Reaktionen provoziert und dann erfaßt werden, sind psychologische Testverfahren Experimenten nicht unähnlich (vgl. Abschnitt 3.1). Ein gewichtiger Unterschied besteht allerdings darin, daß bei Tests grundsätzlich *mehrere* Reaktionen provoziert und anschließend aggregiert werden. Die Grundlagen psychologischer Testverfahren werden in der Testtheorie behandelt (Lienert, 1989). Psychologische Testverfahren genügen im Idealfall einer ganzen Reihe von Kriterien, die vor allem in der Eignungsdiagnostik intensiver diskutiert werden (vgl. die Darstellung in Kapitel 9).

Simulationen

Bei den soeben besprochenen Testverfahren werden in der Regel relativ wenige Reiz- bzw. Stimulusaspekte gezielt vorgegeben und die Reaktionen hierauf erfaßt. Noch extremer verhält es sich bei Experimenten. Dies mag wohl auch der Grund sein, warum Fromkin und Streufert (1976) Simulationen als bemerkenswerte Alternativen zu Experimenten bezeichnen, obwohl die anderen Methoden zur Datengewinnung sich ebenfalls von Simulationen in den nachstehenden Kriterien unterscheiden. Während im Fall von klassischen Experimenten lediglich wenige Stimulusaspekte gezielt variiert werden, steht bei Simulationen die Konfrontation mit komplexen «lebensweltlichen» Ereignissen im Vordergrund. *Beispiele für Simulationen* sind unter anderem

- die Vorgabe einer Rolle und anschließende Durchführung eines *Rollenspiels;*
- die Entwicklung einer *Arbeitsprobe* zur Diagnose von motorischen Fertigkeiten;
- die Beschreibung beruflicher Situationen und anschließende Erfragung des (hypothetischen) Verhaltens;
- die Programmierung einer *komplexen Simulation* auf einem Personal-Computer und Registrierung des «Umgangs» von Untersuchungspersonen mit der Dynamik des Systems.

Gemeinsam ist den meisten Simulationen relativ komplexes Stimulusmaterial, eine Replikation von Systemen und eine Darbietung von Ereignissen. Die abhängigen Va-

riablen bestehen in mehr oder weniger freiem Verhalten der Untersuchungspersonen. In der Regel dauern Simulationen länger als z.B. die Bearbeitung von Testverfahren. Schließlich zeichnen sich Simulationen durch ein hohes Involvement der Personen aus (auf mögliche begriffliche Unterscheidungen sei hier nicht eingegangen; vgl. Fromkin & Streufert, 1976). *Probleme,* die bei der Entwicklung von Simulationen zu berücksichtigen sind, und entsprechend aufwendige Vorprüfungen erfordern, lauten:

Die Komplexität der Simulation kann die Untersuchungsperson überfordern.

Simulationen können als unrealistisch empfunden werden und sind es auch prinzipiell, da nie alle Realitätsausschnitte simuliert werden können.

Die Komplexität der Simulation macht sie eher fehleranfällig.

Das soziale Umfeld während der Auseinandersetzung mit der Simulation ist künstlich.

Die Untersuchungspersonen kommen eher in Versuchung, die Grenzen «auszutesten».

Der meist dynamisch-interaktive Charakter von Simulationen führt zu einem geringen Standardisierungsgrad von Simulationen.

Die zugrundeliegenden Prozesse unterscheiden sich von den in der Realität ablaufenden Prozessen.

(Vor-)Erfahrungen mit dem betreffenden Realitätsausschnitt oder auch dem Simulationsmedium (z. B. Personal-Computer) beeinträchtigen die Gültigkeit der Aussagen.

Für die nahe Zukunft ist eine Zunahme des Interesses an Simulationen zu erwarten. Hier ist z.B. an computerunterstützte Simulationen zu Trainingszwecken zu denken.

Nicht-reaktive Verfahren

Allen bisher dargestellten Methoden der Datengewinnung ist das Wissen der Untersuchungspersonen über die Datenerhebung gemeinsam. Demgegenüber ist das gemeinsame Merkmal der ansonsten heterogenen Verfahren zur Datengewinnung, die zu der nun zu erörternden Gruppe von Meßverfahren zählen, die unterstellte (oder erhoffte) fehlende *Reaktivität* der Datengewinnung (vgl. Abschnitt 4.2). Beispiele sind physische Spuren (z.B. abgelaufene Parkettböden), Archivdaten (z. B. die Geschichte des organisationsinternen Aufstiegs) oder Daten, die mit versteckten Kameras gewonnen werden. Weitere Beispiele finden sich bei Webb, Campbell, Schwartz und Sechrest (1975) sowie Bungard und Lück (1974). Die Grenzen für die Anwendbarkeit nicht-reaktiver Methoden sind deren begrenztes Vorliegen, deren ambivalente Interpretierbarkeit sowie ethische Aspekte (Verletzung der informationalen Selbstbestimmung).

4.2 Gütekriterien von Methoden zur Datengewinnung

Die vorgenommene Trennung bei der Beschreibung verschiedener Methoden der Datengewinnung diente didaktischen Zwekken. In der organisationspsychologischen Praxis werden oftmals mehrere Methoden zu einer kombiniert. So können beispielsweise Beobachtungen von Gruppendiskussionen mittels Beurteilungsskalen vorgenommen werden. In vielen Fällen empfiehlt sich die Kombination von Methoden in einem Stufenmodell: Kennenlernen eines Sachverhalts, aber auch (eventuelle) inhaltliche Vorbereitung einer Intervention profitieren vor allem in neuen Gebieten eher von qualitativ orientierten Methoden (Interview, Beobachtung), während Beschreiben und Aufklären von Zusammenhängen sowie Unsicherheitsreduktion und Entscheidungsvereinfachung eher durch quantitativ standardisierte Verfahren möglich sind. Bereits in den einzelnen Teilabschnitten wurden Vor- und Nachteile der jeweiligen Methoden genannt (vgl. auch Abschnitt 6). Im folgenden wird auf *formale* Aspekte zur Bewertung von Methoden zur Datengewinnung eingegangen.

Betrachten wir ein Beispiel, nämlich daß die Wirkung einer Informationsstrategie auf Einstellungen bestimmt werden soll. Hierzu

wird Bewerbern ein Film über eine Organisation gezeigt und im Anschluß ein Fragebogen vorlegt, um Aussagen über diese Organisation zu erfassen: Der Film «repräsentiert» oder operationalisiert eine Informationsstrategie und der Fragebogen soll die Einstellung zur Organisation erfassen. Die im Einzelfall eingesetzten Methoden (Film, Fragebogeninhalte) können demnach von den Absichten bzw. späteren Interpretationen (Informationsstrategie, Einstellungen) differieren. Diese Differenz entsteht zum einen dadurch, daß immer zusätzlich Meßfehler Verschiedenster Art auftreten können und zum anderen dadurch, daß «tatsächlich» etwas anderes getan bzw. gemessen wurde als geplant oder interpretiert. Ganz allgemein kann man von zwei Meßproblemen sprechen, dem Reliabilitäts- und dem Validitätsproblem.

Reliabilität

Nur in den wenigsten Fällen werden in organisationspsychologischen Untersuchungen perfekte Beobachtungen, Messungen oder Beurteilungen vorgenommen. Einige Methoden zur Reduzierung von Fehlern wurden bereits in Abschnitt 3 besprochen. In diesem Abschnitt geht es um die Methoden zur Bestimmung des Ausmaßes der Fehler. In der Klassischen Testtheorie (z. B. Lienert, 1989) werden hier zunächst zwei Begriffe angeführt, die Objektivität und die Reliabilität (Meßgenauigkeit). Die Objektivität kann untergliedert werden in Durchführungs-, Auswertungs- und Interpretationsobjektivität. Die Objektivität und vor allem Verfahren zu deren quantitativer Bestimmung werden oft der Reliabilität subsumiert. (Manchmal werden auch hohe Reliabilitätskoeffizienten als Objektivitätskoeffizienten bezeichnet.) Mit «Test» wird hier also jede Art von Meßinstrument bezeichnet (Moser & Schuler, 1989).

Die Reliabilität läßt sich über vier verschiedene korrelative Methoden bestimmen, Retestreliabilität, Parallelität, Split-Half-Reliabilität und Interne Konsistenz. In allen Fällen werden mindestens zwei Meßwertreihen bestimmt und ihr Zusammenhang ermittelt (in der Regel als Produkt-Moment-Korrelation). Die Reliabilität kann Werte zwischen 0 und 1 annehmen.

Retestreliabilität (Stabilität)

Eine Möglichkeit zur Bestimmung des Meßfehlers ist die wiederholte Datenerhebung. Zeigen sich die Daten als stabil, so kann der Meßfehler als gering angenommen werden. Probleme sind, daß bei kurzfristiger Wiederholung Erinnerungs-, Lern- oder Ermüdungseffekte auftreten können. Ist der Zeitabstand zu groß, dann kann sich hingegen der Untersuchungsgegenstand (etwa der Arbeitsplatz) verändert haben.

Parallelität (Beurteilerübereinstimmung, Äquivalenz)

Von Fall zu Fall besteht Gelegenheit, von einem «Meßgegenstand» (z. B. einen Arbeitsplatz) mehrere Meßwertreihen zu erhalten. Bei Intelligenztests und einigen Persönlichkeitstests existieren Parallelformen, die den gleichen Personen vorgegeben werden können. Unter der Annahme, daß die Parallelformen tatsächlich das gleiche messen, ist die Korrelation zwischen den Meßwertreihen als Schätzung der Reliabilität auffaßbar. Ein in der Organisationspsychologie häufig verwendetes Maß für Parallelität ist die Beurteilerübereinstimmung. Hier sind dann die verschiedenen Beurteiler (z.B. die Inhaber vergleichbarer Arbeitsplätze) die «parallelen Meßmethoden».

Split-Half-Reliabilität (Testhalbierung)

Unter der Annahme, daß ein Meß- oder Testverfahren eine ausreichende Zahl von Beobachtungen zuläßt, kann es halbiert werden, um dann die Korrelation zwischen den beiden Testhälften zu bestimmen. Diese Hälften können dadurch hergestellt werden, daß die geraden und die ungeraden Items zusammengefaßt werden («odd-even-Methode») oder daß einfach die zuerst verwendete Hälfte des Verfahrens mit der zuletzt verwendeten verglichen wird. Letztere Vorgehensweise empfiehlt sich aber nicht bei solchen Verfahren, die unter Zeitbegrenzung vorgegeben werden. Der Reliabilitäts-

koeffizient, der auf der Testhalbierung basiert, bezieht sich nur auf die Hälfte des Tests. Mit Hilfe der weiter unten beschriebenen Spearman-Brown-Formel kann die Reliabilität für den gesamten Test geschätzt werden. Eine Verallgemeinerung der Testhalbierung ist die folgende Methode.

Interne Konsistenz (Homogenität)

Bei vielen Testverfahren oder deren Teilen beziehen sich mehrere Fragen auf die gleiche «latente» Dimension. Damit sind Testverfahren auch als Folgen von Meßwiederholungen auffaßbar, und die Reliabilität der Verfahren als mittlere Korrelation über alle Items hinweg schätzbar. Der nahezu routinemäßig angegebene Koeffizient für diese Interne Konsistenz oder Homogenität ist Cronbachs α. Hierzu seien einige Anmerkungen gemacht: Zunächst handelt es sich bei diesem Koeffizienten eher um eine *untere Grenze* der Reliabilität. Zudem gleicht seine Verwendung oft einem Paradox: Beispielsweise werden unterschiedliche Fragen zu einem Sachverhalt gestellt, um gerade ein breiteres Spektrum abzudecken, dann aber wird oft so getan, als ob es sich doch um einen homogenen Sachverhalt handelt. Schließlich ist auf das Homogenitätsparadox hinzuweisen, nach dem bei sehr homogenen Skalen die Chance, einen Sachverhalt aufzuklären, wieder abnimmt.

Die angeführten Koeffizienten können unterschiedlich groß sein, da es ja nicht *die* Reliabilität, sondern jeweils nur Reliabilitäts*aspekte* einer Methode zur Datengewinnung gibt. Zudem ist die Reliabilität eines Verfahrens zunächst einfach ein mehr oder weniger abstraktes Gütemaß von Meßmethoden und sollte möglichst groß sein. Eine konkretere Anwendung ist die Möglichkeit, den Effekt der Verlängerung oder Verkürzung einer Methode zur Datengewinnung (z.B. durch Hinzunahme bzw. Wegnahme von Fragen) berechnen zu können. Die Reliabilität ist des weiteren Grundlage der Bestimmung von Vertrauensintervallen der Meßwerte einzelner Personen. Schließlich sei auf die Möglichkeit einer *Korrektur* der Beziehungen zwischen meßfehlerbehafteten Variablen sowie auf einen allgemeinen Zusammenhang mit der Validität einer Methode zur Datengewinnung (die Validität kann höchstens so groß werden wie die Quadratwurzel der Reliabilität) hingewiesen (vgl. für Beispiele Moser & Schuler, 1989).

Konstruktvalidität

Das gemeinsame Merkmal aller Untersuchungsvarianten besteht darin, daß Untersuchungspersonen Reizen ausgesetzt sind und ihre Reaktionen hierauf erfaßt werden. Werden aber diese Reize und Reaktionen richtig oder gültig oder valide interpretiert? Wer diese Frage stellt, die genaugenommen bei *jeder* Untersuchung gestellt werden kann, interessiert sich für Fragen der Konstruktvalidität. Die Konstruktvalidität einer Untersuchungsanordnung oder eines Meßverfahrens bezeichnet die Gültigkeit der jeweils abgeleiteten Schlüsse. So werden beispielsweise der eingangs erwähnte Film als Informationsstrategie und die Werte im Fragebogen als Einstellungen *interpretiert*. Mangelnde Konstruktvalidität kann sich in zwei Problemen ausdrücken: Die Operationalisierung kann nur einen Teil des Konstrukts abdecken (z. B. daß der Film Aspekte der Unternehmenskultur nur unvollständig zu beschreiben vermag) oder sie kann darüber hinausgehen (z.B. daß der Film aufgrund des Auftretens attraktiver Schauspieler den Bewerbern zu viel verspricht).

In den Abschnitten 3.1-3.4 wurden bereits eine Reihe von Kontrolltechniken besprochen, die letztendlich alle Voraussetzungen für die Gültigkeit der Interpretation eines Untersuchungsergebnisses sind. Immer wieder wurde gefragt: Hat A tatsächlich B verursacht bzw. bewirkt? In diesem Abschnitt soll auf ein anderes Problem eingegangen werden, nämlich ob A tatsächlich die UV und B tatsächlich die AV operationalisieren bzw. repräsentieren, oder welche Konstrukte sich hinter A und B «verbergen». Zusammenhänge zwischen A und B werden hierbei in der Regel als gegeben vorausgesetzt. Im folgenden werden nun einige *Quellen mangelnder Konstruktvalidität sowie Maßnahmen zu deren Analyse und Prävention* behandelt.

Operationalisierung

Eine sorgfältige Analyse der vorgenommenen Operationen bzw. Definitionen sollte am Anfang einer Untersuchung stehen. *Beispiele* hierfür sind etwa, ob ein untersuchtes Konzept (z.B. «Organisationsklima») sich auf ein strukturelles Merkmal einer Organisation, die Wahrnehmung dieses Merkmals oder die Einstellung zu diesem Merkmal bezieht (Schwab, 1980). Die Verwendung bewährter Untersuchungsmethoden und ein kompetentes Literaturstudium sind hier geeignete Maßnahmen. Zudem kann eine Prüfung der Operationalisierung durch *Experteneinschätzungen* vorgenommen werden: Es kann die Frage gestellt werden, ob in einem Fragebogen überhaupt die richtigen Inhalte aufgenommen wurden. Wird allerdings mit dem Expertenstatus zu großzügig umgegangen, so kann eher von «Augenscheinvalidität» der Methode gesprochen werden.

Besonderheiten der Treatmentintensität (Instrumentierung)

In organisationspsychologischen Anwendungsgebieten kann eine Variable scheinbar nicht wirksam sein, weil ihre Intensität zu schwach ist. Zum Beispiel mag sich die Vermutung ergeben, daß es nicht die Informa*tionsstrategie* (einseitig vs. zweiseitig), sondern lediglich die Tatsache des Informierens (vs. Nicht-Informierens) überhaupt ist, die eine veränderte Einstellung bewirkt. Deshalb kann es sinnvoll sein, mehrere Abstufungen bzw. Variationen des Filminhalts vorzunehmen (z. B. einseitig, zweiseitig kombiniert mit Kurzform, Langform; vgl. auch Abschnitt 3.3).

Besonderheiten der Methoden zur Datengewinnung

Eine andere Interpretation könnte sein, daß *es* der *Film* ist, der eine positive Einstellung bewirkt, und weniger der konkrete Inhalt. Daher kann es sich anbieten, die gleichen Informationen schriftlich oder über einen Film vorzugeben, um so den Einfluß der Methode untersuchen zu können. Solche Methodenvergleiche gleichen allerdings oft genug kleinen Kunstwerken; wie ist es beispielsweise möglich, nur das Medium und nicht auch noch weitere Komponenten zu verändern (z.B. das Verhältnis von Bild und Text)? Aber auch auf seiten der abhängigen Variablen können Methodeneffekte entstehen. Beispielsweise können Fragebogen für Antworttendenzen anfällig sein und deshalb nicht nur das Konstrukt «Einstellung zum Unternehmen» erfassen, sondern auch die Tendenz, Aussagen (irgendwelcher Art) zuzustimmen. Ein oft diskutiertes Verfahren zur Analyse von Methodeneffekten wird in Informationsbox 5 beschrieben.

Versuchsleitereffekte

Werden Untersuchungen im direkten Kontakt mit den Untersuchungspersonen durchgeführt, so können Untersuchungs- bzw. Versuchsleitereffekte auftreten. Damit sind beabsichtigte oder unbeabsichtigte Beeinflussungen des Verhaltens der Untersuchungspersonen durch den Versuchsleiter gemeint. Möglichkeiten zur Reduzierung dieses Effekts wurden in Abschnitt 4.1 im Zusammenhang mit Interviews besprochen.

Reaktivität als Sensibilisierung oder Placeboeffekt

Bei jeglicher Art von Intervention, die zusätzlich einer Evaluation unterzogen wird, muß mit zwei Problemen gerechnet werden, die weniger durch die konkreten Inhalte als durch die Tatsache der Intervention und Evaluation an sich entstehen (vgl. auch Kapitel 4.2). Die Messung des Ausgangs- oder Endzustandes kann das zu Messende beeinflussen und zu einer *Sensibilisierung* führen. Wenn beispielsweise nach einem Training eine Überprüfung des Effektes dieses Trainings stattfindet, dann kann bei den untersuchten Personen die Vermutung entstehen, daß aus den Ergebnissen für sie mehr oder weniger wünschenswerte Konsequenzen entstehen. So mag ein Anreiz bestehen, den Trainingseffekt zu verschleiern, um keine erhöhten Zielsetzungen erwarten zu müssen oder die Teilnehmer des Trainings wollen diese Art von «Belohnung» auch für ihre Kollegen und versuchen deshalb, den Trai-

Informationsbox 5

Ein Verfahren zur Analyse von Methodeneffekten: Die MTMM-Analyse

Die Multi-trait-multi-method-Analyse (MTMM-Analvse) wurde von Campbell und Fiske (1959) eingeführt, um zu untersuchen, ob es gelingt, Merkmale («traits») unabhängig von der verwendeten Methode zu erfassen. Im Prinzip werden für eine solche Analyse mindestens zwei Methoden benötigt, mit denen jeweils mindestens zwei Merkmale erhoben werden. Betrachten wir ein Beispiel: Ein Personalberater möchte die Merkmale «Durchsetzungsfähigkeit» und «Kommunikationsgeschick» bei Bewerbern um eine Führungsposition ermitteln. Hierzu wendet er zwei Methoden an, Gruppendiskussionen und Interviews zu bisherigen Erfahrungen in der Biographie der Bewerber. Als Ergebnis liegen die Urteile von Beobachtern vor. Diese Meßwerte sind nach Campbell und Fiske (1959) nun vier Fragen zu unterziehen:

(1) Konvergente Validität: Die Korrelation zwischen zwei Methoden, die das gleiche Merkmal erfassen, sollte statistisch und praktisch bedeutsam sein. (Beispiel: Die Werte für Durchsetzungsfähigkeit sollten in der Gruppendiskussion und im Interview ähnlich ausfallen.)

(2) Diskriminante Validität:
- (a) Die Korrelation zwischen den gleichen *Merkmalen* sollte größer sein als die Korrelation zwischen den *Methoden.* (Beispiel: Die Korrelation zwischen den Werten für Durchsetzungsfähigkeit sollte größer sein als die Korrelation zwischen den Gesamtwerten von Gruppendiskussion und Interview.)
- (b) Die Korrelation zwischen verschiedenen Merkmalen, die mit der gleichen Methode erfaßt wurden, sollte nicht größer sein als die Korrelation zwischen den gleichen Merkmalen, die mit verschiedenen Methoden erfaßt werden.
- (c) Die Muster der Merkmalskorrelationen sollten gleich ausfallen, und zwar unabhängig von Methodenhomogenität oder -heterogenität.

Daß auch bei diesem Verfahren noch Interpretationsspielräume bleiben, zeigt sich z.B. daran, daß ein Autor die gleiche Korrelation (r = .52) in einer Studie als Beleg für Konvergenz, in einer anderen Untersuchung aber für Diskriminanz bezeichnete (vgl. Schwab, 1980). Zudem gibt es auch Fälle, in denen die Konvergenz alternativer Methoden bei gleichem Konstrukt nicht erwünscht bzw. gut erklärbar ist. So mag besagter Personalberater etwa die Auffassung vertreten, daß Durchsetzungsfähigkeit in einer Gruppendiskussion etwas anderes ist als das, was in einem Interview zum Vorschein kommt.

ningseffekt als besonders hoch erscheinen zu lassen. Die Reaktion auf die ***Meßsituation*** an sich wird in der weiterführenden Methodenliteratur unter dem Stichwort «sozial erwünschtes Antwortverhalten» untersucht.

Neben den eine Untersuchung begleitenden Messungen können auch ***Untersuchungen «an sich»*** wirken: Die bloße Tatsache einer Intervention kann bereits (erwünschte wie unerwünschte) Verhaltens- und Erlebenskonsequenzen haben. Zum Beispiel wurden die berühmten Hawthorne-Experimente auch so interpretiert, daß sie die Wirksamkeit einer Maßnahme unabhängig vom konkreten Inhalt demonstriert hätten (vgl. Kapitel 2). Die bloße Tatsache einer Maßnahme zeigte den beteiligten Mitarbeiterinnen, daß etwas mit ihnen geschehe. Untersuchungen, die in diesem Sinne wirksam sind, erzeugen ***Placeboeffekte.*** Manchmal wird auch in Anlehnung an besagte Untersuchungen von einem Hawthorne-Effekt gesprochen. In einzelnen Fällen beschränken sich Untersuchungen auf Messungen (z.B. im Falle von Mitarbeiterbefragungen). Damit sind dann Sensibilisierung und Placeboeffekt identisch. Die Reaktivität kann sich vielfältig ausdrücken. Die wichtigsten Aspekte sind Vermutungen über die Untersuchungshypothesen und entsprechendes Agieren sowie Versuche, sich positiv darzustellen.

Mit der Reaktivität von organisationspsychologischen Interventionen kann wie mit allen Artefaktquellen umgegangen werden (Moser, 1987). Zunächst können sie ***ignoriert*** bzw. als wenig relevant bei der jeweiligen Intervention bezeichnet werden. So mögen bei der Bewertung der Wirkung einer Unterweisung in eine neue CNC-Maschine zwar beide Reaktivitätsprobleme auftreten. Entscheidend dürfte aber letztendlich das Verhaltensergebnis sein, daß ein Facharbeiter die Maschine am Ende des Kurses bedienen kann. Selbst wenn Reaktivität auftritt - etwa der «Lerneffekt» vor allem entsteht, weil es interessant ist, an einer Untersuchung teilzunehmen bzw. in Erwartung einer besonders schwierigen Prüfung - so ist sie zunächst Vernachlässigbar. Wenn jedoch die Unterweisung routinemäßig durchzuführen ist und auf die «Prüfungen» verzichtet wird,

dann können mit der dann später fehlenden Reaktivität auch die Lerneffekte ausbleiben und es wäre ratsam gewesen, weniger großzügig mit der Interpretation der ursprünglichen Ergebnisse gewesen zu sein.

Eine Möglichkeit zur Kontrolle von Reaktivitätseffekten besteht in der *experimentellen* Kontrolle der Wirkung der Messung durch zusätzliche Kontrollgruppen in einem *Solomon-Plan* bzw. der Wirkung der Intervention durch einen modifizierten Solomon-Plan (Nachreiner, Müller & Ernst, 1987). Als weitere Varianten seien genannt (vgl. auch Bungard, 1984):

- die Befragung der Untersuchungspersonen nach der eigentlichen Untersuchung;
- das Non-Experiment, d.h. Beschreibung des Untersuchungsvorhabens und Erhebung der vermuteten Reaktionen;
- explizites Nennen der Untersuchungshypothesen in einer (weiteren) Kontrollgruppe;
- Ablenkung vom Untersuchungszweck;
- die Betonung durch den Untersuchungsleiter, daß die *individuellen* Daten weniger interessant sind;
- Anonymität der Untersuchungspersonen herstellen;
- nicht oder falsches Aufklären der Untersuchungspersonen über den generellen Zweck der Untersuchung;
- Verwendung von Skalen zur sozialen Erwünschtheit;
- Verwendung nichtreaktiver Meßverfahren (vgl. Abschnitt 4.1).

Bemerkenswert ist, daß Reaktivität auch beabsichtigt sein bzw. gezielt provoziert werden kann. Wenn Vorgesetzte bekanntgeben, daß sie Leistungsbeurteilungen durchführen, dann kann dies eine motivierende Wirkung haben, die über die bloße Zurschaustellung von «Aktivität» durch den Mitarbeiter hinausgehen mag. Wer sich bei einem Vorstellungsgespräch gut «darstellen» kann, der mag auch gut als Verkäufer geeignet sein. Und wer an einer Personalentwicklungsmaßnahme teilnimmt und lediglich wegen eines «Hawthorne- Effekts» motivierter an seine Arbeit geht, hat durch sein Verhalten die «Wirksamkeit» der Maßnahme belegt.

Insgesamt kann die Validität von Untersuchungen oder Meßinstrumenten noch von einer großen Zahl weiterer Faktoren beeinflußt werden. Da viele dieser Faktoren sehr spezifisch sind, sei auf die ausführliche Diskussion in Cook und Campbell (1979) hier lediglich verwiesen.

Abschließende Bemerkungen

Wie wichtig sind Konstruktvaliditätsfragen in der Organisationspsychologie? So wünschenswert eine klare Antwort hierauf wäre, so strittig oder doch zumindest abhängig von der Vielfalt von Zielen und Schlußfolgerungen ist sie. Stellt man eher den technologischen Charakter von Organisationspsychologie in den Vordergrund, dann könnte etwa folgendes behauptet werden: Wenn es das Ziel ist, daß die Produktivität nach Einführung einer neuen Art der Arbeitsorganisation zunimmt und dies geschieht, dann ist danach das Ziel erreicht. Irgendwelche theoretische (oder empirische) Analysen der «Konstrukte» Produktivität oder Arbeitsorganisation sind müßig. Ebenso kann die Möglichkeit, aufgrund biographischer Merkmale beruflichen Erfolg zu prognostizieren, wichtiger sein als die Erklärung der biographischen Merkmalen zugrundeliegenden Konstrukte (Persönlichkeitsmerkmale, Interessen usw.). Biographische Fragebogen oder Arbeitsproben sind ebenso wie Trainingsmaßnahmen als *multiple* Treatments charakterisierbar. Trotz zweifelhafter Konstruktvalidität werden diese Treatmenttypen deshalb eingesetzt, weil sie den Nachweis der *Wirkung* über die *Erklärung* der Wirkung setzen und oft auch erbringen können (vgl. Kapitel 9 und 10). Methoden zur Datengewinnung oder Intervention könnten auch so charakterisiert werden, daß aufgezählt wurde, was mit ihnen leistbar ist. So kann auch das Konzept der *Kriteriumsvalidität* verstanden werden, nämlich als Auskünfte über die Differenzierungskraft oder Prognosekraft einer Methode zur Datengewinnung.

5. Bewertung organisationspsychologischer Untersuchungen

Gegenstand der bisherigen Ausführungen dieses Kapitels waren Methoden zur Untersuchung organisationspsychologischer Fragestellungen. In diesem Abschnitt soll es nun um die Bewertung von Untersuchungsergebnissen gehen. Begonnen wird mit Verfahren der Implementierungskontrolle, danach werden verschiedene allgemeine Fragen der Ergebnisdarstellung behandelt und schließlich wird auf die Möglichkeit quantitativer Reanalysen bereits vorliegender Untersuchungen eingegangen.

5.1 Implementierungskontrolle

Vor, während und nach einer Untersuchung empfiehlt es sich, sogenannte Implementierungskontrollen vorzunehmen. Implementierungskontrollen sollen zum einen als «Frühwarnsysteme» und damit Gelegenheiten für die eventuelle Modifikation von Untersuchungsteilen verstanden werden. Zum anderen haben sie die Funktion, die Wirkung von Untersuchungen im Nachhinein sicherzustellen.

Eine Möglichkeit zur Implementierungskontrolle besteht in der Durchführung von *Pilotstudien*. Hier wird die Untersuchung vollständig oder teilweise an einer relativ kleinen Zahl von Untersuchungspersonen durchgeführt, um z.B. zu prüfen:

- Verständlichkeit von Datenerhebungsmethoden;
- Ablauf zeitlich ineinandergreifender Untersuchungsteile;
- Akzeptabilität von Inhalt oder Umfang einer Untersuchung;
- neutrales Verhalten von Beobachtern, Interviewern oder Versuchsleitern.

Ein Problem von Pilotstudien ist die in der Regel geringe Stichprobengröße, die zufällige Extremfälle als repräsentativ erscheinen lassen kann. Es soll auch Fälle gegeben haben, in denen methodisch fragliche Studien nachträglich als «Pilotstudien» bezeichnet wurden. Bedenklich ist dies vor allem dann, wenn dann dennoch weitreichende Schlußfolgerungen gezogen werden, so als ob es sich um eine Hauptstudie handeln würde (Wortman, 1983). Ein weiteres Problem ist das vorzeitige Bekanntwerden einer Untersuchung, da hierdurch unkontrollierte Erwartungen und Gerüchte in Organisationen entstehen können. Da in Pilotstudien Fehler durchaus einkalkuliert und manchmal sogar provoziert werden (z.B. das «Austesten» der Akzeptabilität von Interviewfragen), können Akzeptanzprobleme entstehen. Zu bedenken ist schließlich die geringe Realitätsnähe, wenn Personen wissen, daß sie «Testpiloten» sind.

Eine weitere Variante von Implementierungskontrollen sind *Expertenurteile* über Wirkung und Akzeptabilität von Untersuchungen. Dies bietet sich etwa bei vermuteten heiklen Inhalten an. Expertenurteile sind vor allem dann zu empfehlen, wenn der Organisationspsychologe nur in Maßen mit den fachlichen Inhalten eines Untersuchungsgebietes oder auch mit den Gegebenheiten in der betreffenden Organisation vertraut ist.

Die häufig bei Experimenten angewandte Methode zur Untersuchung der Wirkung der Untersuchung im Nachhinein sind die *«manipulation checks»*. Sofern Täuschungen über den Inhalt oder das Ziel einer Untersuchung sinnvoll, unumgänglich und ethisch vertretbar waren, kann nachträglich geprüft werden, ob die beteiligten Personen nicht doch entsprechende (oder auch unangemessene) Vermutungen hatten, die das Ergebnis einer Untersuchung systematisch beeinflußten. Solche Überprüfungen können natürlich auch ohne das Vorliegen einer Täuschung sinnvoll sein. Eine weitere Variante von manipulation check ist, zu erfragen, wie die Manipulation der UV gewirkt hat (bzw. ob sie überhaupt «angekommen» ist). Wenn beispielsweise untersucht werden soll, ob ein strukturiertes Interview anders erlebt wird als ein unstrukturiertes Interview, dann kann vor oder nach der Analyse des Erlebens der Verfahren zusätzlich untersucht werden, ob die *Variation* des Interviewtyps überhaupt als strukturiert bzw. unstruktu-

riert erlebt wurde. Bei manipulation checks ist allerdings zu gewährleisten, daß den Personen überhaupt ein direkter Zugang zur Beschreibung ihrer Hypothesen oder Erlebnisse möglich ist und daß sie offen darüber sprechen (vgl. weiterführend Adair & Spinner, 1981, und Kapitel 4.2).

Als vierte Methode der Implementierungskontrolle sei schließlich die *Prüfung der Datenqualität* angeführt. Diese Prüfung kann z.B. darin bestehen, Untersuchungspersonen die Intervention beurteilen zu lassen (z.B. die Verständlichkeit eines Fragebogens). Gemeint sind hier aber vor allem die eher «technisch» orientierten Aspekte, wie z. B. Prüfung der Vollständigkeit von Unterlagen, Nachfaßaktionen im Falle geringer Rücklaufquoten bei schriftlichen Datenerhebungen, Sicherung der Stichprobengrößen unter den verschiedenen Untersuchungsbedingungen, Anonymisierung der Daten (Datenschutz und Respektierung der Privatsphäre) oder Kontrollen bei der Datenerhebung (z.B. Überwachung der Interviewer) und der Datenaufbereitung (Erstellung eines Codeplans; Kontrolle der Dateneingabe in die EDV). Zu den Fragen der Datenqualitätssicherung gehören auch sogenannte Plausibilitätskontrollen auf Datenfehler.

5.2 Ergebnisdarstellung

Die statistische Auswertung von Untersuchungen führt zu zwei paradigmatischen Resultaten, nämlich zu Unterschiedsmaßen (Experimental- bzw. Untersuchungsgruppe unterscheidet sich von Kontrollgruppe) oder zu Zusammenhangsmaßen (wenn Variable x zunimmt, dann nimmt auch Variable y zu). In empirischen Untersuchungsberichten werden bei Gruppenvergleichen meistens t-, d- oder F-Werte berichtet, als Zusammenhangsmaße meistens Produkt-Moment-Korrelationen (r-Werte).

Die in der Organisationspsychologie am häufigsten gewählte Darstellung der Zusammenhänge von Sachverhalten besteht in der Angabe von Korrelationen, so z.B. als Reliabilitäts- oder Validitätskoeffizienten. Wie ist aber deren Höhe zu interpretieren? Eine Vorgehensweise besteht darin, von einer optimalen Korrelation von 1.0 auszugehen. Alle Korrelationen, die geringer sind, könnten als wenig überzeugend betrachtet werden - eine Überlegung, die allerdings unrealistisch sein dürfte. Eine andere Möglichkeit besteht darin, von Erfahrungen bzw. typischen Koeffizienten auszugehen. So werden Effektmaße gelegentlich als klein (d = .20 bzw. r = .10), mittel (d = .50 bzw. r = .30) oder groß (d = 80 bzw. r = .50) beschrieben (Cohen, 1977). Diese Festlegungen entbehren nicht einer gewissen Subjektivität. Als weitere Variante wird oft angeführt, daß das Quadrat der Korrelation als Ausmaß gemeinsamer Varianz interpretierbar ist (wir hatten weiter oben gesehen, daß es eigentlich immer darum geht, Variation zu erklären bzw. aufzuklären). Zwei Probleme können hier auftreten:

- Da die Reliabilitäten der Messung von UV und AV in der Regel beschränkt sind, wird dadurch auch das maximal mögliche Ausmaß von Korrelationen festgelegt.
- Werden zwei unterschiedlich schwierige Variablen korreliert, so kann die maximal mögliche Korrelation wesentlich geringer als r = 1.0 sein.

Trotz dieser Einschränkungen scheinen viele Korrelationen gering auszufallen (vgl. zur ausführlichen Diskussion Moser, 1991). Rosenthal (1990) ist jedoch der Auffassung, daß die Betrachtung der Größe von Korrelationen irreführend sein kann. Selbst relativ geringe Korrelationen können zu einem wesentlichen Prognosegewinn führen. Die grundsätzliche Frage lautet, wie groß eine Korrelation sein muß, um als *bedeutsam* eingestuft werden zu können. Um die praktische Bedeutung von Korrelationen zu visualisieren, werden sie in binomialen Effektgrößen veranschaulicht (binomial effect size display; BESD). Die Korrelation läßt sich als einfache Differenz in Prozentsätzen ausdrücken. Grundlage ist eine Vierfeldertabelle; unabhängige und abhängige Variable werden jeweils als zweistufig angenommen. Eine Korrelation von 0 entspricht einer Zellenbesetzung von jeweils 50%, d.h. das

BESD ist unabhängig von der Stichprobengröße. Ist die Korrelation von 0 verschieden, so wird sie mit 50 multipliziert und die Werte werden von den Zelleninhalten subtrahiert bzw. zu ihnen hinzu addiert, wobei Zeilen- und Spaltensummen 100% ergeben. Zum Beispiel wird eine Korrelation von r = .30 mit Hilfe des BESD wie in Tabelle 3 umgewandelt, wenn zwei dichotome Merkmale zugrundegelegt werden.

Das BESD veranschaulicht eine Korrelation auch wie folgt: Bei einer Korrelation von r = .30 wird im Falle des Auftretens von Merkmal 1 ca. doppelt so oft das Auftreten von Merkmal 2 im Vergleich zum Nichtauftreten von Merkmal 2 erwartet.

Bei der vorstehenden Veranschaulichung von Effekten als Korrelationen wurden die zugrundeliegenden Untersuchungsgruppengrößen bzw. die Frage, ob es sich um einen überhaupt jenseits des Zufalls zu interpretierenden Effekt handelt, vernachlässigt. Nun können solche Effekte aber auch *zufällig* zustandekommen. Muß also nicht zunächst einmal die *statistische Signifikanz* bestimmt werden? Um es gleich vorwegzunehmen: Die Diskussion entstand vor allem deshalb, weil der Begriff der «Untersuchung» unpräzise gefaßt wird: Kausalwirkung und (differenziert abstufbare) Intensität eines Zusammenhangs werden oft nicht klar genug auseinandergehalten.

Untersucht man eine Hypothese, dann können zwei Fehler passieren: man kann sie für bestätigt halten, obwohl sie falsch ist, oder verwerfen, obwohl sie richtig ist. Das in den meisten Statistiklehrbüchern vorgestellte Modell betont nun, daß der erste Fehler gravierend ist oder daß man prüfen soll, ob ein Effekt tatsächlich auch vorliegt. Der «Erfolg» einer Untersuchung wird dann oft mit einem signifikanten Effekt gleichgesetzt, bzw. mit einer geringen Irrtumswahrscheinlichkeit (sogenannter u-Fehler). Dieses Vorgehen ist aus der Perspektive heraus verständlich, daß Hypothesen (möglichst) streng zu prüfen sind (Gadenne, 1984). Wer eine strenge Prüfung vornimmt, ist aber eher bereit, eine Wirkung zu übersehen. Für angewandte Fragestellungen wird dabei oft unterschätzt, daß der Schaden, der durch den hierdurch akzeptierten großen **β-Fehler** entsteht, ebenfalls gravierend sein kann. So können kleine, aber *praktisch* durchaus *bedeutsame Effekte* übersehen werden. Zudem kann der Schaden, der aus der fälschlichen Ablehnung einer Alternativhypothese entsteht, sogar sehr viel größer sein als der einer fälschlichen Ablehnung einer Nullhypothese. (Man denke beispielsweise an das Erkennen einer geringen Belastbarkeit von Flugpiloten, die im Falle eines Übersehens schwerwiegende menschliche und finanzielle Kosten zur Folge haben kann.) Ein weiterer Kritikpunkt am statistischen Testen ist die Künstlichkeit des gewählten Signifikanzniveaus. Warum sollte beispielsweise, wie dies oft geschieht, α = .05 und nicht .06 gesetzt werden?

Derzeit ist die Frage, ob bzw. unter welchen Bedingungen der statistischen oder der praktischen Signifikanz mehr Wert beigemessen werden sollte, nicht entscheidbar. Die Empfehlung lautet daher, im Zweifelsfall neben Signifikanztests auch Konfidenzintervalle, Stichprobengrößen und Effektstärken in Untersuchungsberichte aufzunehmen.

Tabelle 3: BESD einer Korrelation von r = .30

		Merkmal 2: liegt vor	Merkmal 2: liegt nicht vor
Merkmal 1	liegt vor	65	35
	liegt nicht vor	35	65

Kosten-Nutzen-Analyse

Organisationspsychologische Untersuchungen können die Grundlage für Entscheidungen bilden, und ihre Planung und Durchführung beruht auf Entscheidungen. Kosten-Nutzen-Analysen lassen sich sowohl für den Gegenstand von Untersuchungen als auch für die Untersuchung selbst aufstellen. Am

Beispiel der Fluktuation (vgl. zum folgenden Mobley, 1982) lassen sich für das Individuum, die Organisation und die Gesellschaft Kosten- und Nutzenaspekte unterscheiden (vgl. Tabelle 4).

Betrachten wir nur die monetären Kosten für die Organisation, so lassen sich diese unter anderem unterteilen in Zeit für (Verabschiedungs-)Gespräch, Verwaltungsaufwendungen, Abfindungen, Anwerbekosten für neue Mitarbeiter, eventuell höheres Gehalt für neue Mitarbeiter und Kosten der Informierung und Einarbeitung des neuen Mitarbeiters. Alle diese Komponenten lassen sich zumindest in etwa abschätzen. Damit läßt sich die Wirkung einer Maßnahme oft konkreter kommunizieren als durch die Angabe von Effektmaßen. Dies gilt nicht zuletzt für kleine Effekte. So gibt Wanous (1989) den Effekt der realistischen Information von Bewerbern auf deren wahrscheinlicheren Verbleib in der Organisation mit r = .06 an. Hieraus berechnet sich eine Kostenersparnis für nicht erforderlichen Personalersatz zwischen 6 und 12%.

Die Notwendigkeit einer Kosten-Nutzen-Analyse organisationspsychologischen Handelns ist bis heute umstritten. Eine gewisse Vorreiterrolle haben in den letzten Jahren Eignungsdiagnostiker übernommen (Schuler & Guldin, 1991; vgl. auch Kapitel 9). Aber auch zahlreiche andere Bereiche, in denen organisationspsychologische Maßnahmen stattfinden, wurden auch ökonomisch untersucht. Eine Übersicht sowie zahlreiche Beispiele führt Cascio (1982) an; deutschsprachige Darstellungen finden sich bei Wottawa und Thierau (1990) und Kalkulationsvarianten für die Weiterbildung bei Jeserich (1989).

Ergebnisbericht

Die Ergebnisse einer Untersuchung sind in der Regel mündlich und schriftlich zu berichten. Sie sind Auftraggebern, Vorgesetzten, der Fachöffentlichkeit, oft auch den Untersuchungspersonen zu vermitteln. Vorschläge zum Aufbau bzw. zur Gliederung von schriftlichen Untersuchungsberichten finden sich unter anderem in den Richtlinien von Fachzeitschriften, den Empfehlungen der Deutschen Gesellschaft für Psychologie, der American Psychological Association oder auch bei Bortz (1984). Der Zweck dieser Berichte kann verschiedener Natur sein; entsprechend ausführlich bzw. kurz sollten Teilfragen gehalten werden. Für Forschungsberichte ist zu fordern, daß

- die Fragestellungen klar herausgearbeitet werden;
- das methodische Vorgehen nachvollziehbar beschrieben wird;
- methodische Details so beschrieben werden, daß die Untersuchung prinzipiell durch andere wiederholt werden könnte;
- die Ergebnisdarstellung so vollständig ist, daß eine Weiterverarbeitung z. B. in Form von Metaanalysen möglich ist;
- die Ergebnisse sich auf die Fragestellungen konzentrieren (weniger ist meistens mehr!).

In der Diskussion von Untersuchungen sollten zudem die Generalisierbarkeit und Robustheit der Ergebnisse berücksichtigt werden (Brinberg & McGrath, 1985). Zur Beurteilung des Geltungsbereichs einer Untersuchung wird man zunächst fragen, ob bei naheliegenden Änderungen von Randbedingungen bereits andere Effekte auftreten. Hierzu können zum einen die Ergebnisse

Tabelle 4: Kosten und Nutzen von Fluktuation für Individuum, Organisation und Gesellschaft (in Beispielen nach Mobley, 1982).

	Nutzen	Kosten
Individuum	Karriereziele werden verfolgt	psychische Belastung durch Arbeitslosigkeit
Organisation	weniger leistungsfähige Mitarbeiter werden ersetzt	materieller Aufwand für Suche nach Ersatz
Gesellschaft	Migration in neue ökonomische Bereiche	Produktivitäts-Steigerung wird gemindert

ähnlicher Untersuchungen herangezogen werden, zum anderen geben auch Teilanalysen der erhobenen Daten weiteren Aufschluß. Unter Robustheit soll hier allgemein verstanden werden, inwieweit die erhaltenen Effekte anfällig für eine minimale Variation von Teilaspekten der Untersuchung sind. Generalisierbarkeit und Robustheit von Ergebnissen sind aber nicht immer klar trennbar.

5.3 Metaanalyse

Bei der Planung einer organisationspsychologischen Studie, der Wahl einer Interventionsmethode oder der Überlegung, wie ein Sachverhalt zu diagnostizieren ist, stellt sich die Frage nach dem bisherigen Forschungsstand sowie dessen Bewertung. Beim Einarbeiten in ein Forschungsgebiet bietet sich bei der Betrachtung einzelner Untersuchungen oft ein verwirrendes Bild voller scheinbar widersprüchlicher Ergebnisse. Auf solche Literaturanalysen dürfte man aber dennoch vor allem dann angewiesen sein, wenn der originäre Untersuchungsaufwand minimal gehalten werden muß. Eine Möglichkeit besteht dann darin, bereits existierende Daten nochmals zu analysieren, sie einer *Sekundäranalyse* zu unterziehen. Oftmals sind aber die Originaldaten nicht zugänglich bzw. zwischen verschiedenen Studien nicht klar Verknüpfbar, und man muß sich auf die Ergebnisdarstellungen in den Untersuchungsberichten beschränken. Findet man zudem keine integrativen Zusammenstellungen, dann kann man sich an einer *Überblicksarbeit* versuchen. Traditionelle, eher qualitativ vorgehende Überblicksarbeiten haben jedoch eine Reihe von typischen Schwächen:

- die Kriterien für die Literaturauswahl und -bewertung sind intransparent;
- die Art der Integration verschiedener Untersuchungsergebnisse ist oft subjektiv;
- Besonderheiten einzelner Untersuchungen werden zu wenig beachtet;
- Stichprobenfehler (Besonderheiten der Stichproben in Einzeluntersuchungen) werden vernachlässigt;
- kleine Effekte werden übersehen.

Eine alternative Methode der Integration bzw. Zusammenfassung und Bewertung von Studien stellen sogenannte (quantitative) *Metaanalysen* dar. In den 80er Jahren nahm die Zahl von Metaanalysen, d.h. quantitativer Kombination von Ergebnissen verschiedener Untersuchungen, gerade in der Organisationspsychologie exponentiell zu (Hunter & Schmidt, 1990, S.42). Das Vorgehen bei einer Metaanalyse läßt sich in sieben verschiedene Phasen unterteilen, ist aber zudem auch in mancher Hinsicht dem bei einer Einzeluntersuchung vergleichbar. Der Unterschied ist zunächst lediglich, daß die Beobachtungseinheiten statt Personen, Gruppen, Organisationen oder Arbeitsplätze nun *Untersuchungen* sind.

Eine Metaanalyse beginnt mit einer *Klärung des Problemfeldes* (1) sowie der Formulierung von Hypothesen über die Faktoren, die als Erklärungen für Unterschiede zwischen Untersuchungen herangezogen werden könnten. Die anschließende *Literatursuche* (2) sollte möglichst breit angelegt sein und sich die vielfältigen Möglichkeiten der computerunterstützten Literaturrecherchen zunutze machen (Fricke & Treinies, 1985, S. 36 ff.). Sie sollte aber auf jeden Fall um die Durchsicht einschlägiger Fachzeitschriften, Lehrbücher, Monographien, aber auch Dissertationen und «grauer Literatur» ergänzt werden. Diese Literaturrecherche wird zunächst großzügig ausfallen und daher im nächsten Schritt auf *thematische Angemessenheit (3)* bewertet werden müssen. Zudem stellt sich hier die umstrittene Frage der *Äquivalenz von Operationalisierungen* bzw. Konstruktvalidität (4). Wer beispielsweise die Wirkung von Führungstrainings untersuchen will, muß zunächst klären, was (noch) unter Führungstraining zu verstehen sein soll.

Auch der nächste Schritt ist umstritten, nämlich der Umgang mit methodischen Fehlern bzw. mit der methodischen Qualität (5) von Primäruntersuchungen. Teilweise wird der Ausschluß solcher Untersuchungen befürwortet, teilweise wird vorgeschlagen, die methodische Qualität zu bewerten und als Moderator zu berücksichtigen. Dieser Autor tendiert eher zu letzterer Empfehlung.

Tabelle 5: Übersicht zu Untersuchungsartefakten (nach Hunter & Schmidt, 1990, S. 45).

1. Stichprobenfehler
2. Meßfehler bei der abhängigen Variable
3. Meßfehler bei der unabhängigen Variable
4. Dichotomisierung einer natürlich-kontinuierlichen abhängigen Variable
5. Dichotomisierung einer natürlich-kontinuierlichen unabhängigen Variable
6. Varianzeinschränkung in der abhängigen Variable
7. Varianzeinschränkung in der unabhängigen Variable
8. mangelnde Konstruktvalidität der abhängigen Variable
9. mangelnde Konstruktvalidität der unabhängigen Variable
10. Transkriptions-/Schreibfehler
11. Drittvariableneffekte

Die Auswertungsmöglichkeiten (6) in Metaanalysen bestehen aus der Integration von Signifikanztests sowie der Angabe von Konfidenzintervallen oder Effektstärken. Das Auszählen signifikanter Ergebnisse («vote-counting») setzt dabei am wenigsten Informationen in den Ausgangsuntersuchungen voraus. Dieses Vorgehen besitzt allerdings eine geringe Teststärke und berücksichtigt nicht die Stichprobengröße. Eine nächste Möglichkeit sind *integrative Signifikanztests* (Beispiele bei Fricke & Treinies, 1985, S. 66 f). Signifikanztests werden im allgemeinen aber nur in einem ersten Schritt berechnet. Die Angabe eines durchschnittlichen Effektmaßes sowie Aussagen zur Heterogenität bzw. Streubreite der Effekte betrachten beispielsweise Hunter und Schmidt (1990) als viel wichtiger. Um diese Ziele zu erreichen, müssen zunächst die Effektmaße für einzelne Studien bestimmt und in einen einheitlichen Maßstab transformiert werden. Zudem sind die Effektmaße einzelner Untersuchungen artefaktanfällig und damit korrekturbedürftig. Die Artefaktquellen von Einzeluntersuchungen sind in Tabelle 5 zusammengefaßt.

Nach Hunter und Schmidt (1990) sollten zunächst die Korrekturen auf Stichprobenfehler (1), Meßfehler bei Prädiktoren und Kriterien (2 und 3) sowie Varianzeinschränkungen bei Prädiktoren und Kriterien (4 bis 7) vorgenommen werden. Auch für die Bewältigung der Probleme 8,9 und 11 werden von Hunter und Schmidt (1990) erste Vorschläge gemacht.

In einem nächsten Schritt wird die Heterogenität der Effektmaße geprüft, also die Frage, ob sich die Ergebnisse verschiedener Untersuchungen überhaupt unterscheiden. *Erst dann* werden die inhaltlichen Moderatorhypothesen (7) analysiert, so z. B. der Einfluß der Herkunft der Untersuchungspersonen, der verwendeten Erhebungsinstrumente, usw.. Da aber nicht alle Artefakte quantifizierbar sind (z.B. Fehler bei der Dateneingabe), wurde die 75% -Regel postuliert, wonach bereits dann, wenn 75% der Effektmaßvarianz durch die berechenbaren Artefakte erklärbar sind, von *keinem* inhaltlichen Moderatoreffekt auszugehen ist.

Abschließend seien nach wie vor kritische Punkte beim metaanalytischen Vorgehen angesprochen. Zum einen setzen Metaanalysen konkrete Problemformulierungen voraus, um die Einbeziehung irrelevanter Studien zu vermeiden. Vor allem aber müssen Methoden angewendet werden, um die Operationalisierungen in verschiedenen Untersuchungen vergleichbar zu machen. Zweitens kann über die methodische Qualität von Einzeluntersuchungen nicht in allen Fällen Einigkeit erzielt werden. Schließlich wurde bisher nicht diskutiert, ob es sich bei allen Artefaktquellen um echte «Fehler» handelt. Hier ist insbesonders an die Idee des Stichprobenfehlers zu denken. Wenn beispielsweise davon ausgegangen wird, daß die den verschiedenen Untersuchungen zugrundeliegenden Personen der gleichen Population entstammen, dann läßt sich fragen, ob diese Annahme *inhaltlich* gerechtfertigt ist.

Betrachten wir das (fiktive) Beispiel, daß bei Managern, ungelernten Arbeitern und Ingenieuren der Zusammenhang zwischen Arbeitszufriedenheit und Leistung untersucht worden sei; bei der ersten Gruppe ergebe sich eine positive, bei der zweiten eine negative und bei der dritten eine Null-Korrelation. Aus Sicht der Metaanalyse wird nun argumentiert, daß diese Korrelationen auch stichprobenfehlerbedingt unterschiedlich groß sein könnten. Nun kann aber wohl von den wenigsten Personen erwartet werden, daß sie jeglichen Beruf ergreifen könnten. Den Personen in Stichproben von Managern, angelernten Arbeitern und Ingenieuren zu unterstellen, daß sie aus der gleichen Population stammen, mutet jedenfalls seltsam an.
Trotz dieser letzten Vorbehalte kann die Metaanalyse inzwischen als etablierte Methode bezeichnet werden, die insbesondere zur Überblicksgewinnung und Komplexitätsreduktion unverzichtbar werden könnte.

6. Schlußbemerkung

An vielen Stellen wurde nicht nur auf die Möglichkeiten von Methoden, sondern auch auf Probleme aufmerksam gemacht. Die Schärfung des *Problembewußtseins* sollte aber nicht den Blick für die *Möglichkeiten* trüben. Absehbare Probleme erfordern Kontrolle oder Korrektur, nicht aber Verzicht auf Handeln überhaupt. Zudem konnten gerade methodische Entwicklungen in den letzten Jahren die Leistungsfähigkeit der Organisationspsychologie demonstrieren (z.B. Hunter & Schmidt, 1990). Wenn nun nach einem Fazit bzw. nach den wichtigsten Empfehlungen gefragt wird, die sich aus diesem Kapitel ergeben, dann lauten diese (vgl. auch Cohen, 1990):

- besser mehr zu planen und weniger «explorative» Untersuchungen durchzuführen;
- genauer zu prüfen, ob nicht bereits Methoden der Datengewinnung für das jeweilige Problem existieren, statt ständig neue zu «entwickeln»;
- möglichst experimentelle Untersuchungen durchzuführen;
- Vorerprobungen mehr Gewicht beizumessen;
- insbesondere bei korrelativen Untersuchungen große Stichproben zu wählen und wenige Variablen zu erheben (außer, sie sind *wirklich* theoretisch relevant);
- mehr Theorie statt komplexe Datenanalyseverfahren «einzusetzen»;
- Signifikanztests *und* Effektstärke *und* Konfidenzintervalle (wo angemessen) zu berichten.

Wer Methoden anwendet oder methodisch vorgeht, der geht nach einem Plan vor, hat Gründe für sein Vorgehen und ist dabei systematisch. Der Regelcharakter von Methoden besagt, daß sie *richtig* oder *falsch* angewendet werden können. Die Anwendung bestimmter Methoden kann sehr vielfältig begründet werden, und in diesem Kapitel wurden nur einige Gründe oder Kriterien (wie z.B. Reliabilität oder Validität) ausführlicher erörtert. Weitere Argumente sind von Fall zu Fall beispielsweise die erzielbare Eindeutigkeit von Aussagen über das Ergebnis einer Untersuchung, Praktikabilität und Wirtschaftlichkeit, generelle Anwendbarkeit bzw. Routinisierbarkeit, Standardisierungsgrad, Realitätsnähe oder Nebenwirkungsfreiheit (vgl. auch Kapitel 4.2).
Es gibt keine «besten» Argumente für die Auswahl von Methoden. Einige der angeführten Gründe bzw. Kriterien widersprechen sich sogar oft (z.B. Standardisierungsgrad und Realitätsnähe), und auch die Autoren dieses Buches sind sich nicht immer einig, welche Gründe oder Kriterien besonders stark zu gewichten sind. Wie dem auch sei-keine Methode ist ausschließlich Selbstzweck, sondern vor allem *Hilfsmittel,* um *inhaltliche* Fragen zu beantworten.

7. Literatur

Adair, J.G. & Spinner, B. (1981). Subjects' access to cognitive processes: Demand characteristics and verbal report: *Journal for the Theory of Social Behaviour, 11,* 31-52.

Backhaus, K., Erichson, B., Plinke, W., Schuchard-Ficher, Chr. & Weiber, R. (1987). *Multivariate Analysemethoden.* Berlin: Springer.

Bortz, J. *(1984). Lehrbuch der empirischen Forschung.* Berlin: Springer.

Bortz, J. (1989). *Lehrbuch der Statistik* (3. Aufl.). Berlin: Springer.

Bouchard, T.J. (1976). Field research methods: Interviewing, questionnaires, participant Observation, systematic observation, unobtrusive measures. In M.D. Dunnette (Ed.). *Handbook of industrial and organizational psychology* (pp. 363-214). New York: Wiley.

Brutberg, D. & McGrath, J.E. (1985). *Validity and the research process.* Newbury Park: Sage.

Bungard, W. (1984). *Sozialpsychologische Forschung im Labor*. Göttingen: Hogrefe.

Bungard, W. (1987). Zur Problematik von Reaktivitätseffekten bei der Durchführung eines Assessment Centers. In H. Schuler & W. Stehle (Hrsg.), *Assessment Center als Methode der Personalentwicklung* (S. 99-125). Göttingen: Hogrefe/Verlag für Angewandte Psychologie.

Bungard. W. & Lück. H.E. (1974). *Forschungsartefakte und nichtreaktive Meßverfahren.* Stuttgart: Teubner.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant Validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cascio, W.F. (1982). *Costing human resources: The financial impact of behavior in organizations.* Boston, MA: Kent.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cook, T.D. & Campbell, D.T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed.). *Handbook of industrial and organizational psychology* (pp. 223-326). New York: Wiley.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation. Design and analysis issues for field settings.* Chicago: Rand McNally.-

Ellsworth, PC. (1977). From abstract ideas to concrete instances: Some guidelines for choosing natural research settings. *American Psychologist, 72,* 604-615.

Fricke. R. & Treinies, *G.* (1985). *Einführung in die Metaanalyse.* Bern: Huber:

Fromkin: H.L. & Streufert, S. (1976). Laboratory experimentation. In M.D. Dunnette (Ed.), *Handbook of industrial and organisational psychology* (pp. 415-466). New York: Wiley.

Gadenne, V. (1984). *Theorie und Erfahrung in der psychologischen Forschung.* Tübingen: Mohr.

Gordon, M.E., Slade, L.A. & Schmitt, N. (1986). The «science of the sophomore» revisited: From conjecture to empiricism. *Academy of Management Review. 11,* 191-207.

Hinrichs. J.R. (1964). Communications activity of industrial research personnel. *Personnel Psychology, 17,* 193-204.

Hoffmann, H. *(1987). Kreativitätstechniken für Manager* (2. Aufl.). Landsberg: Moderne Industrie.

Hofstätter, P.R. (1977). *Persönlichkeitsforschung.* Stuttgart: Kröner.

Hunter. J.E. & Schmidt. FL. (1990). *Methods of meta-analysis.* Newbury Park: Sage.

Jeserich, W. *(1989). Top-Aufgabe.* München: Hanser.

Lienert, *G.* A. (1989). *Testaufbau und Testanalyse.* Weinheim: Psychologie Verlaus Union.

Locke, E.A. (1986). Generalizing from laboratory to field: Ecological validity or abstraction of essential elements? In E.A. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 3-9). Lexington: Heath.

Mobley, W.H. (1982). *Employee turnover: Causes, consequences, and control.* Reading, MA: Addison-Wesley.

Moser, K. (1987). Artefaktforschung: Aspekte einer methodologischen Systematisierung. *Gruppendynamik, 18,* 83-98.

Moser, K. *(1991). Konsistenz der Person.* Göttingen: Hogrefe.

Moser. K. & Schuler, H. (1989). The nature of psychological measurement. In P. Herriot (Ed.), *Handbook of assessment in organizations* (pp. 281-305). New York: Wiley.

Nachreiner, F., Müller, G.F. & Ernst, G. (1987). Methoden zur Planung und Bewertung arbeitspsychologischer Interventionsmaßnahmen. In U. Kleinbeck & J. Rutenfranz (Hrsg.), *Enzyklopädie der Psychologie D/III/1* (S. 360-439). Göttingen: Hogrefe.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775-777.

Rosenthal, R. & Rubin, D.B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences, 3,* 377-415.

Schnelle, E. (Hrsg.). (1982). *Metaplan Gesprächstechnik.* Quickborn: Metaplan.

Schuler, H. *(1972). Das Bild vom Mitarbeiter.* München: Goldmann.

Schuler, H. *(1980). Ethische Probleme psychologischer Forschung.* Göttingen: Hogrefe.

Schuler, H. & Guldin, A. (1991). Methodological issues in personnel selection research. In C.L. Cooper & I.T. Robertson (Eds.), *International review of industrial and organizational psychology 1991.* (Vol. 6, pp. 213-264). New York: Wiley.

Schwab, D.P. (1980). Construct validity in organizational behavior. In B.M. Staw & L.L. Cummings (Eds.), *Research in organizational behavior* (Vol. 2, pp. 343). Greenwich, CO: JAI Press.

Staw, B.M. (1984). Organizational behavior: A review and reformulation of the field's outcome variables. *Annual Review of Psychology, 35,* 627-666.

Stelzl, I. (1984). Experiment. In E. Roth (Hrsg.), *Sozialwissenschaftliche Methoden* (S. 220-237). München: Oldenbourg.

Stone, E.F. *(1978). Research methods in organizational behavior.* Glenview, IL: Scott.

Stone, E.F., Stone, D.L. & Gueutal, H.G. (1990). Influence of cognitive ability on responses to questionnaire measures: Measurement precision and missing response problems. *Journal of Applied Psychology, 75,* 418-427.

Wanous, J.P. (1989). Installing a realistic job preview: Ten tough choices. *Personnel Psychology, 42,* 117-134.

Webb, E.J., Campbell, D.T., Schwartz, R.D. & Sechrest, L. (1975). *Nichtreaktive Meßverfahren.* Weinheim: Beltz.

Weick. K.E. (1967). Organizations in the laboratory. In V. Vroom (Ed.), *Methods of organizational research* (pp. 1-56). Pittsburgh: University of Pittsburgh Press.

Wortman, PM. (1983) Evaluation-research: A methodological perspective. *Annual Review of Psychology, 34,* 223-260.

Wottawa, H. & Thierau, H. *(1990). Evaluation.* Bern: Huber.

ZUMA (Hrsg). (1983). *Skalen-Handbücher.* Mannheim: ZUMA.