In: Kühl; Strodtholz; Taffertshofer, Handbuch Methoden der Organisationsforschung. Quantitative und Qualitative Methoden, Wiesbaden 2009: VS Verlag für Sozialwissenschaften, S.534-557.

Experiment

Stefan Kühl

1 Einleitung

Das Experiment wird häufig nicht als Methode der Datenerhebung, sondern als eine spezifische Form des Untersuchungsdesigns verstanden (vgl. z. B. Diekmann 1999, S. 8 ff.). In einem Experiment variiert der Forscher einzelne Bedingungsfaktoren (unabhängige Variablen), um zu sehen, welche Effekte (abhängige Variablen) sich daraus ergeben. Die Veränderung der abhängigen Variablen können dann mit Instrumenten der Datenerhebung, wie mündliche oder schriftliche Befragung, Beobachtung oder Inhaltsanalyse usw., gemessen werden. Wenn mögliche weitere Einflussgrößen (Störvariablen) kontrolliert werden können, müssten bei einer Wiederholung des Experiments die gleichen Effekte auftreten (vgl. z. B. Bortz 1984, S. 35 ff.; Osnabrügge/Frey 1989, S. 180; Czienskowski 1996, S. 23).

Tabelle 1: Unabhängige, abhängige und Störvariablen im Experiment

Unabhängige Variable (UV): Die unabhängige Variable wird vom Experimentleiter absichtsvoll und geplant variiert, um eine Reaktion der abhängigen Variable zu bewirken.
Abhängige Variable (AV): Die Reaktion der abhängigen Variable auf das geplante Variieren der unabhängigen Variablen wird beobachtet. Der Effekt wird in der Hypothese vorausgesagt.
Störvariable (SV): Als Störvariable wird eine Variable bezeichnet, die den Einfluss der unabhängigen auf die abhängige Variable verfälscht. Das Ziel der Experimentalanordnung ist es, mögliche Einflusseffekte der Störvariablen vollständig zu kontrollieren bzw. zu neutralisieren.

Es gibt keine Möglichkeit alle möglichen Störvariablen zu bestimmen und im Wirkzusammenhang zu berücksichtigen. Die einzige Möglichkeit, potenzielle Einflüsse bei einer entsprechenden Stichprobengröße im Durchschnitt zu eliminieren bzw. zu neutralisieren, bietet das Experiment mit einer Zufallsaufteilung der Untersuchungseinheiten (Randomisierung) auf eine Kontroll- und eine Experimentalgruppe. Aus diesem Grund gilt im kausal-wissenschaftlichen Paradigma das Experiment als Königsweg bei der Suche nach Kausalitäten. Bei einer nicht mit einem Experiment kombinierten Befragung, Beobachtung oder Inhaltsanalyse kann man zwar Zusammenhänge zwischen Variablen herausarbeiten, die ursächliche Wirkrichtung der Variablen kann dagegen nicht angegeben werden. Eine Befragung von Führungskräften in Unternehmen kann beispielsweise einen Zusammenhang zwischen Vermögen der Führungskraft und Unternehmensgröße liefern. Die Wirkrichtung zwischen beiden Variablen ist zunächst unklar. Möglicherweise ist der

höhere Verdienst in großen Firmen Ursache für große Vermögen. Ebenso ist aber denkbar, dass Vermögen die Voraussetzung für die Finanzierung hochqualifizierender Ausbildung darstellt, um dann letztlich die begehrteren Stellen in Großunternehmen zu erhalten. Erst durch die Experimentalanordung, die zum Teil auch ex post rechnerisch erzeugt werden kann, lassen sich unabhängige und abhängige Variablen unterscheiden, Störvariablen bestimmen und Kausalität beschreiben.

Labor- und Quasi-Experiment, Planspiel, Feld- und Krisenexperiment

Diese Bestimmung eines Experiments geht vom Idealfall des Laborexperiments aus, das sich durch Kontrolle der Störvariablen mit Hilfe der Randomisierung und durch kontrolliertes Variieren der unabhängigen Variable auszeichnet. Das Laborexperiment biete, so die herkömmliche Auffassung, ein Maximum an gesichertem Erkenntnisgewinn (vgl. Mertens 1975, S. 19 f.). Die von Karl Popper formulierten Gütekriterien der Wissenschaftlichkeit, Reproduzierbarkeit, Standardisierbarkeit und Messbarkeit Aufgrund der erfüllbaren Ansprüche an finden sich in fast in optimaler Ausprägung wieder.

Dabei darf aber nicht übersehen werden, dass unter dem Begriff des Experiments neben dem Laborexperiment auch noch andere Formen fallen können, die den Ansprüchen an Reproduzierbarkeit, Standardisierbarkeit und Messbarkeit nicht in gleicher Form gerecht werden: das Planspiel, das Quasi-Experiment, das Feldexperiment und das Krisenexperiment.

Planspiel: Planspiele ähneln auf dem ersten Blick Laborexperimenten. Bedingungen (unabhängige Variable) werden planmäßig manipuliert und die Effekte (abhängigen Variablen) beobachtet. Ähnlich wie beim Laborexperiment wird auch beim Planspiel den Teilnehmern eine soziale Situation aufoktroyiert. Aber beim Planspiel geht es nicht wie beim Experiment, vorrangig um die Überprüfung von Hypothesen über Kausalzusammenhänge, sondern um eine möglichst plausible Simulation sozialer Realitäten. So ist auch die Komplexität der im Planspiel simulierten Realität in der Regel wesentlich höher als beim Laborexperiment und wird weniger durch Ansprüche an Standardisierung und Messbarkeit beeinflusst (vgl. Berg 1988, S. 150; siehe auch den Beitrag zum Planspiel in diesem Band).

Quasi-Experiment: Das Quasi-Experiment ähnelt dem Laborexperiment insofern, als auch hier die unabhängige Variable vom Experimentleiter aktiv manipuliert wird. Anders als beim Laborexperiment können beim Quasi-Experiment die Testpersonen nicht nach Zufallsauswahl der jeweiligen Experimentalgruppen zugeordnet werden (vgl. Campbell/Stanley 1963). Bei den Hawthorne-Experimenten beispielsweise, eine Experimentreihe, die ab Ende der zwanziger Jahre bei der Western Electric Company in Hawthorne durchgeführt wurde, konnten die Versuchsleiter die Bedingungen in den verschiedenen Arbeitsgruppen verändern. Sie hatten aber nur sehr begrenzte Möglichkeiten alle Störvariablen zu kontrollieren. Sie konnten zum Beispiel nicht sicherstellen, dass die Räumlichkeiten, in denen die Gruppen arbeiteten absolut identisch waren. Auch konnten sie die Arbeiterinnen nicht per Los einer beliebigen Arbeitsgruppe zuordnen, sondern mussten sich an die im Betrieb vorgenommenen Gruppeneinteilungen orientieren (vgl. Roethlisberger/Dickson 1939).

Feldexperiment: Bei Feldexperimenten wird die Testperson nicht wie beim Laborexperiment in eine künstliche Umgebung eines Labors gebracht, sondern die Unter-

suchung wird in einer für die Testpersonen natürlichen Umgebung durchgeführt (vgl. Koch 1976; Bungard/Bay 1982). Muzafer Sherif, der die Methode des Feldexperiments maßgeblich entwickelt hat, führte seine Untersuchungen zu Inter- und Intragruppenkonflikten in einem Jugendlager durch, das für die Testpersonen eine weitgehend natürliche Umgebung darstellte. Wie beim Laborexperiment werden auch beim Feldexperiment die Bedingungen (unabhängigen Variablen) durch die Forscher manipuliert. Es wird dann beispielsweise untersucht, in wiefern eine gemeinsame, übergreifende Aufgabe die Konflikte zwischen zwei gebildeten Jugendgruppen reduziert (vgl. Sherif 1954, 1958).

Krisenexperiment: Bei Krisenexperimenten, in der Regel einer Spielart des Feldexperiments, initiiert der Versuchsleiter für die Testpersonen eine Krisensituation. Das einfachste Beispiel eines Krisenexperiments ist die von Harold Garfinkel (1973, S. 207) entwickelte Reaktion auf die Frage „Wie geht‘s?“. Antwortet man nicht mit einem „Mir geht‘s gut“, sondern fragt nach „Wie geht es mit was? Meiner Gesundheit, meinen Geldangelegenheiten, meinen Aufgaben für die Hochschule, meinen Seelenfrieden“ löscht man sich von Konventionen der Alltagsinteraktionen und löst beim Gesprächspartner eine Krise aus. Die Reaktionen auf die Krise können dann untersucht werden. Aufgrund der Schwierigkeiten bei der Kontrolle von Störvariablen werden Krisenexperimente nicht zum klassischen Repertoire sozialwissenschaftlicher Experimentalforschung gezählt und vorrangig in der qualitativen Sozialforschung eingesetzt (vgl. Gstettner 1984, S. 440 ff.; Cordes 1994, S. 155 ff.; für die Nähe zum politischen Straßentheater siehe Boal 1979).

Die hier vorgestellte Unterscheidung in Laborexperiment, Planspiel, Quasi-Experiment, Feldexperiment und Krisenexperiment ist idealtypisch. Die verschiedenen Formen des Experiments sind teilweise miteinander kombinierbar. So gibt es Laborexperimente, die mit einer simulierten Krise arbeiten. Es existieren Planspiele, die in der „natürlichen“ Arbeitsumgebung von Organisationsmitgliedern durchgeführt werden und so Feldexperimenten ähneln.

Die Herausforderungen einer experimentellen Organisationsforschung

Führt man mit standardisierten Fragebögen eine Umfrage in einem Unternehmen durch, kann man davon ausgehen, dass man eine Organisation untersucht (siehe den Beitrag zur schriftlichen Befragung in diesem Band). Ebenso kann man bei einer standardisierten Beobachtung von Konferenzen in einer Verwaltung sicher sein, Interaktionen in Organisationen zu beforschen (siehe den Beitrag zu SYMLOG und zur standardisierten Befragung).

Die Gewissheit über den gerade untersuchten Typus eines sozialen Systems hat man bei einem Laborexperiment – aber auch bei einem Planspiel – nicht. Bei einem Experiment handelt es sich zunächst „nur“ um eine Face-to-face-Interaktion zwischen dem Experimentleiter und einer oder mehrerer Testpersonen. Ob dabei spontane Face-to-face-Interaktionen, Interaktionen in Gruppen oder Interaktionen in Organisationen abgebildet werden, lässt sich erst durch eine nähere Betrachtung des Experimentaufbaus und der Reaktionen der Testpersonen erschließen (vgl. auch Zelditch/Hopkins 1961).

Während bei Feldexperimenten und bei Krisenexperimenten der Forscher an den schon vorstrukturierten Organisationsmerkmalen „parasitieren“ kann, müssen im Laborexperiment über Organisationen die Merkmale von Organisationen künstlich erzeugt werden.

Was sind nun die Merkmale von Organisationen, die in Experimenten über Organisationen auf alle Fälle simuliert werden müssen? Karl Weick hat in einem nach wie vor grundlegenden Artikel über Laborexperimente in Organisationen eine Liste von Organisationsmerkmalen aufgeführt, die durch das Experiment simuliert werden müssen: Hierarchische Strukturierung der sozialen Situation, Bildung von Untergruppen, Vernetzung zwischen den Aufgaben, die von den Untergruppen erledigt werden, nicht nur Face-to-face-Interaktionen, sondern auch „vermittelte" Interaktionen, Orientierung der Testpersonen an Mitgliedschaften und Karriere, Motivation der Testpersonen durch Anweisungen, Bezahlung, Rückmeldung ihrer Leistungen und durch Wettbewerb (vgl. Weick 1965).

Die Liste von Karl Weick lässt sich auf drei zentrale Merkmale von Organisationen zusammenfassen: Zwecke, Mitgliedschaften und Hierarchien. Besonders durch die entscheidungstheoretische und die systemtheoretische Organisationsforschung wurde aufgezeigt, dass beim Übergang von der ständischen zur modernen Gesellschaft die Bedeutung von Zwecken, Mitgliedschaften und Hierarchien zwar zur Strukturierung der Gesamtgesellschaft abnimmt, diese Elemente aber als zentrale Strukturierungsmerkmale von Organisationen einen zunehmend prominenten Platz einnehmen (vgl. in Anlehnung an Luhmann hierzu besonders Kieserling 1994; siehe ausführlich die Argumentation in Strodtholz/Kühl 2002, S. 11; Kühl 2003a, S. 251).

Demnach verzichten Gesellschaften seit dem Übergang von der stratifizierten zu einer funktional differenzierten Gesellschaft darauf, sich übergeordneten *Zwecken*, etwa der Befolgung göttlicher Gebote, zu verschreiben. Ganz anders Organisationen: Egal, ob es sich um eine Verwaltung, ein Unternehmen oder eine Kirche handelt, konkrete Zwecke, wie eine mehr oder minder freundliche Befriedigung von Anfragen nach Aufenthaltsgenehmigungen oder die Eroberung des Markts, spielen eine zentrale Rolle in der Ausrichtung von Organisationen (vgl. Luhmann 1973, S. 87 ff., 1997, S. 826 ff.).

Auch das Management des Ein- und Austritts von Personal – die Bestimmung von *Mitgliedschaften* – handhaben Organisationen anders als moderne Gesellschaften. Ein totaler Ausschluss aus der Gesellschaft findet seit der weitgehenden Abschaffung von Verbannung, Ausbürgerung und Todesstrafe nur noch in Ausnahmefällen statt. Das Management der Mitgliedschaft ist dagegen ein zentrales Merkmal von Organisationen geworden. Über die Mitgliedschaft wird trennscharf festgelegt, wer zu einer Organisation gehört und wer nicht. Dadurch werden Grenzen geschaffen, in denen sich die Mitglieder (und eben nur die Mitglieder) den Regeln der Organisation zu unterwerfen haben (Luhmann 1964, S. 16).

Schließlich verlieren auch *Hierarchien* in der Gesellschaft an Bedeutung, während sie für die Strukturierung von Organisationen zentral bleiben. Es gibt in modernen Gesellschaften keine Personen mehr, die über Befehls- und Anweisungsketten in die verschiedenen Lebensbereiche der Bevölkerung hineinregieren könnten. Eine solche Gesellschaft gilt heutzutage als diktatorisch und unmodern. Im Gegensatz zu modernen Gesellschaften sind Organisationen zentral über Hierarchien strukturiert. Erst die Hierarchie stellt sicher, dass die Anweisungen und Zusagen der Spitze auch umgesetzt werden. Sie gewährleistet somit, dass Verbände, Verwaltungen und Unternehmen überhaupt als berechenbare kollektive Akteure auftreten können (Luhmann 1997, S. 834).

Da die Laborsituation von ihrer Struktur her die zeitlich stark befristete Face-to-face-Interaktion zwischen Unbekannten abbildet, ist die Simulation von Organisationsprozessen durch Laborexperimente nicht ganz einfach. Es muss durch den Experimentaufbau deutlich

werden, dass die anwesenden Personen auf Zwecke ausgerichtet sind und in einem legitimierten hierarchischen Verhältnis zu einander stehen. An die Zweckerfüllung und die Akzeptanz der Hierarchie müssen sie durch eine simulierte Einwilligung zur Mitgliedschaft in der „Kurzzeit-Organisation" des Experiments gebunden werden.

2 Datenerhebung und Datenaufbereitung

Im Gegensatz zur so genannten nichtexperimentellen Forschung (Befragung oder Beobachtungen) wird bei Experimenten die soziale Situation durch das Experiment gezielt beeinflusst (Feldexperiment) oder durch den Experimentleiter überhaupt erst geschaffen (Laborexperiment). Deswegen ist der Datenerhebung bei Experimenten immer ein Versuchsdesign vorangestellt.

Im Einzelnen lassen sich bei der Experimentplanung und der Datenerhebung drei Phasen unterscheiden: erstens die Operationalisierung, also die Übersetzung einer sprachlich formulierten Sachhypothese in eine mit mathematischen Mitteln auswertbare, statistische Hypothese; zweitens die Versuchsplanung, die unter anderem darin besteht, verschiedene Versuchs- und Kontrollgruppen zu bilden; drittens die Kontrolle der Störvariablen, also all der Effekte, die den Kausalzusammenhang zwischen den vom Experimentleiter manipulierten Bedingungen und den zu beobachtenden Effekten verzerren könnten (eine sehr gute Darstellung zu den Schritten des Experiments findet sich bei Mittenecker 1964; Huber 2000).

Operationalisierung: Von der Sachhypothese zur statistischen Hypothese

Die Schwierigkeit bei einem Experiment ist – ähnlich wie bei anderen quantitativen Methoden auch – die in Sprache formulierte Sachhypothese in eine mit mathematischen Mitteln zu prüfende, statistische Hypothese zu übersetzen (vgl. Henning/Muthig 1979, S. 18 ff.). Eine Sachhypothese ist eine sprachlich formulierte Aussage. Ein Beispiel für eine solche Sachhypothese wäre „Je genau einer Mitarbeiter überwacht wird, desto bessere Leistungen erbringt er". Zur Übersetzung in eine statistische Hypothese muss ein messbares Kriterium festlegt werden, um die gemessenen Effekte mit statistischen Berechnungen analysieren zu können. So kann man zum Beispiel festlegen, dass sich der Grad der Überwachung durch die räumliche Nähe des Vorgesetzten zum Arbeiter messen lässt. Die Anzahl der Karten, die ein Arbeiter innerhalb einer halben Stunde sortiert, dient als messbares Kriterium der Leistungsfähigkeit. Da sowohl die räumliche Nähe als auch die Anzahl der Karten gemessen werden, kann dann die Sachhypothese in eine statistisch prüfbare Hypothese übersetzt werden: „Je näher der Vorgesetzte beim Arbeiter sitzt, desto mehr Karten sortiert dieser innerhalb einer halben Stunde."

Die Übersetzung von Sachhypothesen in statistische Hypothesen ist alles andere als einfach, kann man doch den generellen Verdacht hegen, dass das Verhalten in Experimenten wenig über das Verhalten außerhalb des Laboratoriums besagt (vgl. Greenwood 1989, S. 177 ff.). Es muss sehr genau geprüft werden, ob durch die Operationalisierung auch die Kategorien der Sachhypothese getroffen werden. Bildet die räumliche Nähe zwischen Vorgesetzten und Arbeiter den Grad der Überwachung ab? Gibt es nicht andere

Kriterien, die besser den Grad der Überwachung abbilden können? Ist das Sortieren von Karten ein adäquater Gradmesser für Leistungsfähigkeit eines Arbeiters? Reicht eine halbe Stunde aus, um die Leistungsfähigkeit eines Arbeiters zu messen?

Zur Messung steht dem experimentell arbeitenden Forscher das breite Spektrum der quantitativen Sozialforschung zur Verfügung. Er kann das Verhalten der Testperson beobachten und dabei beispielsweise die Methoden der strukturierten Beobachtung oder Beobachtung mit SYMLOG anwenden (siehe die Beiträge zur strukturierten Beobachtung und zu SYMLOG in diesem Band). Er kann den Testpersonen anbieten nach dem Experiment einen standardisierten Fragebogen auszufüllen und dabei die verbreiteten Befragungstechniken anwenden (siehe den Beitrag zur schriftlichen Befragung in diesem Band). Oder er kann die Testperson bitten einen Aufsatz zu verfassen und dann den so produzierten Text nach dem Vorkommen bestimmter Worte untersuchen.

Versuchsplan: Die Bildung von Versuchs- und Kontrollgruppen

Um herauszubekommen, ob die vom Experimentleiter manipulierte Bedingung (unabhängige Variable) für bestimmte Effekte (abhängige Variable) verantwortlich ist, muss er prüfen, was passiert, wenn die Bedingung nicht manipuliert werden. Dies kann er im Prinzip durch zwei – auch kombinierbare – Strategien erreichen (vgl. Hagmüller 1979, S. 165 ff.; Bortz 1984, S. 400 ff.).

Die erste Strategie ist, dass er die Gruppe der Versuchspersonen zu einem Zeitpunkt mit einer unmanipulierten Bedingung konfrontiert, die Effekte misst und dann zu einem anderen Zeitpunkt die gleiche Gruppe der Versuchspersonen der manipulierten Bedingung aussetzt und dann wiederum die Effekte bestimmt. Diese Vorgehensweise ist jedoch problematisch, weil die erste Phase des Experiments (bei unmanipulierter Bedingung) die Ergebnisse in der zweiten Phase (bei manipulierter Bedingung) beeinflussen kann. Diese kann an dem Experiment von Richard Tracy Lapiere (1934) über Diskrepanz zwischen Einstellung und Verhalten gegenüber ethnischen Minderheiten verdeutlicht werden. Lapiere reiste in den frühen dreißiger Jahren mit einem jungen chinesischen Paar durch die Vereinigten Staaten, übernachtete mit ihnen in vielen Hotels und aß mit ihnen in einer Vielzahl von Restaurants. Während der ganzen Zeit wurde ihnen in weniger als 1 % aus fremdenfeindlichen Gründen die Bedienung verweigert. Nach seiner Reise wandte er sich mit einem Fragebogen an die 250 Inhaber der Restaurants und Unterkünfte an, die er mit dem chinesischen Pärchen besucht hatte. Über 90 % der Hotel- und Restaurantbesitzer gaben bei der Beantwortung des Fragebogens an, dass sie Chinesen keine Unterkunft oder Verpflegung gewährten. Das Problem war, dass Lapiere nicht ausschließen konnte, dass das ablehnende Verhalten erst durch den Kontakt mit dem jungen chinesischen Paar – also durch seine Experimentbedingungen – ausgelöst wurde.

In der zweiten Strategie soll diesem Zweifel durch die Bildung einer zweiten Gruppe begegnet werden, die der manipulierten Bedingung nicht ausgesetzt ist. Bei der Gruppe, die den Manipulationen des Experimentleiters unterzogen ist, spricht man von der Versuchs- oder Experimentalgruppe, bei der Gruppe, die die gleiche Beobachtung ohne die Manipulationen des Experimentleiters erfährt, spricht man von der Kontrollgruppe. Lapiere schickte beispielsweise auch an eine Kontrollgruppe von hundert Hotels und Restaurants, die er nicht besucht hatte, den gleichen Fragenbogen, die er auch an die besuchten Hotels

versandt hatte. Da auch hier der weitgehende Teil der Besitzer, die Aufnahme von Chinesen ablehnte, konnte er davon ausgehen, dass die Einstellung seiner Versuchsgruppe nicht durch die vorangegangene praktische Erfahrung mit ihm und seinen chinesischen Gästen verzerrt worden war.

Wenn die Versuchspersonen nicht wissen, ob sie der Versuchsgruppe oder der Kontrollgruppe zugeordnet sind, spricht man von einem Blindversuch. Wenn auch der Forscher nicht weiß, ob eine Person zur Versuchs- oder zur Kontrollgruppe gehört, dann spricht man von einem Doppelblindversuch. Besonders bei medizinischen Forschungen wird sichergestellt, dass nicht nur die Versuchsperson in Unkenntnis darüber ist, ob sie ein Placebo oder ein Medikament nimmt. Auch der die Wirkung testende Arzte weiß nicht, ob die Versuchsperson zur das Medikament nehmenden Versuchsgruppe oder zur das Placebo einnehmenden Kontrollgruppe gehört. Durch einen Blindversuch soll verhindert werden, dass die Versuchsperson durch Selbstsuggestion oder der Versuchsleiter durch unbewusste Beeinflussungsmechanismen die Ergebnisse des Experiments verzerren.

Häufig haben wir es in Experimenten nicht mit Kontroll- und Vergleichsgruppen im engeren Sinne zu tun. Wenn man beispielsweise in einem Unternehmen, die Auswirkungen von Rationalisierungsmaßnahmen auf die Arbeitsproduktivität untersucht, könnte man mit einer Versuchsgruppe und einer Kontrollgruppe arbeiten. In einer Versuchsgruppe wird beispielsweise ein kontinuierlicher Verbesserungsprozess eingeführt und danach die Produktivität gemessen. Diese Produktivität wird dann mit einer Kontrollgruppe verglichen, in der der kontinuierliche Verbesserungsprozess nicht eingeführt wurde. Wenn wir jedoch in einer Gruppe den kontinuierlichen Verbesserungsprozess durchführen und in einer anderen Gruppe die Vorarbeiterposition auflösen, dann vergleichen wir im engeren Sinne nicht eine Versuchsgruppe mit einer Kontrollgruppe. Durch den Vergleich der Produktivität der beiden Gruppen ist jede Gruppe gleichermaßen Kontroll- und Versuchsgruppe (vgl. Diekmann 1999, S. 297).

Kontrolle der Störvariablen: Parallelisieren und Randomisierung

In jedem Experiment kann es vorkommen, dass neben den vom Experimentleiter gezielt beeinflussten Variablen auch noch andere Variablen Einfluss auf die gemessenen Ergebnisse hatte. Diese Störvariablen stellen das ganze Experiment in Frage, weil der Forscher jetzt nicht mehr bestimmen kann, ob die Effekte durch die von ihm geplant manipulierten unabhängigen Variablen oder durch die Störvariablen ausgelöst wurden.

Eine klassische Störvariable ist der *Wissenschafts-Effekt (auch Hawthorne-Effekt)*. Damit werden verzerrende Einflüsse bezeichnet, die durch den wissenschaftlichen Kontext des Experiments entstehen. In den Hawthorne-Werken der Western Electric Company führte in den zwanziger Jahren eine Forschungsgruppe um den Sozialpsychologen Elton Mayo Untersuchungen zur Leistungssteigerung durch. Ausgangspunkt war eine Untersuchung der Firma, ob eine bessere Beleuchtung die Arbeitsproduktivität in der Montage erhöht. Wie erwartet stieg die Produktivität mit gesteigerter Beleuchtung an. Paradoxerweise stieg die Produktivität aber auch an, als das Management die Beleuchtung gleich ließ oder sie reduzierte. In einer Vielzahl von Experimenten wurde von der Forschergruppe die Erklärung herausgearbeitet, dass sich die Produktivität deshalb verbesserte, weil die Testpersonen in den Mittelpunkt wissenschaftlicher Aufmerksamkeit

geraten waren und sie sich dadurch eine größere Mühe gaben (vgl. Roethlisberger/Dickson 1939; kritisch Bramel/Friend 1981; Moldaschl/Weber 1998).

Ein andere typische Störvariable ist der *Verlierer-Effekt.* Damit werden Verzerrungen benannt, die durch die experimentelle Zuweisung einer Personengruppe auf eine schlechter angesehene Position erzeugt werden. Eine Versicherung in Mannheim führte in den neunziger Jahren ein Assessment-Center durch, um eine Gruppe von neuen Versicherungsvertretern auszuwählen. Die im Assessment-Center am besten bestehenden Personen wurden von der Versicherung eingestellt. Ein plötzlicher Nachfrageboom führte dazu, dass auch die ursprünglich nicht qualifizierten Versuchspersonen von der Versicherung eingestellt wurden. Der Vergleich der beiden Gruppen in Bezug auf die Verkaufszahlen ergab, dass die ursprünglich nicht ausgewählte Gruppe tendenziell bessere Verkaufsergebnisse brachte als die durch das Assessment-Center bestimmten. Diese (leider nicht als Artikel publizierten) Ergebnisse können mit dem Versagen von Assessment-Center erklärt werden (vgl. Kühl 2003b); es ist aber auch vorstellbar, dass die Zurechnung der Vertreter zur „Verlierer"-Gruppe deren Leistungsbereitschaft besonders angespornt hat.

Eine weitere häufig vorkommende Störvariable ist der *Selbstselektions-Effekt.* Damit werden die Verzerrungen in einem Experiment benannt, die durch die Selbstselektion der Testpersonen für ein Experiment oder gar für eine bestimmte Gruppe entstehen können. Eine Forschungsgruppe um den Sozialpsychologen David Seidman untersuchte in den fünfziger Jahren, ob Menschen in Gruppen oder alleine besser Elektroschocks ertragen können. Für dieses Experiment wurden als Versuchspersonen knapp über hundert Wehrpflichtige gewonnen, die gerade ihren Grundwehrdienst abgeschlossen hatten. Die Versuchsanordnung sah vor, dass die Versuchspersonen die Höhe der Elektroschocks selbst mit Hilfe eines Einstellknopfes festlegen konnten. Das Ergebnis war, dass die Versuchspersonen bereit waren, sich höhere Elektroschocks zu setzen, wenn ein anderer Soldat gleichzeitig sich Elektroschocks verabreichte als wenn die Testpersonen allein im Raum waren. Die Frage ist jedoch, ob nicht die freiwillige Meldung für ein schmerzhaftes Experiment im Rahmen einer militärischen Ausbildungsinstitution, nicht die Ergebnisse so weit verzerren, dass keine allgemeine Rückschlüsse gezogen werden können (vgl. Seidman et al. 1957; siehe auch Mann 1999, S. 117).

In einem Experiment müssen sowohl die durch die wissenschaftliche Untersuchungssituation erzeugten (vgl. früh Kintz et al. 1965) als auch der die Testperson bedingten Störvariablen (vgl. früh Schultz 1969) kontrolliert werden. Besonders der Kontrolle der durch die Testpersonen bedingten Störvariablen, muss hohe Aufmerksamkeit gewidmet werden. Es muss sichergestellt werden, dass durch die Zuteilung der Testpersonen auf die Versuchsgruppe und die Kontrollgruppe keine Verzerrungen entstehen. Dafür lassen sich die beiden Standardmethoden Parallelisieren und Randomisierung unterscheiden (vgl. auch Heller/Rosemann 1974, S. 71 ff.; Czienskowski 1996, S. 62; Huber 2000, S. 93 ff.).

Beim Parallelisieren werden durch dem Experiment vorgeschaltete Tests sichergestellt, dass die Versuchs- und die Kontrollgruppen sich nicht in für das Experiment zentralen Experimenten unterscheiden. Will man die Auswirkung der räumlichen Nähe eines Vorgesetzten auf die Schnelligkeit bei der Sortierung von Karten messen, sollte man sicherstellen, dass die Fingerfertigkeiten sich in der verschiedenen Gruppe nicht allzu stark unterscheiden. Dafür kann man die Versuchspersonen vor dem eigentlichen Experiment eine ähnliche Fertigkeiten erfordernde Übung machen lassen und dann darauf achten, dass

im Durchschnitt die Testpersonen in die Kontroll- und die Versuchsgruppe sich in ihren Grundfähigkeiten nicht unterscheiden.

Bei der Randomisierung werden die Versuchspersonen zufällig auf Versuchs- und Kontrollgruppe (oder den unterschiedlichen Versuchsgruppen) verteilt. Die Zufallsgenerierung sollte nicht durch Ad-hoc-Zuteilungen des Experimentleiters erfolgen, weil sich unbewusst Selektionskriterien des Experimentleiters einschleichen könnten. Die zufällige Zuteilung sollte vielmehr durch Auszählen, durch Münzwurf oder durch Auslosen vorgenommen werden. Der Vorteil der Randomisierung ist, dass anders als beim Parallelisieren die Störvariable nicht im Einzelnen bekannt sein muss. Man geht davon aus, dass durch die zufällige Zuteilung sich die Testpersonen der Versuchs- und die Kontrollgruppe in allen relevanten Aspekten ähneln und die zu messenden Effekte allein durch die Manipulationen des Versuchsleiters entstehen.

3 Datenanalyse und Dateninterpretation

Nach der Durchführung des Experiments hat der Forscher die Rohdaten seines Experiments zur Verfügung. Die Analyse seiner Daten verläuft in drei Schritten: erstens der statistischen Auswertung der Daten; zweiten der Bestimmung des Zusammenhangs von statistischer Hypothese und Sachhypothese; drittens der Bestimmung der Reichweite des Experiments.

Statistische Auswertung des Experiments

Der erste Schritt besteht in der statistischen Prüfung der Hypothesen. Die statistischen Auswertungsverfahren unterscheiden sich nicht von den Prüfverfahren, die bei einer Befragung oder bei einer quantifizierenden Beobachtung eingesetzt werden können. Wie bei Befragungen und Beobachtungen muss geprüft werden, ob die Anzahl der Stichproben ausreichend gewesen ist, um eine statistische Validität zu erreichen. Wie bei anderen quantitativen Methoden muss auch bei Experimenten Signifikanztests durchgeführt werden. Wie bei anderen quantitativen Untersuchungen bietet es sich auch bei Experimenten an, über multivariate Analysen die Zusammenhänge zwischen drei oder mehr Variablen zu prüfen (einen guten Überblick vermittelt Czienskowski 1996, S. 91 ff.).

Wichtig ist zu unterscheiden, ob durch das Experiment eine Vielzahl von Fällen generiert wird, die dann mit statistischen Methoden überprüft werden, oder ob das Experiment aus einem einzigen Fall besteht, in dem lediglich das Verhalten der Teilnehmer quantitativ gemessen und dann statistisch ausgewertet wird. Das Stanford-Prison-Experiment, das häufig ganz selbstverständlich im Kontext von quantitativen Experimenten aufgeführt wird (vgl. z. B. Bierbrauer 1997), gehört zum zweiten Fall. In diesem Experiment teilte der Experimentleiter eine Gruppe von „normalen“ Männern nach dem Zufallsprinzip in eine Gruppe von Gefängniswärtern und eine Gruppe von Gefangenen auf. In einem fiktiven Gefängnis in der Universität von Stanford sollten die beiden Gruppen für einige Tage die Rollen von Gefängniswärtern und Gefangenen spielen. Das für zwei Wochen geplante Experiment wurde von den Experimentleitern nach sechs Tagen abgebrochen, weil sich bei der Hälfte der Gefangenen starke Anzeichen von Passivität und Depression ausbildeten, während einige Wärtern sadistische Verhaltensweisen entwickelten (vgl.

Haney/Banks/Zimbardo 1973, S. 69 ff.; siehe auch Zimbardo et al. 1973, 1975). In dem Experiment wurden die Einstellungen der Testpersonen quantitativ erhoben und auch die Aggressionen der Personen quantitativ gemessen. Dies darf aber nicht darüber hinwegtäuschen, dass das Experiment nur ein einziges Mal durchgeführt wurde und damit nicht die Minimalanforderungen an eine ausreichende Stichprobenzahl erfüllt.

Zusammenhang von statistischer Hypothese und Sachhypothese (interne Validität)

Der zweite Schritt ist die Überprüfung des Zusammenhangs zwischen der statistischen Hypothese und der Sachhypothese. Häufig wird, so Oswald Huber, in Fachzeitschriften suggeriert, dass die Bestätigung der statistischen Hypothese mit der Bestätigung der Sachhypothese identisch ist. Dabei besagt die Bestätigung der statistischen Hypothese zunächst nichts anderes, als dass auf der Basis richtig gerechneter statistischen Verfahren die Hypothese plausibel erscheint (Huber 2000, S. 132 f.). Aber dem Forscher geht es ja nicht vorrangig um die statistische Hypothese, sondern er ist an der Sachhypothese interessiert.

Wenn in der Phase der Datenanalyse der Zusammenhang zwischen statistischer Hypothese und Sachhypothese geprüft wird, geht es letztlich darum die eigene Operationalisierung noch einmal kritisch zu überprüfen. Wurde durch das Experiment wie geplant der Einfluss von kontinuierlichen Verbesserungsprogrammen auf Produktivität gemessen? Bildet das Experiment wirklich die Anpassung an Gruppendruck ab oder war es den Testpersonen vielleicht völlig egal, wie sie sich selbst in der Experimentalsituation verhalten?

Bestimmung der Reichweite des Experiments (externe Validität)

Der dritte Schritt besteht darin die Reichweite des Experiments zu klären. Der Forscher kreiert im Labor eine eigene soziale Situation. Er legt fest, wie lange ein Experiment dauert, wie viele Personen daran teilnehmen und unter welchen Regeln die Kontakte zwischen den Personen ablaufen. Es ist damit eine offene Frage, ob die Laborsituation einer spontanen Interaktion zwischen Personen, einer Interaktion in stabilen Gruppen von Freunden, einer Interaktion in Familien oder einer Interaktion in Organisationen ähnelt. Erfahrungsgemäß gibt es bei Laborexperimenten zwei potenzielle Fehler bei der Bestimmung der Reichweite eines Experiments.

Der erste Fehler ist die Übergeneralisierung eines Experiments: Bei einer Befragung oder einer Beobachtung wird die Reichweite der Argumentation schon dadurch beschränkt, dass der Forscher sich bewusst ist, in welchem sozialen Kontext er seine Untersuchung durchführt. Wenn ein Sozialforscher die Dynamik in Familien beobachtet, steht er in einer Begründungspflicht, wenn er seine Ergebnisse nicht nur für Familien, sondern auch für Freundschaftsbeziehungen für relevant erklären will. Wenn eine Forscherin in einer Verwaltung eine Befragung zu Über- und Unterordnungsverhältnissen durchführt, müsste sie begründen, wenn sie ihre bestätigten Hypothesen auch für Über- und Unterordnungsverhältnisse in Familien als gültig betrachtet. Da besonders in Laborexperimenten von Sozialpsychologen häufig nicht spezifiziert wird, welche Art von sozialem System untersucht wird, besteht die Gefahr der vorschnellen Generalisierung. Eine Tendenz zur Über-

generalisierung lässt sich beispielsweise beim ursprünglich als Planspiel konzipierten Deportationsexperiment feststellen. Beim Deportationsexperiment handelt es sich um die Simulation einer Deportation von mehreren hunderttausend Gastarbeitern aus dem Osten Deutschlands in ein radioaktiv verseuchtes Gebiet in Süddeutschland. Für diese Massendeportation muss eine Gruppe von Testpersonen nächtliche Transporte durch Deutschland planen, die Bahnwaggons für den Transport einer großen Anzahl von Personen entwickeln und ausstatten, eine möglichst kostengünstige Verpflegung organisieren und die Arbeitsfähigkeit der Personen nach ihrer Ankunft im strahlenverseuchten Gebiet untersuchen. Als Ziel der simulierten Operationen wird den Teilnehmern die offizielle Zweckformulierung eines Bahnunternehmens, also die möglichst effektive Abwicklung von Güter- und Personentransporten, genannt. Dass es sich bei den Transporten um Zwangsdeportationen von Ausländern in ein strahlenverseuchtes Gebiet handelt, kann jeder Stelleninhaber aber aus den mitgelieferten Informationen erschließen (Kraus 2003, S. 3). Der Generalisierungsfehler besteht jetzt darin, dass die hohe Folgebereitschaft in dem Experiment – nur in einem einzigen von über dreihundert Fällen war der Widerstand so stark, dass das Experiment abgebrochen wurde – als Indiz dafür angesehen wird, dass in der modernen Gesellschaft Menschen als ein „Rädchen im Getriebe“ zur Teilnahme an einem Massenmord fähig sind (vgl. Kraus 1987, S. 50 ff.; Berg 1988, S. 121 ff.). Es lässt sich jedoch mit Gründen annehmen, dass durch das Deportationsexperiment „lediglich“ eine Organisation simuliert wird. Die Aussage kann nur dahingehend generalisiert werden, dass Menschen dann zur Beteiligung an einem Massenmord bereit sind, wenn sie in ein System aus Hierarchien, Zweckvorgaben und Regeln eingebunden sind, durch die sie sich als Mitglieder der Organisation gebunden sehen (vgl. Kühl 2005).

Der zweite Fehler kann die Falschzurechnung eines Experiments sein: Während der erste Fehler in der Übergeneralisierung der Ergebnisse besteht, stellt der zweite Fehler die vorschnelle Zurechnung der experimentellen Ergebnisse auf eine bestimmte soziale Situation dar. In jeder sozialwissenschaftlichen Disziplin, die sich Experimenten bedient, gibt es thematische Moden. Einmal ist die Beforschung von Gruppen angesagt, ein andermal von Familien. Einmal ist die Organisationsforschung aktuell, ein andermal die Beforschung spontaner Face-to-face-Interaktionen. Die Gefahr besteht jetzt darin, dass ein Experiment vorschnell dem gerade aktuellen Modethema der Disziplin zugerechnet wird, ohne zu überprüfen, ob nicht durch die Operationalisierung des Experiments eine ganz anderes soziales System abgebildet wird. Ein Beispiel für eine solche Falschzurechnung könnte das Experiment von Solomon Asch angesehen werden. Die in der ersten Hälfte des zwanzigsten Jahrhunderts durchgeführten sozialwissenschaftlichen Experimente verstanden sich als Forschungen zur Funktionsweise von Gruppen. Solomon E. Asch, einer der prominentesten „Gruppenforscher“, zeigte, wie stark Personen sich dem Druck anderer Personen unterordnen. In seinem Experiment wurden sieben Personen aufgefordert, die Länge dreier Linien einzuschätzen. Sechs der sieben Personen waren Mitglieder des Forschungsteams, die – ohne dass es die siebte Person wusste – Strohmänner des Versuchsleiters waren und systematisch falsche Einschätzungen abgaben. Das Ergebnis war, dass unter dem Druck der sechs Personen die eigentliche Testperson den falschen Einschätzungen der anderen Personen folgte (Asch 1951, S. 177 ff., 1955, S. 31 ff.). Aus einer differenzierungstheoretischen Perspektive würde man die Experimente von Asch heutzutage nicht mehr als Experimente zu Gruppenprozessen, sondern zu unmittelbaren Face-to-face-Interaktionen betrachten. Face-to-face-Interaktionen ergeben sich schon alleine auf-

grund von gegenseitigen Wahrnehmungen, während Gruppen darüber hinaus ein Gefühl von „Zusammengehörigkeit“ entwickeln (vgl. Tyrell 1983, S. 83). Da bei Solomon Asch die Testperson mit den sechs Lockvögeln nur eine sehr spontane Beziehung aufgebaut hat, spricht vieles dafür die Aussagen von Aschs Experiment für Interaktionen außerhalb von Gruppen und nicht für Interaktionen in Gruppen gelten zu lassen (vgl. Kieserling 1999, S. 17).

4 Anwendungsbeispiel

Vermutlich die bekannteste sozialwissenschaftliche Experimentreihe ist die von Stanley Milgram in den frühen sechziger Jahren durchgeführte Untersuchung zur Gehorsamsbereitschaft gegenüber Autoritäten. Sein Buch über das Experiment wurde in elf Sprachen übersetzt, in Magazinen wie Harper’s und Esquire wurde über die Experimente berichtet. Es entstanden Fernsehsendungen über Milgrams Versuchsreihe und das Experiment bildete sogar die Grundlage für einen Spielfilm (vgl. Miller 1986, S. 7 ff.; Blass 1992b, S. 293 ff.).

Datenerhebung

Im Grundexperiment erklärt ein mit zentralen Insignien der wissenschaftlichen Autorität ausgestatteter Experimentleiter der Testperson, dass diese im Rahmen eines Experiments zur Lernfähigkeit von Schülern die Rolle eines Lehrers zu übernehmen hätte. Wenn ein im Nebenraum sitzende Schüler eine falsche Antwort gegeben hat, sollte die Testperson dem Schüler Elektroschocks in kontinuierlich zunehmender Stärke verabreichen. Die Testperson wusste nicht, dass der Schüler von einem Mitarbeiter des Forschungsteams gespielt wurde und seine Reaktionen auf die Stromstöße, wie Schmerzenschreie, Proteste und plötzliches Verstummen, lediglich simuliert wurden (vgl. Milgram 1963, S. 372 ff.).

Operationalisierung: Von der Sach- zur statistischen Hypothese

Die Herausforderung für Milgram war es, Gehorsamkeit so zu operationalisieren, dass die Ergebnisse der verschiedenen Experimente miteinander verglichen werden konnten. Dies leistete er darüber, dass die Testpersonen die vermeintlichen Stromstöße über einen Apparat mit insgesamt dreißig Schockstufen versetzen sollten. Die Aufschrift beinhaltete einmal die um jeweils 15 Volt steigende Voltstufen bis hin zu 450 Volt und kurze Erklärungen der Schockstufen. Die Aufschriften lauteten: leichter Schock (15–60 Volt), mäßiger Schock (75–120 Volt), mittlerer Schock (135–180 Volt), kräftiger Schock (195–240 Volt), schwerer Schock (255–300 Volt), sehr schwerer Schock (315–360 Volt), Gefahr! Bedrohlicher Schock (375–420 Volt) und abschließend XXX (435–450 Volt) (vgl. Milgram 1974, S. 44 f.).

Die Ergebnisse der verschiedenen Varianten des Experiments konnten jetzt in zweifacher Form gemessen werden. Die erste Messung bestand darin, die durchschnittlich maximale Schockstufe zu bestimmen, also den Durchschnitt der Spannungsstufen, an dem die Versuchspersonen sich weigerten weitere Stromstöße zu setzen. Ein Wert von 10 be-

deutete dann beispielsweise, dass sich die Versuchspersonen durchschnittlich bei der zehnten Schockstufe (150 Volt) weigerten weitere Stromstöße zu setzen. Mit der zweiten Messung wurde bestimmt, wie viel Prozent der Testpersonen bereit waren den höchsten Stromschlag von 450 Volt zu setzen. Ein Wert von 25 Prozent bedeutete beispielsweise, dass ein Viertel aller Testpersonen bereit waren 450 Volt Stromstöße zu versetzen, während drei Viertel an irgend einer vorigen Stufe sich geweigert hatten weiter zu machen.

Versuchsplanung: Bildung von Versuchs- und Kontrollgruppen

Insgesamt entwickelte Milgram achtzehn Varianten seines Experiments. In einer Reihe von Experimenten holte er das Opfer immer näher an die Testperson heran, um zu messen, ob die räumliche Nähe zum Opfer die Gehorsamsbereitschaft reduziere (Experimente 1–4). In einer weiteren Versuchsreihe testete er, welchen Einfluss die Persönlichkeit und räumliche Nähe des Experimentleiters als Autoritätsperson auf die Gehorsamsbereitschaft hatte (Experimente 6 und 7). In einer am Ende des Projektes durchgeführten Testreihe, untersuchte Milgram dann auch noch, welchen Einfluss rebellierende und zustimmende Peers auf das Verhalten der Testperson hatten (Experimente 16 und 17).

Am ehesten erfüllten zwei Erhebungen die Funktion einer Kontrollgruppe (Milgram 1963, S. 373 ff., 1974, S. 45). Erstens ließ er Gruppen von Psychologen, Studenten und Erwachsenen der Mittelschicht die Beschreibung des Experimentaufbaus lesen und dann schätzen, bei welchem Stromstoß Testpersonen wohl den Versuch abbrechen würden. Seine später bestätigte Vermutung war, dass Personen, die lediglich die Situation geschildert bekommen, die reale Gehorsamsbereitschaft in den Experimenten stark unterschätzen würden.

Zweitens testete er, wie viele Testpersonen bereit sind Stromstöße von 450 Volt zu setzen, wenn es keine akustischen Rückmeldungen des vermeintlich leidenden Schülers gibt. Schon bei seinen Pretest stellte Milgram fest, dass unter diesen Bedingungen fast alle Testpersonen Stromstöße von bis zu 450 Volt setzten, vermutlich weil sie sich das Leiden der Testperson nur schwer vorstellen konnten. Beim Experiment 1 variierte Milgram das Experiment mit der Kontrollgruppe, indem er zwar keine akustischen und visuellen Rückkopplungen initiierte, aber der vermeintliche Schüler bei 300 Volt gegen die Wand hämmerte.

Kontrolle der Störvariablen

Methodisch ist Milgrams Experiment deswegen beachtlich, weil er über die drei Jahre dauernde Erhebungsphase die durch die Versuchspersonen und den Versuchsaufbau bedingten Störvariablen weitgehend kontrollieren konnte. Eine sinnvoll erscheinende Veränderung im Experimentaufbau oder ein notwendig gewordener Wechsel des Versuchsleiters wurde jeweils daraufhin getestet, ob sich die Ergebnisse dadurch veränderten.

Als eine Störvariable konnte angesehen werden, dass die Testpersonen möglicherweise die simulierte Situation mit den gespielten Reaktionen der Schüler durchschauten. Besonders Martin T. Orne und Charles H. Holland (1968) stellten in einer längeren Auseinandersetzung mit Milgram heraus, dass die Testperson die Experimentsituation nicht als

eine Realsituation, sondern als eine „Als-ob-Situation“ wahrnahmen. Ihre Annahme sei es gewesen, dass in einer Experimentsituation schon niemand zu Schaden kommen würde.

Diese Störvariable wurde in den Experimenten von Milgram und in den an Milgram angelehnten Experimenten auf dreifache Weise kontrolliert. Erstens erhob Milgram – wenn auch nicht systematisch – die körperlichen Reaktionen der Testpersonen während des Experiments. Nervosität, Schweißausbrüche und Augenzwinkern sah er als ein Indiz an, dass die Testpersonen die Situation als real annahmen (vgl. Milgram 1963, 1965a). Zweitens ließ er in zeitlicher Distanz seine Testpersonen befragen, ob sie die Situation als real oder als nicht real eingeschätzt hatten. In dieser Umfrage gaben nahezu alle Befragten an, dass sie das Experiment als real eingeschätzt hatten (vgl. Milgram 1972). Drittens führten Charles F. Sheridan und Richard G. King ein Experiment durch, in dem einem Hundewelpen reale Stromstöße versetzt wurden. Auch hier wurde eine ähnliche Gehorsamsbereitschaft nachgewiesen wie bei dem ursprünglichen Milgram-Experiment (vgl. Sheridan/King 1972).

Datenanalyse

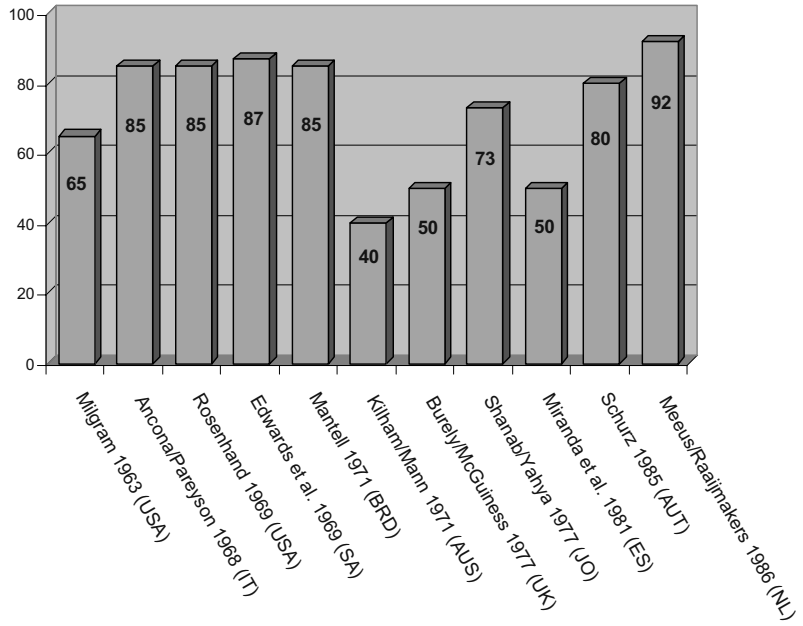
In dem durch Milgram durchgeführten Grundexperiment waren über zwei Drittel der Testpersonen bereit, Stromschläge von 450 Volt zu verabreichen (vgl. Milgram 1963). Legen wir jetzt die drei Kriterien an, mit der die Ergebnisse in der Datenanalyse überprüft werden müssen.

Statistische Auswertung des Experiments

Die relativ einfachen, vorrangig nur mit einer Variablen arbeitenden statistischen Auswertungen Milgrams wurden in der Diskussion seines Experiments kaum angezweifelt. Die auch für Laien nachvollziehbaren Angaben machten es leicht, seine Ergebnisse mit denen von ähnlichen Experimenten in anderen Ländern zu vergleichen.

In Replikationen des Milgram’schen Grundexperiment wurden seine Ergebnisse weitgehend bestätigt (s. *Abbildung 1*). In (manchmal leicht variierten) Experimenten in Italien (Ancona/Pareyson 1968), den USA (Rosenhand 1969), Südafrika (Edwards et al. 1969), Deutschland (Mantell 1971a, 1971b), Australien (Kilham/Mann 1971), Großbritannien (Burely/McGuiness 1977), Jordanien (Shanab/Yahya 1977), Spanien (Miranda et al. 1981), Österreich (Schurz 1985) und den Niederlanden (Meeus/Raaijmakers 1986) ergaben sich ähnlich hohe Prozentsätze wie bei Milgram. Die Variation der Ergebnisse von 40 bis 92 Prozent sind wohl eher Variationen im Versuchsaufbau, denn auf nationale Besonderheiten in der Gehorsamsbereitschaft zurückzuführen.

Abbildung 1: Ergebnisse des Baseline-Experiment von Milgram und dessen Replikationen: Anteil der Versuchspersonen, die bis zur Obergrenze von 450 Volt Stromstöße versetzten (siehe die Übersichten bei Smith/Bond 1993, S. 20; Blass 2000, S. 59; Neubacher 2002, S. 54)



Zusammenhang von Sachhypothese und statistischer Hypothese

Bei der Prüfung des Zusammenhangs von Sachhypothese und statistische Hypothese ist die Hauptfrage, ob durch das Experiment wirklich Gehorsamsbereitschaft gemessen wurde. Nach den ersten Veröffentlichungen von Milgram in den sechziger Jahren (Milgram 1963, 1964a, 1964b, 1965a, 1965b) hätte man noch mit guten Gründen behaupten können, dass Milgrams Untersuchungen sowohl methodisch als auch statistisch korrekt seien, er aber nicht Gehorsamsbereitschaft, sondern Aggressionspotenziale gemessen habe (vgl. Blass 1992b, S. 299 f.).

Milgram war sich schon während der Durchführung der Experimentreihe bewusst, dass er durch eine Reihe von Experimenten nachweisen musste, dass von ihm Gehorsamsbereitschaft und nicht Aggression getestet wurde. Im Einzelnen plausibilisierte Milgram in drei Kontrollexperimenten, dass die Stromstöße aus Gehorsamsbereitschaft und nicht aus Aggression gesetzt wurden.

Milgram ließ im Experiment 11 die Versuchsperson die Schockhöhe selbst wählen. Die Testpersonen mussten also nicht bei jedem neuen Fehler des vermeintlichen Schüler die Höhe der Stromstöße kontinuierlich steigern, sondern konnten bei jedem Fehler selbst die Höhe des Stromstosses frei wählen. Seine später bestätigte Vermutung war, dass bei

diesem Experimentaufbau sowohl die Höhe der durchschnittlichen gesetzten Stromschläge als auch die Anzahl der Personen, die Stromschläge von 450 Volt setzen würden, geringer sein würden als im Fall der Steigerung der Stromschlaghöhen.

Weiterhin baute Milgram im Experiment 12 die Versuchsbedingung so um, dass nach dem 150-Volt-Schlag, der Experimentleiter die Einstellung des Experiments verlangte, während aber der vermeintliche Schüler lauthals die Fortsetzung der Bestrafung forderte. Weil ausnahmslos die Testpersonen der Anweisung des Experimentleiters und nicht des Schülers folgten, hatte Milgram ein weiteres Indiz, dass das Verhalten der Testpersonen sich nicht durch Freude an der Bestrafung, sondern durch Gehorsamsbereitschaft erklären lässt.

Ferner veränderte Milgram im Experiment 13 die Bedingungen so, dass der Experimentleiter die Testperson nicht überwachte. In dieser Variation führte der Experimentleiter die Testperson in das Experiment ein, verließ dann den Raum und gab seine Anweisungen per Telefon. Es war zu beobachten, dass eine Anzahl von Testpersonen dem Experimentleiter meldete, die angesetzte Höhe des Stromschlages gesetzt zu haben, in Wirklichkeit aber einen deutlich geringeren oder gar keinen Stromstoß verabreichten. Auch dies konnte als Beleg dafür gewertet werden, dass die ursprüngliche, auf Gehorsamsbereitschaft ausgerichtete Sachhypothese durch den Experimentaufbau korrekt wiedergegeben wurde.

Bestimmung der Reichweite des Experiments

Stanley Milgram tendierte dazu, die Ergebnisse seiner Experimente als Aussagen über Gehorsamsbereitschaft und Autoritätshörigkeit in der modernen Gesellschaft insgesamt zu begreifen. Milgram erklärt das Verhalten der Versuchspersonen damit, dass sie in seinem Experiment in gesellschaftliche Strukturen wie Wertsysteme und Autoritätsbeziehungen eingebunden sind, aus denen sie nur mit großen Schwierigkeiten aussteigen können. Charakteristisch für den breiten Erklärungsanspruch ist Milgrams Frage, wozu eine Regierung mit all ihrer Autorität und ihrem Prestige fähig ist, wenn bereits ein unbekannter Experimentleiter Erwachsene dazu bringen kann, einen fünfzigjährigen Mann zu unterdrücken und ihm schmerzhafte Elektroschocks zu versetzen (Milgram 1965, S. 75; siehe Kroner 1988, S. 19 für eine Verallgemeinerung auf die Konfliktsituationen zwischen Staaten). Die Methodenkritik stellte aus zwei Positionen die Generalisierung Milgrams in Frage.

Auf der einen Seite wurde behauptet, dass die Experimente lediglich das Verhalten gegenüber wissenschaftlichen Autoritäten widerspiegelten. Die Experimente würden nur zeigen, welch enorm wichtige Rolle die Wissenschaft in modernen Gesellschaften spiele, so dass sich kaum eine Person vorstellen könne, im Namen der Wissenschaft würde „falsch“ gehandelt. Genauso wie Abraham sich nicht vorstellen konnte, dass Gott sich irre, als er von ihm verlangte, seinen Sohn zu töten, hätten sich die Testpersonen im Milgram-Experiment nicht vorstellen können, dass im Namen der Wissenschaft Unrecht geschehe (vgl. Fromm 1973, S. 74; Patten 1977a, S. 438 f., 1977b, S. 350 ff.). Milgram konnte diese Kritik jedoch teilweise entkräften, weil er im so genannten Bridgeport-Experiment (Experiment 10) zeigte, dass die Verlagerung in Gebäude außerhalb der Universität und der Verzicht auf

einige Insignien der wissenschaftlichen Autorität die Gehorsamsbereitschaft nicht signifikant reduzierte (Milgram 1974, S. 72 ff.).

Aus der anderen Position wurde die These aufgestellt, dass das Milgram-Experiment nicht wie häufig impliziert allgemeines Verhalten in Gesellschaften, sondern nur das Verhalten in Organisationen simuliert. In den Experimenten lässt sich, so die Argumentation, die Selbstbindung von Personen an eine, wenn auch kurzfristige Organisationsmitgliedschaft beobachten. Gerade weil der „Eintritt" in das Experiment freiwillig ist, fällt der „Austritt" so schwer (vgl. Indizien bei Milgram 1974, S. 140 ff.; Miller 1986, S. 225 f.). Weil der Eintritt nicht erzwungen wird, binden sich die Mitglieder an eigene Entscheidungen. Sie „verlieren ihr Gesicht", wenn sie kurz nach dem Einstieg in eine Organisation schon wieder aussteigen (siehe auch Silver/Sabini/Parrott 1987, S. 47 ff.). Die Schlussfolgerung, die daraus gezogen werden kann ist, dass Organisationen, die auf einer Freiwilligkeit des Ein- und Austritts aufbauen, in ihren Verhaltenserwartungen an Mitglieder weiter gehen können als Organisationen, die sich des Mechanismus der Zwangsmitgliedschaft bedienen (vgl. Kühl 2005).

Bewertung

Kaum ein Experiment ist so intensiv analysiert und heftig kritisiert worden wie das von Milgram. Es ist beachtlich, dass Milgrams Untersuchungen diese Auseinandersetzung fast unbeschadet überstanden. Es herrscht weitgehende Einigkeit darüber, dass Milgrams Experimente nicht nur originell konzipiert waren, sondern auch methodisch sauber durchgeführt wurden. Aufgrund der gelungenen Täuschung der Testpersonen über das Leiden des Schülers bestehen kaum Zweifel daran, dass die Testpersonen die Experimente tatsächlich als eine reale Situation begriffen hatten.

Das Milgram-Experiment zeigt in fast idealtpyischer Weise, die Vorteile der Kombination einer Beobachtung oder einer Befragung mit einem Experiment gegenüber einer Beobachtung oder Befragung ohne vorgeschaltete experimentelle Situation. Es herrscht weitgehende Übereinstimmung, dass eine nicht mit einem Experiment kombinierte schriftliche oder mündliche Befragung zur Gehorsambereitschaft ungeeignet ist, weil die Antworten durch soziale Erwünschtheit verzerrt würde. Befragungen hätten nur allgemeine Einstellungen zur Gehorsamkeit reproduziert und nicht die in sozialen Situationen wirkenden Kräfte (vgl. Neubacher 2002, S. 46).

5 Möglichkeiten und Grenzen der Methode

Letztlich ist ein Experiment eine Befragung oder Beobachtung mit einem vorgeschalteten Impuls. Weil dieser Impuls genauso sorgfältig geplant, durchgeführt und kontrolliert werden muss, wie die anschließende Datenerhebung bedeutet ein Experiment immer mehr Aufwand als eine einfache Befragung oder Beobachtung. Warum sollte man überhaupt diesen Aufwand betreiben?

Der erste Vorteil ist, dass man genau bestimmen kann, was die Ursache und was der Effekt ist. Durch Befragungen oder Beobachtungen ohne vorheriges Experiment kann man lediglich feststellen, dass zwei Variablen miteinander korrelieren. Beispielsweise kann man

beobachten, dass in Burschenschaften, die brutale Initiationsriten haben, die gegenseitigen Sympathien zwischen den Burschenschaftlern besonders ausgeprägt sind. Aber man kann sich nicht sicher sein, ob die große Sympathie durch die Quälereien in der Initiationsphase ausgelöst wird. Es wäre ja auch vorstellbar, dass sich die Gruppe der Burschenschaftler zuerst besonders sympathisch ist, in dieses Bündnis enger Freunde nicht jeden aufnehmen will und deswegen besonders hohe Hürden der Aufnahme legt (vgl. Aronson/Carlsmith 1968, S. 7). Erst im Laborexperiment ist es dem Forscher möglich zu bestimmen, ob die Quälerei oder die Sympathie die Ursache war, weil er ja selbst den Faktor Quälerei initiiert.

Der zweite Vorteil ist, dass die Einflüsse, die einen Effekt produzieren, durch den Forscher kontrolliert werden können. Bei einer Befragung oder Beobachtung ohne vorheriges Experiment kann man zwar Effekte wie Arbeitszufriedenheit, Produktivität oder politische Einstellung messen, es ist aber schwierig genau zu bestimmen, was diese Effekte beeinflusst hat. Im Laborexperiment hat man – idealer Weise – alle anderen Faktoren (die Störvariablen) so im Griff, dass man die Arbeitszufriedenheit, Produktivität oder politische Einstellung auf eine einzige, vom Versuchsleiter manipulierte Variable zurückführen kann.

Die Hauptkritik an Laborexperimenten richtete sich gegen die mangelhafte Repräsentanz der Experimentalsituation. Schon Muzafer Sherif, einer der Urväter der sozialwissenschaftlichen Experimentalforschung, verwies auf die Gefahr der Künstlichkeit einer Laboratmosphäre. Die Laborsituation könnte so verkünstelt sein, dass die beobachten Prozesse wenig mit denen zu tun haben, die wir im „wirklichen Leben“ beobachten (vgl. Sherif 1936, S. 68).

Im Einzelnen lassen sich drei Aspekte unterscheiden, durch die die Situation im Experiment verkünstelt wird (vgl. Tunnell 1977, S. 426 ff.). Erstens erzeugt bereits die Manipulation durch den Experimentleiter eine künstliche Situation. Es ist ja nicht gesagt, dass die Manipulationen des Experimentleiters auch in der Alltagsrealität so auftreten. Zweitens erzeugt die Verortung im Labor eine künstliche Situation. Es kann gut sein, dass sich Personen in einem Labor anders verhalten als an ihrem Arbeitsplatz am Fließband, in einem Konferenzraum oder im Büro eines Kunden. Sie wissen, dass sie sich in einer artifiziellen Situation befinden und ihr Verhalten keine schwerwiegenden Folgen hat. Drittens kann instruiertes Verhalten zu einer künstlichen, verzerrenden Situation führen. Wenn man Testpersonen bietet ihre Handlungen immer auch mündlich oder schriftlich zu kommunizieren, dann kann dieses „Multi-Tasking“ dazu führen, dass die Handlungen ganz anders durchgeführt werden.

Die sozialwissenschaftliche Experimentalforschung hat auf zwei Arten versucht, die Künstlichkeit der Situation zu reduzieren (vgl. dazu Aranson/Carlsmith 1968, S. 22). Die eine Strategie dient der Steigerung des „weltlichen Realismus„. Die Experimente sollten so gestaltet werden, dass sie möglichst stark einer Alltagssituation ähneln. Die zweite Strategie hat das Ziel den „experimentellen Realismus„ zu steigern. Dafür muss für die Testperson ein echtes Interesse mit dem Experiment verbunden sein und sie darf ihr Verhalten nicht durch die Künstlichkeit der Situation erklären können. Beide Strategien müssen sich nicht ausschließen, in der Praxis befindet sich der Sozialwissenschaftler jedoch häufig in zwei methodischen Dilemmas.

Das Dilemma zwischen interner und externer Validität

Schon frühzeitig führte die Unzufriedenheit mit dem Relevanzproblem Forscher dazu, nach Möglichkeiten zu suchen, wie Experimente in einem „natürlichen Kontext" durchgeführt werden können. Statt soziale Ereignisse in „das Schnürband eines aseptischen, gekünstelten Designs zu zwängen" und damit notgedrungen einen „Verlust des Informations- und Bedeutungsgehaltes" zu riskieren sollten die Experimente, so die Vorstellung, in die Alltäglichkeit des sozialen „Feldes verlagert" werden (vgl. Kordes 1994, S. 150 ff.; zur Künstlichkeit von Feldexperimenten siehe jedoch Bungard/Bay 1982, S. 192 ff.). Feldexperimente bringen jedoch das Problem mit sich, dass sich die Randbedingungen nicht gut kontrollieren lassen. Auch bei größten Bemühungen gelingt es nicht die alltägliche Welt von Testpersonen so zu konstruieren oder zu kontrollieren wie in einem Laborexperiment.

Hinter der Frage zwischen Feld- und einem Laborexperiment steckt ein generelles Dilemma der experimentellen Forschung (vgl. Campbell 1957; Cook/Campell 1976). Eine Erhöhung der externen Validität, also der Realitätsnähe eines Experiments, macht es schwieriger das Experiment zu standardisieren und reduziert dadurch die interne Validität. Eine Erhöhung der internen Validität, also das Ausschließen aller möglichen Störvariablen, verringert notgedrungen die Realitätsnähe des Experiments und reduziert die externe Validität (vgl. Schnell/Hill/Esser 1992, S. 238 f.). Mit den Bemühen um zunehmende Kontrolle der Störvariablen, geht eine wachsende Irrelevanz einher (vgl. auch Holzkamp 1970, S. 11 ff.).

Tabelle 2: Labor- und Feldexperiment zwischen interner und externer Validität

	Vorteil	**Nachteil**
Labor-experiment	*hohe interne Validität:* man kann mit hoher Sicherheit sagen, dass die beobachteten Effekte auf die Variationen des Experiment-leiters zurückgehen	*niedrige externe Validität*: da der Kontext des Experiments stark standardisiert wird, kann das Experiment nur schwer Alltags-situationen widerspiegeln
Feld-experiment	*hohe externe Validität*: da Feld-experimente in der natürlichen Umgebung der Testperson durchgeführt wird, bilden sie die Realität recht genau ab	*niedrige interne Validität*: da Feld-experimente in der natürlichen Um-gebung der Testpersonen durch-geführt werden, lassen sich die be-obachteten Effekte kausal nur schwerlich auf die Variationen des Experimentleiters zurechnen

Das Dilemma zwischen externer Validität und Aufklärung der Testperson

Das Problem der externen Validität (Repräsentanzproblem) von Experimenten führte dazu, dass die Experimentalforscher Methoden ersannen, mit denen auch Laborexperimente möglichst realitätsnah gestalten werden konnten. Eine der vielversprechendsten Strategien war es, die Testpersonen über den Aufbau des Experiments und die realen Auswirkungen ihrer Handlungen zu täuschen. Der Clou von Stanley Milgrams Experiment bestand beispielsweise darin, dass die Testpersonen über die wirklichen Auswirkungen ihrer Handlungen im Unklaren gelassen wurden. Die Testpersonen mussten aufgrund der Infor-

mationen des Experimentleiters davon ausgehen, dass der Testperson reale Stromstöße versetzt wurden. Damit wurde ihnen die Möglichkeit genommen, ihre Handlungen damit zu rechtfertigen, dass es sich ja nur um ein Spiel handele.

Genau an dieser Täuschung setzen jedoch forschungsethische Bedenken an. Es wird es als problematisch angesehen, wenn Testpersonen über die Ziele des Versuches getäuscht und sie dadurch in extreme Stresssituationen gebracht werden. Diane Baumrind kritisierte beispielsweise am Milgram-Experiment, dass durch die Täuschungen der Testpersonen deren Würde, Selbstbewusstsein und Vertrauen in Autoritäten gestört wurde und dadurch langfristig Schäden bei der Testperson hervorgerufen werden könnten (vgl. Baumrind 1964; Erwiderung von Milgram 1964b).

Als forschungsethisch korrekte Alternative wurde vorgeschlagen, Testpersonen vollständig über die Ziele des Experimentes aufzuklären, sie dann zu bitten die Experimente zu spielen und sich dabei so zu verhalten, als wären sie über die Konsequenzen ihres Verhalten nicht aufgeklärt worden. Don Mixon (1971) setzte beispielsweise bei seinen Untersuchungen zum Milgram-Experiment auf nichtaktive Rollenspielprozeduren. Den Testpersonen wurde der erste Teil des Milgram-Experiments vorgestellt und die Testpersonen dann gebeten, einzuschätzen, wie sich die Testpersonen wohl weiter verhalten würden. Das Problem dieser forschungsethisch unbedenklichen Vorgehensweise ist jedoch, dass die externe Validität des Experiments leidet. Gespielter Experimente („ich erzähle Ihnen jetzt den Aufbau des Milgram-Experiments und bitte Sie dann, sich so zu verhalten, als ob Sie das alles nicht wüssten“) drohen von den Testpersonen nicht in der gleichen Weise ernst genommen zu werden wie Experimente, in denen sie über die Auswirkungen ihres Handelns getäuscht werden.

Ausblick

Die Verbindung des Experiments zu anderen Methoden der Organisationsforschung stellt – wie am Anfang gezeigt – eine Besonderheit dar. Ein Experiment spricht nie für sich selbst. Die durch das Experiment erzeugten Reaktionen der Testpersonen müssen durch Methoden der Beobachtung, der Befragung oder der Dokumentenanalyse erst erhoben werden. Weit entwickelt ist dabei auch die Kombination verschiedener Methoden (Triangulation) bei der Auswertung von Experimenten (zu dieser „between-method triangulation“ siehe Denzin 1978, S. 301 ff.).

Was auffällt ist, wie selten experimentell arbeitende Untersuchungen mit anderen, nichtexperimentellen Untersuchungen kombiniert werden. Während es ausgearbeitete Ansätze gibt, wie qualitative Feldstudien und quantitative Erhebungen miteinander kombiniert werden können (siehe z. B. Vidich/Shapiro 1955; Sieber 1973; Freter/Hollstein/Werle 1992), gibt es solche Überlegungen zur Kombination von experimenteller mit nichtexperimenteller Forschung nur sehr vereinzelt. Es fällt zum Beispiel beim Milgram Experiment auf, dass die experimentellen Ergebnisse ad hoc mit anderem Datenmaterial (z. B. historische Akten über den Holocaust) in Verbindung gebracht werden, es aber keine Versuche gegeben hat, die experimentelle Forschung mit einem nichtexperimentellen Forschungsansatz zu kombinieren.

Die Gründe für diese Berührungsängste sind vielfältig. Ein erster Grund liegt sicherlich darin, dass die experimentelle Forschung so aufwändig ist. Forscher „erschöpfen“ sich

in der Durchführung und Auswertung der Experimente und für die Entwicklung eines zweiten Untersuchungsdesigns fehlt ihnen dann die Kraft. Ein zweiter Grund mag sein, dass sich die sozialwissenschaftliche Experimentalforschung in den letzten hundert Jahren ein hohes Maß an Spezialisierung erreicht hat. Es gibt viele Forscher, deren Kompetenzen in der Durchführung und Auswertung von Experimenten liegen und die keine Gründe sehen – so lange diese Spezialisierung als „Experimentalforscher" nicht kritisiert wird – mit nichtexperimentellen Forschungsansätzen zu arbeiten.

Verschenkt wird dadurch die Möglichkeit das Problem der externen Validität durch den Einbezug weiterer Methoden in den Griff zu bekommen. Die systematische Kombination von experimenteller und nichtexperimenteller Forschung böte die Chance, dass die Experimentalforschung sich von dem Hauptkritikpunkt der Künstlichkeit (externe Validität) ihrer Experimentalsituationen wenigstens teilweise befreit und für ihre Ergebnisse ein noch höheres Maß an Validität erzeugt.

6 Literatur

Ancona, L./Pareyson, R. (1968): Contribute alle studie della aggressione: La dinamica della obedienza distruttiva, in: Archivio di Psicologia, Neurologia e Psichiatria, 29, S. 340–372

Aronson, Elliot/Carlsmith, J. Merrill (1968): Experimentation in Social Psychology, in: Lindzey, Gardner/Aronson, Elliot (Hrsg.), The Handbook of Social Psychology, Bd. 2, 2. Auflage, Reading, S. 9–79

Asch, Solomon E. (1951): Effects of Group Pressure upon the Modification and Distortion of Judgements, in: Guetzkow, Harold (Hrsg.), Groups, Leadership, and Men, Pittsburg, S. 177–190

Asch, Solomon E. (1955): Opinions and Social Pressure, in: Scientific American, 5/1995, S. 31–35

Baumrind, Diana (1964): Some Thoughts on Ethics of Research. After Reading Milgram´s Behavorial Study of Obedience, in: American Psychologist, 19, S. 421–423

Berg, Perdita (1988): Das Verhalten von Schülern in dem Planspiel „Das Dritte Reich – bewältigte Vergangenheit?". Empirische Untersuchung und Interpretation unter Berücksichtigung psychologischer Faschismustheorie, Hamburg: Diplomarbeit Fachbereich Psychologie der Uni Hamburg

Blass, Thomas (1992): The Social Psychology of Stanley Milgram, in: Zanna, Mark P. (Hrsg.), Advances in Experimental Social Psychology, 25, San Diego, S. 277–329

Blass, Thomas (2000): The Milgram Paradigm after 35 Years: Some Things we now Know about Obedience to Authority, in: Blass, Thomas (Hrsg.), Obedience to Authority. Current Perspectives on the Milgram Paradigm, Mahwah, S. 39–59

Boal, Augusto (1979): Theater der Unterdrückten, Frankfurt a. M.

Bortz, Jürgen (1984): Lehrbuch der empirischen Forschung für Sozialwissenschaftler, Berlin

Bramel, Dana/Friend, Ronald (1981): Hawthorne, the Myth of the Docile Worker, and Class Bias in Psychology, in: American Psychologist, 36, S. 867–878

Bungard, Walter, Rolf Bay (1982): Feldexperimente in der Sozialpsychologie, in: Patry, Jean-Lux (Hrsg.), Feldforschung, Bern, S. 183–205

Burley, Peter M./McGuiness, John (1977): Effects of Social Intelligence on the Milgram Paradigm, in: Psychological Reports, 40, S. 767–700

Campbell, Donald T. (1957): Factors Relevant to Validity of Experiments in Social Settings, in: Psychological Bulletin, 54, S. 297–312

Campbell, Donald Thomas/Stanley, Julian C. (1963): Experimental and Quasi-Experimental Designs for Research on Teaching, in: Gage, Nathaniel L. (Hrsg.), Handbook of Research on Teaching, Chicago, S. 171–246

Cook, Thomas/Campbell, Donald Thomas (1976): Quasi-experimentation, Chicago

Czienskowski, Uwe (1996): Wissenschaftliche Experimente: Planung, Auswertung, Interpretation, Weinheim
Diekmann, Andreas (1998): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, 4. Auflage, Reinbek
Edwards, D. M. et al. (1969): An Experiment on Obedience, Johannesburg: Unpublished Student Report, University of the Witwatersrand
Fromm, Erich (1973): The Anatomy of Human Destructiveness, Greenwich
Garfinkel, Harold (1973): Das Alltagswissen über soziale und innerhalb sozialer Strukturen, in: Arbeitsgruppe Bielefelder Soziologen (Hrsg.), Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit, Bd. 1, Reinbek, S. 189–261
Greenwood, John D. (1982): On the Relation Between Laboratory Experiments and Social Behavior: Causal Explanation and Generalization, in: Journal for the Theory of Social Behavior, 12, S. 225–250
Greenwood, John D. (1989): Explanation and Experiment in Social Psychological Science. Realism and the Social Constitution of Action, New York
Hagmüller, Peter (1979): Empirische Forschungsmethoden. Eine Einführung für pädagogische und soziale Berufe, München
Haney, Craig/Banks, Curtis/Zimbardo, Philip G. (1973): Interpersonal Dynamics in a Simulated Prison, in: International Journal of Criminology and Penology, 1, S. 69–97
Heller, Kurt/Rosemann, Bernhard (1974): Planung und Auswertung empirischer Untersuchungen. Eine Einführung für Pädagogen, Psychologen und Soziologen, Stuttgart
Henning, Hans Jörg/Muthig, Klaus (1979): Grundlagen konstrutiver Versuchsplanung, München
Holzkamp, Klaus (1964): Theorie und Experiment in der Psychologie. Eine grundlagenkritische Untersuchung, Berlin
Holzkamp, Klaus (1970): Zum Problem der Relevanz psychologischer Forschung für die Praxis, in: Psychologische Rundschau, 21, S. 1–22
Huber, Oswald (2000): Das psychologische Experiment: Eine Einführung, 3. Auflage, Bern
Kieserling, André (1994): Organisationssoziologie und Unternehmensberatung, Bielefeld: unveröff. Ms.
Kieserling, André (1999): Kommunikation unter Anwesenden. Studien über Interaktionssysteme, Frankfurt a. M.
Kilham, Wesley/Mann, Leon (1974): Level of Destructive Obedience as a Function of Transmitter and Executant Roles in the Milgram Obedience Paradigm, in: Journal of Personality and Social Psychology, 29, S. 696–702
Kordes, Hagen (1994): Das Aussonderungs-Experiment. Rechenschaftsbericht zum „Krisenexperiment“ der Aussonderung von „Deutschen“ und „Ausländern“ durchgeführt vor einer Mensa der Universität Münster am 28. Januar 1994, Münster
Kraus, Andreas (1987): Das Dritte Reich – bewältigte Vergangenheit. Ein erfahrungsbezogenes Unterrichtsprojekt zur schulischen politischen Bildung in einer 11. Klasse, Hannover
Kraus, Andreas (2003): Das Dritte Reich – bewältigte Vergangenheit. Ein Planspiel, Hannover
Kroner, Bernhard (1988): Gegen den Pessimismus des Milgram-Experiments, in: Bielefelder Arbeiten zur Sozialpsychologie Nr. 139, Bielefeld
Kühl, Stefan (2003a): Organisationssoziologie. Ein Ordnungs- und Verortungsversuch, in: Soziologie, 1/2003, S. 37–47
Kühl, Stefan (2003b): Assessment-Center. Teures Alibi, in: Management & Training 8,/2003, S. 11
Kühl, Stefan (2005): Ganz normale Organisationen. Organisationssoziologische Interpretationen simulierter Brutalitäten, in: Zeitschrift für Soziologie, 34, erscheint in Heft 1
Lapiere, Richard Tracy (1934): Attitudes vs. Actions, in: Social Forces, 14, S. 230–237
Luhmann, Niklas (1964): Funktionen und Folgen formaler Organisation, Berlin
Luhmann, Niklas (1973): Zweckbegriff und Systemrationalität. Über die Funktion von Zwecken in sozialen Systemen, Frankfurt a. M.
Luhmann, Niklas (1997): Die Gesellschaft der Gesellschaft, Frankfurt a. M.
Mann, Leon (1999): Sozialpsychologie, Weinheim

Mantell, David (1971a): The Potenzial for Violence in Germany, in: Journal of Social Issues, 27, S. 101–112
Mantell, David (1971b): Das Potenzial zur Gewalt in Deutschland. Eine Replikation und Erweiterung des Milgramschen Experiments, in: Der Nervenarzt, 5, S. 252–257
Meeus, Wim H. J./Raaijmakers, Quinten A. W. (1986): Administrative Obedience. Carrying Out Orders to Use Psychological-Administrative Violence, in: European Journal of Social Psychology, 16, S. 311–324
Mertens, Wolfgang (1975): Sozialpsychologie des Experiments. Das Experiment als soziale Interaktion, Hamburg
Milgram, Stanley (1963): Behavioral Study of Obedience, in: Journal of Abnormal and Social Psychology, 67, S. 371–378
Milgram, Stanley (1964a): Group Pressure and Action Against a Person, in: Journal of Abnormal and Social Psychology, 69, S. 137–143
Milgram, Stanley (1964b): Issues in the Study of Obedience. A Reply to Baumrind, in: American Psychologist, 19, S. 848–852
Milgram, Stanley (1965a): Some Conditions of Obedience and Disobedience to Authority, in: Human Relations, 18, S. 57–76
Milgram, Stanley (1965b): Liberating Effects of Group Pressure, in: Journal of Personality and Social Psychology, 1, S. 127–134
Milgram, Stanley (1972): Interpreting Obedience. Error and Evidence (A reply to Orne and Holland), in: Miller, Arthur G. (Hrsg.), The Social Psychology of Psychological Research, New York, S. 138–154
Milgram, Stanley (1974): Obedience to Authority. An Experimental View, New York
Miller, Arthur G. (1986): The Obedience Experiments, New York
Miranda, Francisca S. et al. (1981): Obediencia a la autoridad, in: Psiquis, 2, S. 212–221
Mittenecker, Erich (1964): Planung und statistische Auswertung von Experimenten, Wien
Mixon, Don (1971): Further Conditions of Obedience and Disobedience to Authority, in: Disseration Abstracts International, 32, No 4646B
Moldaschl, Manfred/Weber, Wolfgang G. (1998): The „Three Waves" of Industrial Group Work. Historical Reflections on Current Research on Group Work, in: Human Relations, 51, S. 347–388
Neubacher, Frank (2002): Verbrechen aus Gehorsam – Folgerungen aus dem Milgram-Experiment für Strafrecht und Kriminologie, in: Neubacher, Frank/Walter, Michael (Hrsg.), Sozialpsychologische Experimente in der Kriminologie. Milgram, Zimbardo und Rosenhan kriminologisch gedeutet, mit einem Seitenblick auf Dürrenmatt, Münster, S. 43–68
Orne, Martin T./Holland, Charles H. (1968): On the Ecological Validity of Laboratory Deceptions, in: International Journal of Psychiatry, 6, S. 282–293
Osnabrügge, Gabriele/Frei, Dieter (1989): Experiment, in: Endruweit, Günter/Trommsdorff, Gisela (Hrsg.), Wörterbuch der Soziologie, Stuttgart, S. 180–187
Patten, Steven (1977a): The Case That Milgram Makes, in: Philosophical Review, 86, S. 350–364
Patten, Steven (1977b): Milgram´s Shocking Experiments, in: Philosophy, 52, S. 425–440
Roethlisberger, Fritz Jules/Dickson, William J. (1939): Management and the Worker. An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago, Cambridge
Rosenhan, David L. (1969): Some Origins of Concern to Others, in: Mussen, P. H./Langer, J./Covington, M. (Hrsg.), Trends and Issues in Developmental Psychology, New York, S. 134–153
Schnell, Rainer/Hill, Paul B./Esser, Elke (1992): Methoden der empirischen Sozialforschung, 3. überarb. und erw. Auflage, München
Schultz, Diane P. (1969): The Human Subject in Psychological Research, in: Psychological Bulletin, 72, S. 214–228
Schurz, Grete (1985): Experimentelle Überprüfung des Zusammenhangs zwischen Persönlichkeitsmerkmalen und der Bereitschaft zum destruktiven Gehorsam gegenüber Autoritäten, in: Zeitschrift für Experimentelle und Angewandte Psychologie, 32, S. 160–177

Seidman, David et al. (1957): Influence of a Partner on Tolerance for Self-administered Electric Shock, in: Journal of Abnormal and Social Psychology, 54, S. 210–212
Shanab, Mitri E./Yahya, Khawla A. (1977): A Behavioral Study of Obedience in Children, in: Journal of Personality and Social Psychology, 35, S. 530–536
Shanab, Mitri E./Yahya Khawla A. (1978): A Cross-Cultural Study of Obedience, in: Bulletin of the Psychonomic Society, 11, S. 267–269
Sheridan, Charles L./King Richard G. (1972): Obedience to Authority with an Authentic Victim, in: Proceedings of the American Psychological Association, S. 165–166
Sherif, Muzafer (1936): The Psychology of Social Norms, New York
Sherif, Muzafer (1954): Integrating Field Work and Laboratory in Small Group Research, in: American Sociological Review, 19, S. 759–771
Sherif, Muzafer (1958): Superordinate Goals in the Reduction of Intergroup Conflict, in: American Journal of Sociology, 63, S. 349–356
Silver, Maury/Sabini, John/Parrott, W. Gerrod (1987): Embarrassment: A Dramaturgic Account, in: Journal for the Theory of Social Behavior, 17, S. 47–61
Smith, Peter B./Bond Michael H. (1993): Social Psychology. Across Cultures. Analysis and Perspectives, New York
Strodtholz, Petra/Kühl, Stefan (2002): Qualitative Methoden der Organisationsforschung – ein Überblick, in: Kühl, Stefan/Strodtholz, Petra (Hrsg.), Methoden der Organisationsforschung. Ein Handbuch, Reinbek, S. 11–32
Tunnell, G. B. (1977): Three Dimensions of Naturalness. An Expanded Definition of Field Research, in: Psychological Bulletin, 84, S. 426–437
Tyrell, Hartmann (1983): Zwischen Interaktion und Organisation. Gruppe als Systemtyp, in: Neidhardt, Friedhelm (Hrsg.), Gruppensoziologie – Perspektiven und Materialien, Opladen, S. 75–87
Weick, Karl E. (1965): Laboratory Experimentation with Organizations, in: March, James G. (Hrsg.), Handbook of Organizations, Chicago, S. 194–260
Zelditch, Morris/Hopkins, Terrence K. (1961): Laboratory Experiments with Organizations, in: Etzioni, Amitai (Hrsg.), Complex Organizations. A Sociological Reader, New York, S. 464–478
Zimbardo, Philip G. et al. (1973): The Mind is a Formidable Jailer. A Pirandellian Prison, in: New York Times Magazine, 8.4.1973, S. 38–60
Zimbardo, Philip G. et al. (1975): The Psychology of Imprisonment: Privation, Power, and Pathology, in: Rosenhan, David/London, Perry (Hrsg.), Theory and Research in Abnormal Psychology, 2. Auflage, New York, S. 270–287