Sozialwissenschaftliche Arbeitsmethoden für Mediziner, Soziologen, Psychologen

Lamnek, Siegfried

Veröffentlichungsversion / Published Version
Monographie / monograph

Sozialwissenschaftliche Arbeitsmethoden

Für Mediziner, Soziologen, Psychologen

Von Siegfried Lamnek

1. Auflage 1980

edition medizin
Weinheim · Deerfield Beach, Florida · Basel

Dr. Siegfried Lamnek
Konradstraße 6
D-8000 München 40

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Lamnek, Siegfried:
Sozialwissenschaftliche Arbeitsmethoden für Mediziner, Soziologen, Psychologen / von Siegfried Lamnek. – 1. Aufl. – Weinheim, Deerfield Beach (Florida), Basel: Edition Medizin, 1980.
ISBN 3-527-15002-1

Druck: Mono-Satzbetrieb, D. Betz GmbH, D-6100 Darmstadt 12
Bindung: Aloys Gräf, D-6900 Heidelberg
Printed in West Germany

VORWORT

Das vorliegende Lehrbuch versucht, das für den Medizinstudierenden und den Mediziner relevante methodische Wissen zu vermitteln. Es bezieht sich auf die im Gegenstandskatalog für die Ärztliche Vorprüfung angegebenen Lernziele, die im Kapitel 1. unter 1.1., 1.2., 1.3., 1.4., 1.5. und 1.6. genannt und unter die Medizinische Psychologie und Medizinische Soziologie subsumierbar sind. Darüber hinaus werden wichtige Lernziele zum Fachgebiet Biomathematik und Statistik behandelt.

Die als Vorbereitung auf die medizinische Vorprüfung gedachte Darstellung verfährt zweigleisig: einmal wird abstrakt und theoretisch das im Gegenstandskatalog geforderte Wissen erarbeitet; zum anderen wird jeweils ein ausführliches Beispiel für die Anwendung des abstrakt Vorgestellten gegeben, um dem Leser die Möglichkeit zu eröffnen, die entwickelten Überlegungen pragmatisch nachzuvollziehen.

Die Herauslösung des „Methodischen" aus den Gebieten der Medizinischen Soziologie und Psychologie soll eine ganzheitliche Sicht der methodischen Probleme fördern, die von der Wissenschaftstheorie über die Methoden bis hin zur Statistik reicht und eine untrennbare Einheit bildet, was allzu leicht in dem Wust der Einzelprobleme aus dem Auge verloren wird.

S. Lamnek

INHALT

1 METHODOLOGISCHE GRUNDLEGUNG

Will man sich die Methoden einer Wissenschaft durch Lektüre aneignen und/oder sie in praktischer Forschung anwenden, so sollte man sich über ihre wissenschaftstheoretischen und methodologischen Grundlagen und Voraussetzungen im klaren sein. Dieses einleitende Kapitel verfolgt die Absicht, die basalen, methodologischen Kriterien empirischer Forschung knapp zu umreißen, um Stellenwert und Relativität der einzelnen Methoden erkennen zu können.

Die lange Zeit in den Wissenschaften herrschende Antinomie zwischen Theorie und Empirie bzw. zwischen philosophisch spekulativ gewonnenen „Erkenntnissen" und jenen, die für sich das Prädikat „empirisch" in Anspruch nehmen konnten, hat sich zugunsten einer Gemeinsamkeit von Theorie und Empirie aufgelöst, was sowohl für die Naturwissenschaften wie auch für die Geisteswissenschaften gilt. Unterschiede sind nur noch hinsichtlich der Priorität des einen oder anderen Elementes zu verzeichnen, was in Abhängigkeit von dem Objektbereich der jeweiligen Wissenschaften beurteilt werden muß.

Der Begriff des Empirischen meint, daß theoretische Überlegungen nicht im Spekulativen stecken bleiben dürfen, sondern daß sie auf die Erfahrung zu beziehen sind und mit den Tatsachen der Realität konfrontiert werden müssen, für die sie Aussagen machen. Die Konfrontation mit den Tatsachen der Realität erfolgt über die Beobachtung der theoretisch beschriebenen Phänomene, weil die Beobachtung als die allgemeinste Form von Erfahrung gesehen wird. (In diesem Sinne gelten alle Methoden, wie z.B. Experiment, Befragung, Testverfahren u.a. als spezielle Methoden der Beobachtung). Die Beobachtung als Mittel der Konfrontation theoretisch formulierter Sachverhalte mit der Realität ist somit Grundlage aller empirischen Wissenschaften.

1.1 Die Stellung der Methoden im wissenschaftlichen Forschungsprozeß

Akzeptiert man die kursorisch vorgestellte Prämisse, daß theoretische Aussagen einer empirischen Validierung bedürfen, so müssen Methoden entwickelt werden, die eine gültige Überprüfung der theoretischen Aussagen an der Realität zulassen. Solche Methoden können selbstverständlich nicht unabhängig von den einzelwissenschaftlichen Disziplinen gedacht und konstruiert werden; zwar ist nicht auszuschließen, daß es übergreifende Methoden für alle Wissenschaften gibt oder geben könnte, doch wird der jeweilige Objektbereich der Einzeldisziplinen Modifizierungen oder gar vollständig andere empirische Methoden vorschreiben. *Somit determiniert der Objektbereich einer jeden wissenschaftlichen Disziplin auch deren empirische Methoden.*

Die normativ geforderte Beobachtung als Kontrollinstanz der theoretischen Überlegungen kann nicht nur zu unterschiedlichen Methoden einzelwissenschaftlicher

Art gerinnen, sondern sie wird auch durch den jeweiligen Erkenntnis- oder Objektbereich in entscheidender Weise bestimmt. So ist z.B. die Kontrollwirkung der Beobachtung im Falle der *Fremdbeobachtung des Verhaltens* eine weitaus stärker abgesicherte, als im Falle der *Selbstbeobachtung des Erlebens*. Erlebensvorgänge sind intrapersonale, psychische Abläufe, die von außen nicht beobachtbar sind, sondern bestenfalls über bestimmte Theorien erschlossen werden können. Die Feststellung einer inneren Befindlichkeit mittels Beobachtung, kann daher durch Fremdbeobachtung nur über nach außen dringende overte Verhaltensweisen (z.B. Mimik, Gestik, verbale Äußerungen, Stimmungen etc.) erschlossen werden. Solche äußeren Verhaltensweisen sind der Fremdbeobachtung unmittelbar zugänglich. Die Selbstbeobachtung des eigenen Erlebens – durchaus als wissenschaftliche Methode anerkannt – ist jedoch weit weniger zuverlässig, als die „objektive" Fremdbeobachtung. Das innere Erleben (Gefühle, Neigungen, Triebe etc.), muß ja durch das betroffene Subjekt mitgeteilt, etwa verbalisiert werden, um dessen empirische Erfassung zu ermöglichen. Abgesehen davon, daß sich das Subjekt über seine innere Befindlichkeit täuschen kann, besteht die Gefahr, daß bei deren Mitteilung subjektive Verzerrungsfaktoren – gleich welcher Art – eingehen, die aber, da das eigentlich festzustellende Merkmal keiner unmittelbaren Beobachtung zugänglich ist, nicht erkannt und überprüft werden können.

Somit sind unterschiedliche Gütegrade der beiden Methoden der Selbstbeobachtung und Fremdbeobachtung zu verzeichnen. Generalisiert bedeutet dies, daß es nicht die schlichte Dichotomie hier wissenschaftliche dort unwissenschaftliche Methoden, sondern daß es graduelle Abstufungen der einzelnen Methoden auf dem Kontinuum zwischen wissenschaftlich und unwissenschaftlich gibt.

Die normative Aussage, daß einzelwissenschaftliche Methoden in der Lage sein müssen, die theoretisch behaupteten Sachverhalte empirisch zu überprüfen, bleibt als solche noch relativ inhaltsleer, wenn keine spezifischen Kriterien dafür angegeben werden, wann und unter welchen Bedingungen eine Methode dieses Ziel erreicht. Es sind daher im weiteren Kriterien zu entwickeln, die es erlauben, einzelwissenschaftliche Methoden der empirischen Feststellung und Überprüfung von theoretischen Tatsachenbehauptungen als wissenschaftlichen Anforderungen genügend zu beurteilen. Solche Kriterien liefern die Wissenschaftstheorie und die Methodologie der einzelnen Disziplinen.

1.1.1 Ziele der Wissenschaft

Jede Wissenschaft macht Aussagen über den ihr eigenen Objektbereich. Nun wird man schon vom Alltagsverständnis her und qua Dezision nicht jeden Inhalt und nicht jede Form einer Aussage als wissenschaftlich gelten lassen. So wird dogmatischen, ideologischen, mystischen und anderen „nicht-rationalen" Aussagen die Wissenschaftlichkeit abgesprochen werden müssen. Auch wird man akzidentell gewonnene Erkenntnisse, so richtig sie auch immer sein mögen, zwar als Erkenntnisse, aber nicht als wissenschaftliche Erkenntnisse anerkennen. Vor jeder methodischen Überlegung wird daher eine Grundsatzentscheidung darüber zu stehen ha-

ben, was Wissenschaftlichkeit ausmacht. In sehr allgemeiner Formulierung könnte man behaupten, daß diese Grundsatzentscheidung darin besteht, daß metaphysisch-spekulative und/oder akzidentelle Erkenntnisse nicht als wissenschaftlich bezeichnet werden können. *Wissenschaftliche Erkenntnis bedarf stets einer systematischen und empirischen Fundierung.*

Während die bisherige Festlegung dessen, was als wissenschaftlich gelten kann, eine formale und methodische Angelegenheit war, kann man das Ziel der Wissenschaft auch inhaltlich fassen. Wissenschaft sucht nicht nur nach Informationen und Erkenntnissen, die die Realität als solche beschreiben (so wichtig dies auch sein mag, was aber nur zur „Verdoppelung der Wirklichkeit" (Adorno) führen würde), sondern solche *Deskription ist nur eine notwendige Vorstufe für das eigentliche Ziel der Wissenschaften*, nämlich die Frage nach dem „Warum" zu beantworten. Kann diese Frage gelöst werden, so können Zusammenhänge zwischen Tatsachen erklärt werden. Erklärung bedeutet dabei, daß bestimmte Sachverhalte von anderen Sachverhalten, die mit ersteren in nachweisbarer Beziehung stehen, mit Hilfe von Theorien und empirischen Tatsachen als nicht unabhängig voneinander gesehen werden. (Hat man z.B. ermitteln können, daß bestimmte Hormone die Ovulation steuern, so kann letztere mit bestimmten Hormonkonstellationen im weiblichen Körper erklärt werden.)

Erklärung um der Erklärung willen ist jedoch nicht alleiniges Ziel der Wissenschaft. Vielmehr ist die Erklärung selbst notwendige Voraussetzung, um bestimmte lebenspraktische Erkenntnisse als Anwendung der Erklärung zu gewinnen. Zwei Formen praktischer Anwendung wissenschaftlicher Erkenntnisse, können unterschieden werden: Die *Prognose* und die *technologische Anweisung.* Die technologische Anweisung fragt danach, was getan werden muß, damit ein bestimmter Zustand unter der Voraussetzung eintritt, daß die zugrundeliegende Theorie, die auch für die Erklärung herangezogen wird, gültig ist. (So kann man danach fragen, welche Hormone in welcher Dosis verabreicht werden müssen, damit der Eisprung verhindert wird). Auf ähnlicher Ebene liegt das Ziel der Prognose. Man fragt dabei danach, was geschehen wird, wenn bestimmte Bedingungen realisiert werden. (Wenn wir bestimmte Hormone in bestimmter Dosis verabreichen, können wir vorhersagen, daß eine Schwangerschaft ausgeschlossen sein wird.)

Beschreibung, Erklärung, Prognose und technologische Anweisung auf systematisch wissenschaftlicher Grundlage, können damit als allgemeinste Ziele der Wissenschaften gelten. Wie aber können nun diese Ziele konkret erreicht werden? Die Antwort darauf gibt für die Wissenschaften allgemein die Wissenschaftstheorie, für die Einzeldisziplinen deren Methodologie.

1.1.2 Wissenschaftstheorie und Methodologie

Zunächst scheint eine begriffliche Abklärung und Differenzierung zwischen Wissenschaftstheorie und Methodologie erforderlich zu sein. Beide haben gemeinsam, daß sie Metatheorien darstellen; sie sind Theorien über Theorien. *Während jedoch*

die Wissenschaftstheorie sich mit der Frage beschäftigt, wie und mit Hilfe welcher Vorgehensweisen wissenschaftliche Erkenntnis überhaupt erzielbar ist, beschäftigt sich die Methodologie als Spezialfall oder Anwendungsfall der Wissenschaftstheorie mit der Frage, unter welchen Bedingungen wissenschaftliche Erkenntnis auf einen bestimmten Erkenntnis- und Objektbereich bezogen, möglich ist. Diese analytische Trennung kann im Einzelfalle nicht immer durchgehalten werden, weil allgemein wissenschaftstheoretische Implikationen auch spezifisch methodologischer Art und umgekehrt sein können.

Der Vollständigkeit halber sei noch der Begriff der *Erkenntnistheorie* erläutert: Er beschäftigt sich mit der Frage, *wie menschliche Erkenntis, gleichgültig ob wissenschaftlich, vorwissenschaftlich oder unwissenschaftlich, überhaupt möglich ist.* Wissenschaftstheorie wäre also ein Spezialfall der allgemeinen Erkenntnistheorie. Es kann daher die folgende begrifflich-inhaltliche Struktur festgestellt werden: Erkenntnistheorie → Wissenschaftstheorie → Methodologie → Methoden und Techniken.

1.1.3 Der Stellenwert der Wissenschaftstheorie

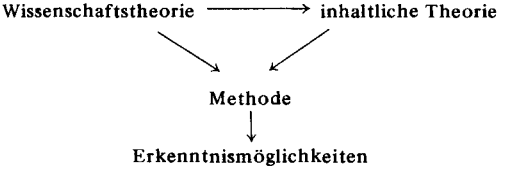
Die Wissenschaftstheorie befaßt sich mit den Überlegungen, auf welche Weise wissenschaftliche Erkenntnis zu gewinnen sei. Somit legt die Wissenschaftstheorie fest, wie und auf welche Weise vorzugehen ist. Da sie dieses generell beabsichtigt, also nicht auf bestimmte Erkenntnisobjekte ausgerichtet ist, muß sie eine Strategie wissenschaftlicher Erkenntnis festlegen, die so allgemein ist, daß sie für alle Einzeldisziplinen gelten kann, die andererseits aber auch so konkret ist, daß sie als klare Handlungsanweisung aufgefaßt werden kann.

Diese Schwierigkeit in der Vermittlung zwischen Allgemeinem und Konkretem zeigt sich darin, daß die wissenschaftstheoretischen Überlegungen logisch vor dem eigentlichen Erkenntnisprozeß liegen. Wenn sie aber vor dem Erkenntnisprozeß angesiedelt sind, dann müssen wissenschaftstheoretische Erwägungen quasi unabhängig von dem jeweiligen Erkenntnisziel auf allgemeinste Vorstellungen davon bezogen sein, wie die Erkenntnisobjekte geartet sind. Wissenschaftstheoretische Überlegungen müssen also gleichwohl das zu Erkennende in seiner potentiellen Gestalt antizipieren. Da man jedoch nie wissen kann, wie die Erkenntnisobjekte selbst aussehen, (sonst wären ja die durchzuführenden wissenschaftlichen Untersuchungen über diesen Objektbereich unsinnig, mindestens aber überflüssig), trifft die Wissenschaftstheorie bereits eine Vorentscheidung darüber, was erkannt werden kann und was aus dem Erkenntnisprozeß ausgeschlossen ist. An einem einfachen Beispiel kann dieser Sachverhalt demonstriert werden: Je nach dem, welches wissenschaftstheoretische, theoretische, oder meßtechnische Raster ich auf einen Objektbereich anlege, ergeben sich unterschiedliche Erkenntnismöglichkeiten: Die Beobachtung von Zellstrukturen mit dem menschlichen Auge allein, wird den Erkenntnisspielraum gegenüber einer mikroskopischen Betrachtung bedeutend einengen. Um jedoch das feinere Meßinstrument anlegen zu können, bedarf es vorher einer theoretischen Vermutung (Hypothese) darüber, daß kleinere Einheiten, als

sie das menschliche Auge erfassen kann, vorhanden sind. Vor dieser inhaltlich-theoretischen Dimension würden wissenschaftstheoretische Überlegungen darüber stehen müssen, das Mikroskop als Beobachtungsinstrument, als wissenschaftlichen Kriterien entsprechend, zuzulassen.

Die bisherigen Überlegungen lassen folgende Determinationsstruktur erkennen: Wissenschaftstheorie bestimmt jene Methoden, mit deren Hilfe wissenschaftliche Erkenntnis gewonnen werden soll. Die Methoden selbst bestimmen aber auch, welche Erkenntismöglichkeiten durch sie gegeben und welche ausgeschlossen sind. Auf einer dritten Ebene kann festgestellt werden, daß die inhaltliche, auf den Untersuchungsgegenstand bezogene Theorie, die Forschung ebenfalls determiniert, weil sie nur das empirisch erkennen läßt, was theoretisch-hypothetisch über den Objektbereich vermutet wurde. (Stellt man z.B. nach einer empirischen Untersuchung durch weitere theoretische Überlegungen fest, daß es sinnvoll gewesen wäre, die Variable x in den Untersuchungszusammenhang einzubeziehen, dies aber bei der Datenerhebung nicht geschehen ist, so können keine Aussagen über diese Variable gemacht werden. Es kann immer nur das empirisch erkannt und beobachtet werden, was theoretisch vorher als erkenntnisrelevant abgesteckt wurde). *Empirische Forschung und Erkenntnismöglichkeiten sind also grundsätzlich theoriegeleitet.* Das folgende Schema verdeutlicht die Abhängigkeiten:

Abb. 1: Die Abhängigkeiten zwischen wissenschaftlicher Erkenntnis.



1.2 Kritisch-rationale Wissenschaftstheorie

Aus den allgemeinen, wissenschaftstheoretischen Überlegungen sollte deutlich geworden sein, daß aus der Tatsache, daß Wissenschaftstheorie immer im Vorgriff auf einen noch unbekannten Objektbereich hin entwickelt wird, also unterschiedliche Annahmen und Vorstellungen über diesen Objektbereich möglich und denkbar sind, sich unterschiedliche wissenschaftstheoretische Auffassungen ergeben können (z.B. kritischer Rationalismus, Phänomenologie etc.). Wenn im weiteren eine Beschränkung und Einengung auf die wissenschaftstheoretische Position des kritischen Rationalismus erfolgt, (häufig auch Positivismus, Neopositivismus, analytische Wissenschaftstheorie, Empirismus genannt), so geschieht dies mit der folgenden Begründung:

1. Unabhängig davon, welche wissenschaftstheoretische Position im einzelnen vertreten wird, zeigt sich in der empirischen Forschung, daß im *Begründungszusammenhang* (= jene Phase, in der die empirische Untersuchung zum Zwecke der Überprüfung der theoretischen Hypothesen mit erklärender Absicht durchgeführt wird) immer jene Vorgehensweise (mit nur zum Teil

minimalen Modifikationen) praktiziert wird, wie sie der kritische Rationalismus methodologisch vorschreibt.

2. Der kritische Rationalismus nimmt für sich in Anspruch (wie praktisch keine andere wissenschaftstheoretische Position), die allgemeinste und für alle Objektbereiche gültige Wissenschaftstheorie entwickelt zu haben. Daher kann man die Überlegungen des kritischen Rationalismus sowohl in den Naturwissenschaften wie auch in den Sozialwissenschaften anwenden.

3. Die Position des kritischen Rationalismus ist kritisch und rational, d.h. sie stellt Kriterien zur Verfügung (wie kaum eine andere wissenschaftstheoretische Auffassung), deren intersubjektive Nachprüfbarkeit im konkreten Forschungsprozeß gerade nicht auf Affirmation der theoretischen Überlegungen abstellt, sondern sie strebt eine besonders kritische und auf Widerlegung ausgerichtete Prüfung der Aussagen an.

1.2.1 Die Suche nach Gesetzen

Der kritische Rationalismus geht im Vorgriff auf die unbekannten Objektbereiche davon aus, daß es bestimmte Strukturen im Objektbereich gibt, die in historisch und zeitlich invarianter Weise, immer und überall, anzutreffen sind. Man vermutet also gesetzesartige Beziehungen, deren Auffinden das Interesse des Forschers gewidmet ist. Hier zeigt sich, daß der Wissenschaftstheorie des kritischen Rationalismus das „Muster" der Naturwissenschaften zugrundeliegt: Das Auffinden von Gesetzmäßigkeiten entspricht in der Tat dem Ziel der Wissenschaft, zu erklären, zu prognostizieren und technologische Anweisungen zu geben (s. hierzu 1.2.2).

Diese vermuteten Gesetzmäßigkeiten, die empririsch festgestellt werden sollen, müssen natürlich in Hypothesen, bzw. Theorien gefaßt, d.h. in eine Satzform gebracht werden. Solche typischen Sätze haben in etwa die folgende Form: „Wenn x, dann y" oder „Je größer x, desto kleiner y" usw. Die Allgemeinheit der Formulierung solcher Sätze müßte im Hinblick darauf, was die Gesetzesaussagen eigentlich meinen, spezifiziert werden: Sie sollen nämlich immer und überall gelten; es werden deterministische Beziehungen zwischen den Merkmalen, die in der Gesetzesaussage formuliert sind, vermutet. *Deterministische, ahistorisch immer und überall geltende Gesetzmäßigkeiten bezeichnet man als nomologische Aussagen.*

Kann man in Naturwissenschaften davon ausgehen, daß es eine Fülle solcher nomologischer Aussagen gibt, (wobei der naturwissenschaftliche Laie vermutlich Gesetze annimmt, wo naturwissenschaftliche Experten längst nachgewiesen haben, daß das Attribut „nomologisch" nicht zutrifft), so gilt für die Sozialwissenschaften (und natürlich auch für die Medizin), daß wir es nur in den wenigsten Fällen mit deterministischen Gesetzen zu tun haben. Tatsächlich verlaufen viele theoretisch vermutete Abhängigkeiten zwischen Merkmalen nur probabilistisch, d.h. nach bestimmten Wahrscheinlichkeiten. Probabilistische oder stochastische Aussagen haben etwa die folgende Form: „Wenn x, dann mit einer bestimmten Wahrscheinlichkeit y". *Probabilistische (stochastische, statistische) Aussagen geben nur Wahrscheinlichkeiten für die Beziehung zwischen Variablen an;* diese sind immer kleiner als 1. Bei nomologischen Aussagen, die deterministischen Charakter haben, ist die Wahrscheinlichkeit immer gleich 1. *Daher können deterministisch-nomologische Aussagen als Spezialfall von probabilistischen angesehen werden.* Wenn wir solche stochastischen Aussagen unseren Erklärungen zugrundelegen, haben wir na-

türlich nur statistische Erklärungen und keine nomologisch-gesetzesartigen.

Da wir nun wissen, daß in den Sozialwissenschaften nicht mit nomologischen Aussagen gerechnet werden kann, probabilistische jedoch den Annahmen des kritischen Rationalismus widersprechen, könnte als Konsequenz daraus abgeleitet werden, daß die wissenschaftstheoretische Position des kritischen Rationalismus „weltfremd" oder realitätsfern, damit ad absurdum geführt ist und aufgegeben werden sollte. Daß gleichwohl diese Position beibehalten wird, hat nicht zuletzt pragmatische Gründe: Es zeigt sich nämlich, daß durch probabilistische Aussagen zwar Reduzierungen der Erklärungskräftigkeit hingenommen werden müssen, daß die Logik des kritischen Rationalismus gleichwohl beibehalten werden kann, weil seine Anwendung im konkreten Forschungsprozeß durch die Praxis erkenntnistheoretisch und pragmatisch legitimiert wird.

1.2.2 Die logische Struktur der Erklärung

Da der kritische Rationalismus von nomologischen Aussagen ausgeht, besteht sein Ziel darin, *kausale Erklärungen* zu finden. Sein explizites Ziel ist die Formulierung und empirische Überprüfung von universell gültigen Theorien, die der wissenschaftlich-kausalen Erklärung und Prognose konkreter Sachverhalte dienen, und die bei der Lösung sozialtechnologischer Probleme herangezogen werden können. Der kritische Rationalismus glaubt nun, kausale Erklärungen durch logische Ableitbarkeitsbeziehungen zwischen Gesetzesaussagen und empirischen Bedingungen herstellen zu können. Er verwendet dabei das sog. HEMPEL-OPPENHEIM-Schema deduktiv-nomologischer Erklärung. HEMPEL und OPPENHEIM machen sich den *Syllogismus der formalen Logik zunutze, wo aus zwei vorgegebenen Prämissen ein weiterer Satz abgeleitet wird,* etwa derart:

Alle Menschen sind sterblich
Sokrates ist ein Mensch

Folglich ist Sokrates sterblich

Aus zwei Sätzen wird also ein dritter Satz – formallogischen Kriterien genügend – abgeleitet. Man bezeichnet dabei den ersten Satz als Gesetzesaussage, den zweiten als Anfangsbedingung oder Randbedingung und beide zusammen als Explanans. Der abgeleitete Satz Ereignis oder Explanandum. Dieses deduktiv-nomologische Schema kann verwandt werden, um zu erklären, zu prognostizieren und technologische Anweisungen zu formulieren. (Die Erklärung, daß Sokrates sterblich ist, besteht darin, daß Sokrates ein Mensch ist (Randbedingung) und alle Menschen sterblich sind (Gesetz).)

Abb. 2: Das deduktiv-nomologische Modell

Explanans	Gesetz (Hypothese, Theorie) Anfangsbedingungen (Randbedingungen)
Explanandum	Prüfungshypothese (Ereignisse)

Die These der Strukturidentität von Erklärung, Prognose und technologischer Anweisung besagt, daß das deduktiv-nomologische Schema allen drei Zielen dient. (Der Leser konstruiere sich ein Beispiel und überlege dabei, wie erklärt, prognostiziert und wie technologische Anweisungen gegeben werden können.)

Nach POPPER kann man den Zustand, den die Anfangsbedingungen beschreiben, als Ursache bezeichnen und das Explanandum als Wirkung. *Erklärung besteht also darin, einer bestimmten Wirkung eine bestimmte Ursache logisch und empirisch korrekt zuzuschreiben* (vgl. hierzu 1.4.4).

Logisch korrekt ist eine Ableitung immer nur dann, wenn die als Gesetz oder Hypothese formulierte Aussage im Explanans nomologischen, d.h. deterministischen Charakter besitzt. Ist jedoch eine solche Theorie oder Hypothese nur als probabilistisch geltende Aussage anzusehen, so kann nicht mit Sicherheit das Explanandum abgeleitet werden (Beispiel: Appendix-Operationen verlaufen mit einer Wahrscheinlichkeit von 99 % positiv. Die Person X wird sich einer Blinddarmoperation unterziehen; daraus kann nun nicht abgeleitet werden, daß auch die Operation bei Person X positiv verlaufen wird. Vielmehr kann nur eine bestimmte Wahrscheinlichkeit (nämlich 99 %) für den positiven Verlauf angegeben werden.) Da wir jedoch grundsätzlich nicht mit kausal-nomologischen Aussagen rechnen können, sich offenbar viele Sachverhalte nur stochastisch abspielen und absolute Sicherheit nur in den wenigsten empirischen Fällen gegeben sein wird, können und müssen wir uns wissenschaftstheoretisch mit probabilistischen Aussagen, also auch mit probabilistischen Erklärungen, Prognosen und technologischen Anweisungen zufriedengeben. Zwar haben HEMPEL und OPPENHEIM jene Bedingungen angegeben, unter denen die deduktiv nomologische Erklärung (und mithin wegen der logischen Strukturidentität auch Prognose und technologische Anweisung) richtig sind, doch handelt es sich dabei um idealtypische Bedingungen, die realiter nicht einzuhalten sind, insbesondere was den deterministischen Charakter der Gesetzesaussagen betrifft (Bedingung 2):

1. Das Argument, welches vom Explanans zum Explanandum führt, muß korrekt sein, d.h., es muß eine korrekte, logische Folgerung sein.

2. Das Explanandum muß mindestens ein allgemeines Gesetz enthalten (Nomologieforderung). Daß dieses realiter nicht einzulösen ist, wurde auch von HEMPEL/OPPENHEIM erkannt. Sie sprechen daher, wenn keine nomologischen Aussagen zugrundeliegen, von Erklärungsskizzen statt von Erklärungen.

3. Das Explanans muß empirischen Gehalt besitzen, d.h. es muß sich auf prinzipiell beobachtbare und überprüfbare Phänomene beziehen.

4. Die Sätze, aus denen das Explanans besteht, müssen wahr sein und dürfen sich nicht logisch widersprechen.

Diese Adäquanzbedingungen für Erklärungen gelten relativiert auch für probabilistische Erklärungsversuche. Da man also realiter keine räumlich und zeitlich unabhängigen Gesetzmäßigkeiten (Nomologien) vorfinden wird, ist vorgeschlagen worden, von sog. *Quasigesetzen* zu sprechen, die für historisch-kulturell abgegrenzte Raum-Zeit-Gebiete gültig sind und dort durchaus auch deterministisch ablaufen können. Damit bekommt man das Problem zwar besser in den Griff, kann es jedoch nicht vollständig lösen.

1.2.3 Induktion und Deduktion

Der kritische Rationalismus hat sich durch das deduktiv-nomologische Erklärungsmodell für eine deduktive Vorgehensweise entschieden. Insoweit unterscheidet er sich von den Naturwissenschaften, die induktive Erfahrungen (z.B. durch Versuchsanordnungen und Experimente, die mehrfach wiederholt werden), nicht ausschließen. Diese induktive Vorgehensweise wird jedoch vom kritischen Rationalismus mit der Begründung abgelehnt, daß aus Einzeltatsachen niemals auf allgemeine Gesetze geschlossen werden kann. (Das berühmte Rabenbeispiel verdeutlicht diesen Sachverhalt: Wenn alle Raben, die je beobachtet wurden, schwarz waren, dann schließt man induktiv daraus, daß alle Raben schwarz sind.) Im Gegensatz zu den deduktiven Schlüssen der Mathematik und der formalen Logik, sind solche induktiven Schlüsse immer unsicher, denn solange nicht jeder Einzelfall geprüft ist, muß damit gerechnet werden, daß es Tatsachen gibt, die der allgemeinen, induktiv gewonnenen Gesetzmäßigkeit widersprechen. (So könnte durchaus der Fall eintreten, daß irgendwann wenigstens ein nicht schwarzer Rabe beobachtet wird, die induktiv gewonnene Gesetzmäßigkeit damit falsifiziert wäre.)

1.2.4 Das Falsifikationsprinzip

Mit demselben logischen Argument, das für die Induktion vorgebracht wurde, lehnt der kritische Rationalismus es ab, Hypothesen zu verifizieren. Vielmehr setzt er an die Stelle der Verifikation die Falsifikation. Selbst die erklärungskräftigsten Theorien können niemals als endgültig positiv bewiesen gelten, weil prinzipiell gedacht werden kann, daß im Universum aller Möglichkeiten noch solche Fälle existieren, die der Hypothese oder der Theorie widersprechen, die aber bislang nicht zur Überprüfung herangezogen worden sind. PRIM/TILMANN geben hierfür ein sehr plastisches Beispiel:

„ ‚Immer, wenn Menschen Zyankali essen, sterben sie.' Da diese Aussage für ‚alle Menschen' gelten soll, müßten wir für die Vergangenheit, Gegenwart und die Zukunft an *jedem* Menschen den Nachweis führen, daß er Zyankali gegessen hat und gestorben ist. Bereits aufgrund der Erstreckung des Wahrheitsanspruchs auf die Zukunft werden wir ‚gegenwärtig' das Ziel der Verifikation nicht einmal in Angriff nehmen können. Selbst wenn wir uns auf die Gegenwart beschränken wollten, macht dieses Beispiel besonders deutlich, vor welchen technischen Schwierigkeiten wir stünden: Es wäre nötig, daß die beweisenden Forscher sich in das Beweismaterial einbeziehen. Falls die Verifikation ‚gelingt', kann das Ergebnis aufgrund der Beteiligung aller Menschen weder formuliert, noch mitgeteilt werden." (s. 86)

Somit kann generell gefolgert werden, daß nomologische Aussagen prinzipiell nicht verifiziert werden können, da sie in ihrer Aussage auf unendlich viele Fälle abstellen, realiter jedoch nur eine endliche Menge von verifizierenden Überprüfungen vorgenommen werden kann. Aus diesem logischen Grunde verwirft POPPER das Prinzip der Verifikation als Überprüfungsmöglichkeit für nomologische Aussagen und setzt an deren Stelle das Falsifikationsprinzip. Dabei wird davon ausgegangen, daß bei der Überprüfung von theoretischen Hypothesen an der Realität solche empirischen Fälle gesucht werden müssen, die besonders geeignet sind, deren Scheitern zu ermöglichen. Werden keine solchen Fälle gefunden, wird also die Hypothese nicht durch die Empirie falsifiziert, so geht man davon aus, daß die

Hypothese vorläufig weiterhin Gültigkeit besitzt, bis weitere Überprüfungen eventuell deren Falsifikation herbeiführen. Ist jedoch die Hypothese einmal falsifiziert, an der Realität gescheitert, so ist sie als falsch zu verwerfen.

Das Falsifikationsprinzip setzt demnach an die Stelle der vermeintlichen Gewißheit der Verifikation die schwebende Ungewißheit über den Status der Hypothesen und Theorien. Falsifikation ist also ein negatives Abgrenzungskriterium, das in besonderer Weise auf die Vorläufigkeit und prinzipielle Ungewißheit der Richtigkeit von Aussagen aufmerksam macht.

1.2.5 Die Approximation der Wahrheit

Da in der deduktiven und falsifikatorischen Vorgehensweise der Hypothesenüberprüfung keine endgültige Sicherheit über den Status der Aussagen gewonnen werden kann, sind wir, obwohl wir danach streben, niemals im Besitze der „Wahrheit". Was wir bestenfalls erreichen können, ist eine Aproximation der Wahrheit, d.h. mit zunehmendem Erkenntnisfortschritt werden wir uns der Wahrheit immer stärker annähern, werden sie aber aus rein logischen Gründen nicht voll erreichen können. Dieses allgemeine Argument kann durch empirische Tatsachen gefestigt werden. So stellt man gerade in der Sozialpsychologie häufig fest, daß zu ein und demselben Sachverhalt, sich widersprechende empirische Erkenntnisse vorliegen. Bei strenger Interpretation des Falsifikationsprinzips müßten bei nur einer Widerlegung die zugrundeliegenden Hypothesen verworfen werden. Gegen deren Verwerfung spricht jedoch, daß sie in einer bestimmten Häufigkeit verifiziert worden sind. (In diesem Zusammenhang sei der Einfachheit halber unterstellt, daß sowohl bei Verifikationen und Falsifikationen, keine Fehlerquellen aufgetreten sind und sich die Aussagen auf jeweils gleiche Sachverhalte beziehen.) Da man die vorliegenden Verifikationen nicht einfach vom Tisch wischen und nur die Falsifikationen als gültige empirische Überprüfungen der zugrundeliegenden Hypothesen ansehen kann und vice versa, muß ein Lösungsweg gefunden werden, wie die sich widersprechenden Resultate interpretiert werden können. Hierfür hat POPPER das Konzept des *Bewährungsgrades* entwickelt, das eine Relativierung des Falsifikationsprinzips darstellt. Es sollen jene Theorien als bewährte und damit wahrheitsähnliche gelten, die jeweils unter ceteris-paribus-Bedingungen weniger Falsifikationen und/oder mehr Verifikationen aufweisen. Dabei wird davon ausgegangen, daß jede Hypothese oder Theorie einen bestimmten *Wahrheitsgehalt* enthält (Verifikationen) und einen bestimmten *Falschheitsgehalt* (Falsifikationen). Sinkender Falschheitsgehalt und steigender Wahrheitsgehalt führen zu einer stärkeren *Wahrheitsähnlichkeit* und zu einer besseren Bewährung der Theorie. (Zur Vertiefung sei auf Weiterentwicklungen und Kritiken des kritischen Rationalismus verwiesen, die z.T. erhebliche Differenzierungen am Falsifikationsprinzip, wie es hier basal entwickelt worden ist, anbringen).

1.3 Wissenschaft und Sprache

„Der Wissenschaftler hat es nicht mit irgendeiner Wirklichkeit ‚an sich' zu tun, sondern mit einer mehr oder weniger absichtsvoll, durch Begriffe vorstrukturierten Erfahrungswelt. Er erlebt seinen Gegenstand nicht unmittelbar und unreflektiert, sondern nimmt ihn, indem er ihn benennt und damit begrifflich ordnet, bewußt und distanziert wahr. Das löst ihn gleichzeitig aus dem Zwang unmittelbar reflexartigen Reagierens auf Umweltreize und gibt ihm die Freiheit zum Denken. Sofern Menschen über Sprache und damit über Begriffe verfügen, gilt dies natürlich nicht nur für den Wissenschaftler. Aber für den wissenschaftlichen Erkenntnisprozeß ist diese begriffliche Vermittlung zwischen Subjekt und Objekt der Erfahrung conditio sine qua non". (MAYNTZ, R., u.a., Einführung in die Methoden der empirischen Soziologie, Köln und Opladen 1969, S. 9).

Sprache strukturiert das menschliche Denken und ist andererseits Vehikel, Gedanken mitteilbar zu machen. Gerade die Wissenschaft ist darauf angewiesen, in dreifacher Weise die Sprache zu verwerten:

1. Wissenschaftlich interessierende Tatsachen begegnen uns nicht als solche, sondern sind durch Sprache vermittelt. So bedient man sich in der Wissenschaftsentwicklung zunächst durchaus der Alltagssprache, in der Absicht, bestimmte Phänomene in den Griff zu bekommen, um sie dann immer stärker zu präzisieren, zu denaturieren und zu objektivieren, um intersubjektiv nachprüfbare Aussagen zu erhalten.

2. Wissenschaftliche Aussagen sind sprachliche Manifestationen, die nur durch die Sprache selbst verständlich und mitteilbar werden. Gerade wissenschaftliche Aussagen sind darauf angewiesen, eine breite Öffentlichkeit und Mitteilungswirkung zu erzielen, denn schließlich sollen wissenschaftliche Erkenntnisse auch pragmatisch-praktisch genutzt werden. (Die Flut von Publikationen verweist auf diesen Gesichtspunkt.)

3. Gerade bei der empirischen Überprüfung theoretischer Aussagen, die ja sprachlich geronnen sind, ergibt sich die Notwendigkeit, die in den Aussagen verwendeten Begriffe klar und eindeutig bestimmten empirischen Sachverhalten zuzuordnen. (Vgl. die deduktiv-nomologische Erklärung, in der sowohl in den allgemeinen Gesetzesaussagen, wie auch in den Randbedingungen und dem Explanandum Begriffe erscheinen.)

Die wissenschaftstheoretische und methodologische Beschäftigung mit den Beziehungen zwischen Wissenschaft und Sprache schafft mittels der daraus gewonnenen Erkenntnisse die notwendigen Vorbedingungen für korrektes, wissenschaftliches Arbeiten. Dieses besteht darin:

1. Mittels präziser Sprachformulierungen, die theoretisch erfaßten Sachverhalte empirisch nachprüfbar zu machen (Theorie-Realität).

2. Eine kritische (auch theoretische) Kommunikation zwischen den Wissenschaftlern herbeizuführen „ohne aneinander vorbeizureden".

Während die alltagssprachlichen Begriffe, die auch in der Wissenschaft verwandt werden, durch den alltagssprachlichen Umgang im Regelfall für alle verständlich sind, (gemeinsamer Zeichenvorrat = Idiolekt), *sind wissenschaftliche Begriffe häufig Spezifizierungen oder Abstraktionen der Alltagssprache, sodaß ihnen spezifische Konnotationen anhaften, die zum Zwecke des Verstehens gesondert mitgeteilt werden müssen. Hierzu dient die Definition.*

1.3.1 Begriffe und Beobachtung

Die in den Theorien bzw. Hypothesen als sprachliche Fassungen der Merkmale bzw. Variablen aufscheinenden Begriffe (vgl. Abschnitt 1.3.2) sollen als theoretische einer empirischen Überprüfung zugeführt werden. Jede empirische Überprüfung setzt jedoch voraus, daß die Beobachtbarkeit der Variablen gewährleistet ist, weil Beobachtung jene grundlegende Methode ist, die die Konfrontation des Theoretischen mit dem Erfahrbaren ermöglicht. Die in den Theorien verwendeten Begriffe weisen jedoch unterschiedliche Grade der Beobachtbarkeit (empirische Feststellbarkeit) auf.

So sind *Beobachtungsbegriffe* (observational terms) aufgrund relativ einfacher, unmittelbarer und direkter Beobachtung anwendbar. Sie werden auch als empirisch, phänotypisch oder deskripitiv bezeichnet. Die unmittelbare Beobachtbarkeit der Phänomene, die mit den Begriffen bezeichnet werden, erleichtert natürlich deren empirische Erhebung immens. Direkte Beobachtungsbegriffe sind jedoch selten.

Weitaus häufiger tauchen die sog. *indirekten Beobachtungsbegriffe* auf (indirect observables). Hier handelt es sich um solche Begriffe, die aufgrund von Beobachtung angewendet werden, wo aber zu den Beobachtungen legitime Schlußfolgerungen hinzutreten. Da diese Begriffe nicht unmittelbar beobachtbar sind, müssen normalerweise aufgrund von theoretischen oder plausiblen Annahmen Indikatoren konstruiert werden, die diese zu erheben gestatten.

Noch abstrakter sind die *Konstrukte*. Hierbei handelt es sich um solche Begriffe, die weder aufgrund direkter noch indirekter Beobachtung angewendet werden können, die aber doch mit Hilfe von Beobachtung definiert werden. Hypothetische Konstrukte sind z.B. Intelligenz, Integration, Mobilität etc. Ein Konstrukt ist im Regelfalle auch so komplex (vgl. Intelligenz), daß es nicht mit einfachen Indikatoren erhebbar erscheint, sondern daß normalerweise eine Fülle von Indikatoren und Relationen zwischen diesen herangezogen werden muß, um es empirisch zu bestimmen.

Theoretische Begriffe beziehen sich auf Zusammenhänge zwischen einzelnen Variablen, liegen also noch eine Stufe abstrakter als die Konstrukte.

Von den Beobachtungsbegriffen bis zu den theoretischen Begriffen lassen sich unterschiedliche Grade der Beobachtbarkeit der Variablen oder Merkmale feststellen. Je mehr wir uns von den Beobachtungsbegriffen entfernen, desto stärker problembehaftet wird die Beobachtung der mit den Begriffen gemeinten Sachverhalte in der Realität, weil nie garantiert werden kann, daß das, was empirisch beobachtet wird, völlig deckungsgleich mit dem ist, was theoretisch mit den Begriffen gemeint ist. Daraus jedoch folgern zu wollen, daß nur direkte Beobachtungsbegriffe als wissenschaftliche Begriffe zugelassen werden sollten, wäre verfehlt. Man würde sich damit jegliche theoretische Entwicklungs- und Entfaltungsmöglichkeit nehmen und einen Erkenntnisverzicht hinnehmen, der nicht legitimierbar erscheint. Der *klassische Behaviorismus* wird in diesem Zusammenhang insbesondere deswegen

kritisiert, weil er sich in der Tat auf unmittelbar Beobachtbares beschränkt hat, was den Erkenntnisspielraum erheblich reduzierte. Vielmehr muß aus der abnehmenden unmittelbaren Beobachtbarkeit zwischen Beobachtungsbegriffen und theoretischen Begriffen abgeleitet werden, daß besondere methodische Sorgfalt dazu verwandt werden soll, die weniger gut beobachtbaren Begriffe empirisch zu erfassen.

Jede Beobachtung ist, da sie theoriegeleitet ist, nicht als solche zu interpretieren, sondern im Lichte von Theorien zu sehen. Da Beobachtungen selbst sich nicht unmittelbar auf Zusammenhänge beziehen, sondern immer nur einzelne Tatbestände erheben, bedürfen die Beobachtungen einer Interpretation, die sich einmal auf die Beobachtung selbst, zum anderen auf die theoretisch-hypothetisch vermuteten Relationen zwischen den Beobachtungen bezieht. Bei dieser Interpretation nun ist in besonderer Weise zu berücksichtigen, daß die unmittelbare, auf Beobachtung beruhende Interpretation eigentlich im Deskriptiven steckenbleibt. Daher ist es notwendig, die Beobachtungsresultate in den theoretischen Zusammenhang zu stellen, wie er vor der Untersuchung hypothetisch formuliert wurde. Hierzu gehört auch, daß die beobachteten Einzeltatsachen, die in theoretisch-hypothetisch festgelegter Weise zusammengefügt zu theoretischen Konstrukten sich ausweiten und qualifizieren, durch das Raster dieser Konstrukte interpretiert werden. Daher erscheint es legitim, ja geradezu notwendig, daß mit Hilfe solcher Konstrukte bestimmte beobachtete Variablen erklärt werden. So können z.B. bestimmte Verhaltensweisen durch das theoretische Konstrukt der Persönlichkeitsstruktur erklärt werden, wie auch bestimmte Testleistungen auf das hypothetische Konstrukt der Intelligenz zurückgeführt werden können. Die Ablehnung solcher, nicht unmittelbar beobachteter Tatsachen, wäre nur zu rechtfertigen, wenn es solche Sachverhalte gäbe, die dasselbe Erklärungspotential wie die hypothetischen Konstrukte besitzen, jedoch zusätzlich die Qualität der unmittelbaren Beobachtbarkeit aufweisen würden. Da dies nicht angenommen werden kann, müssen auch hypothetische Konstrukte als Erklärungsmöglichkeit herangezogen werden. Es bleibt aber immer zu bedenken, daß die empirische Erfassung dieser Konstrukte stärker problem- und fehlerbehaftet ist, als die Erhebung unmittelbar beobachtbarer Sachverhalte.

1.3.2 Die Variabilität der Merkmale

Jede Wissenschaft stellt darauf ab, Hypothesen zu überprüfen. In Hypothesen werden Merkmale miteinander in Beziehung gesetzt. Will man sie überprüfen, muß man die in ihnen genannten Merkmale beobachten und intersubjektiv nachprüfbar messen können. Dies setzt wiederum voraus, daß die Merkmale, die in den Hypothesen in begrifflicher Form und in unterschiedlichen Ausprägungen vorliegen, empirisch erfaßt werden können. Würde man unterstellen, daß ein Merkmal immer nur in gleicher Weise auftritt, so wäre dieses Merkmal für die wissenschaftliche Analyse völlig uninteressant, denn alle Hypothesen müßten sich in gleicher Weise auf das Merkmal beziehen. (So ist z.B. bei ungeschlechtlichen oder bigeschlechtlichen Lebewesen eine Differenzierung des Merkmals „Geschlecht" nach

männlich oder weiblich sinnlos, weil beide Ausprägungen in gleicher Weise zutreffen, sodaß das Geschlecht offensichtlich keinen Einfluß auf die Variation anderer Merkmale ausüben kann). Damit Merkmale für die Wissenschaft interessant werden, müssen sie variabel sein, ihre Ausprägungen also in mindestens zwei Formen vorliegen (dichotomisierte Merkmale). Die Zahl der Merkmalsausprägungen kann aber auch beliebig groß sein (Körpergröße, Einkommen etc.). *Da die interessierenden Merkmale variabel sind, spricht man auch von Variablen. Die Merkmalsausprägungen bezeichnet man – technisch gesehen – als Werte der Variablen.*

Man differenziert nun die *Variablen nach unabhängigen und abhängigen Variablen, je nach dem, welche Stellung sie in den Hypothesen einnehmen und welche Wirkungsrichtung ihnen hypothetisch zugeschrieben wird.* Werden die Hypothesen in „wenn ..., dann ..."-Form geschrieben, sind jene Variablen, die in der Wennkomponente der Hypothese aufscheinen, die unabhängigen und jene, welche in der Dannkomponente genannt sind, die abhängigen Variablen. Im Falle von „je ..., desto ..."-Aussagen gilt der gleiche Sachverhalt: Die unabhängigen Varialben stehen im ersten Halbsatz, die abhängigen im zweiten. Theoretisch wird dabei immer angenommen, daß die unabhängigen Variablen, (die selbst in anderem Zusammenhang durchaus als abhängige gesehen werden können), in dem Kontext der Hypothese Wirkungen in anderen Variablen erzeugen, diese Variablen determinieren.

Zur Verdeutlichung sei noch einmal rekapituliert: *Untersuchungsgegenstand oder sog. Beobachtungseinheiten sind die Merkmalsträger. Solche Merkmalsträger weisen bestimmte Merkmale in unterschiedlichen Merkmalsausprägungen auf.* Das empirische Interesse richtet sich auf die Erfassung, das Messen solcher Merkmalsausprägungen. In etwas abstrakterer und technischer Formulierung lautet derselbe Sachverhalt: *Bestimmte Fälle werden im Hinblick auf die Werte bestimmter Variablen untersucht.*

Die in den Hypothesen der Theorien aufscheinenden theoretischen Begriffe können sich sowohl auf Beobachtungseinheiten, auf Merkmale, wie auch auf Merkmalsausprägungen beziehen. In jedem Falle legen sie fest, an welchen Fällen welche Variablen mit welchen Werten zu beobachten sind. Von daher ist deutlich, daß dann, wenn die Begriffe unscharf und ungenau definiert worden sind, erhebliche meßtheoretische Probleme der Gültigkeit auftreten müssen. Notwendige Absicht wissenschaftlich-empirischer Untersuchungen muß es daher sein, die Begriffe so zu definieren, daß bei deren empirischer Erhebung keinerlei Zweifel über die Richtigkeit der Umsetzung der theoretischen Begriffe in solche, die der Beobachtung zugänglich gemacht werden können, auftreten sollten. Dies gelingt natürlich umso mehr, je eher es sich um direkt beobachtbare Begriffe handelt.

1.3.3 Indikatorisierung und Operationalisierung

Wir kehren zu dem Problem, das eingangs schon einmal angedeutet wurde, zurück, nämlich inwieweit in einem deduktiv-nomologischen Erklärungsschema, das im Hinblick auf seine Gültigkeit und Wahrheit empirisch überprüft werden soll, zwi-

schen empirischer Überprüfung und logischer Erklärung eine Kompatibilitäts- oder Adäquanzbeziehung besteht. In der deduktiv-nomologischen Erklärung und deren empirischer Überprüfung stellten die in den Aussagen verwandten Begriffe ein entscheidendes Bindeglied dar. Hieraus ergeben sich zwei wesentliche Schwierigkeiten: Die eine bezieht sich darauf, inwieweit die Begrifflichkeit an die Erfahrbarkeit herangeführt werden kann, d.h., wie man die theoretischen Begriffe möglichst nahe an die von ihnen gemeinte Erscheinung heranrücken könnte. Eine zweite Schwierigkeit ergibt sich aus der Beziehung zwischen Theorie und Begriff, denn es besteht nicht a priori ein Kompatibilitätsverhältnis zwischen beiden.

Die von rigiden Sozialforschern vorgeschlagene Lösung des ersten Problems besteht darin, die Bedeutung des Begriffs für identisch zu erklären, mit den spezifischen Operationen, die die gemeinte Erscheinung messen (fatales Beispiel: Intelligenz ist, was der Intelligenzquotient mißt, oder die Definition von Lohn und Strafe in der Lerntheorie). Für die empirische Forschung genügt es nicht, die zentralen Begriffe zu definieren, sondern es müssen darüber hinaus präzise Anweisungen für Forschungsoperationen gegeben werden, mit deren Hilfe entscheidbar ist, ob ein mit dem betreffenden Begriff bezeichnetes Phänomen vorliegt oder nicht. *Diese Festlegung von Forschungsoperationen zum Zwecke der empirischen Erfassung eines Begriffs nennt man die operationelle Definition oder Operationalisierung.* Die operationale Definition ist ein in der empirischen Forschung für jeden Begriff notwendiger Übersetzungsvorgang in Forschungsoperationen. Anders formuliert: Die Operationalisierung eines Begriffes erfolgt durch Einbeziehung des Beobachtungsverfahrens in seine Definition.

Die Operationalisierungsmöglichkeit eines Begriffes setzt voraus, daß dieser Begriff einen empirischen Bezug aufweist. Wie aus der Klassifikation von theoretischen Begriffen und Beobachtungsbegriffen bereits ableitbar war, kann dieser empirische Bezug sehr unterschiedlich, nämlich direkt oder aber indirekt geartet sein. Der empirische Bezug eines Begriffes ist direkt, wenn das vom Begriff bezeichnete Phänomen unmittelbar beobachtet werden kann. Auch bei direktem empirischen Bezug eines Begriffes bedarf dieser der operationalen Definition. So erfordert z.B. die „theoretische" Variable Bibliotheksbesucher, deren Transformation in Forschungsoperationen, indem angegeben werden muß, wann, wo und wie welche Personen (wer als Bibliotheksbesucher gelten kann) gezählt werden sollen.

Viele soziale Phänomene, die theoretisch gefaßt sind, können nicht unmittelbar beobachtet werden. Aus solchen Begriffen geht normalerweise auch nicht hervor, wie man die mit den Begriffen angesprochenen Phänomene empirisch erfassen kann. So ist Intelligenz ein eindeutig nur indirekt empirisch feststellbarer Sachverhalt. Bei solchen Begriffen mit indirektem empirischen Bezug, kommt es darauf an, Indikatoren für das nicht unmittelbar empirisch beobachtbare Phänomen angeben zu können. Auch hier gilt wieder, daß die Benennung von Indikatoren selbst, noch keine Operationalisierung darstellt. Hierzu müssen die ausgewählten Indikatoren selbst in Forschungsoperationen transformiert werden. *„Unter Indikatoren sind direkt wahrnehmbare Phänomene ('Ersatzgrößen', 'Stellvertreter') zu ver-*

stehen, mit deren Hilfe man begründet auf das Vorliegen des nicht unmittelbar wahrnehmbaren Phänomens schließen zu dürfen glaubt." (PRIM/TILMANN, Grundlagen einer kritisch rationalen Sozialwissenschaft, Heidelberg 1975, S. 55).

PRIM und TILMANN zitieren an derselben Stelle MAGER, der für den theoretischen Begriff des Musikverständnisses – eher persiflierend als ernst gemeint – die folgenden Indikatoren vorschlägt:

„1. Der Lernende seufzt ekstatisch, wenn er Bach hört
2. Der Lernende kauft Hi-Fi-Einrichtung und Schallplatten im Wert von 500 S
3. Der Lernende beantwortet 95 Auswahlantwortfragen zur Musikgeschichte richtig
4. Der Lernende schreibt einen flüssigen Aufsatz über die Bedeutung von 37 Opern
5. Der Lernende sagt: Mann, glaube mir, ich bin Fachmann, es ist einfach großartig."

Für den doch etwas vagen Begriff der Feindschaft werden folgende Indikatoren diskutiert:

„1. A gibt B Ohrfeigen
2. A sagt, ich habe einen Haß auf B
3. A beleidigt B bei einem Wortwechsel
4. A lehnt eine von B erbetene Hilfeleistung ab."

Aus beiden Beispielen läßt sich unschwer ableiten, daß es keine eindeutigen Indikatoren für die Feststellung theoretischer Sachverhalte in der empirischen Realität gibt. Gleichwohl sind wir in der empirischen Forschung darauf angewiesen, Indikatorisierungen und Operationalisierungen vorzunehmen, weil ohne diese, die empirische Überprüfung theoretischer Aussagen nicht eingelöst werden kann.

Wenn in der Literatur gefordert wird, daß empirischer Bezug und Operationalisierbarkeit für theoretische Begriffe erfüllt sein müssen, damit eine intersubjektive Verständigung zwischen den Wissenschaftlern möglich erscheint – auch weil sie für die empirische Überprüfung theoretischer Aussagen unverzichtbar sind – so ist andererseits doch ein konsequenter Operationalismus für die Verständigung nicht gerade förderlich, denn jede, noch so geringfügige Veränderung der Meßoperation führt ja strenggenommen zu einem neuen Begriff. Kritiker des Operationalismus wenden daher auch ein, daß eine strenge, eindeutige und strukturidentische Transformation des theoretischen Begriffes in empirischen Forschungsoperationen nicht möglich ist (vgl. hierzu auch 1.5), weil der Forscher immer zu bestimmten Abstraktionen gezwungen wird, die nicht in die Messung eingehen. Im übrigen bestimme auch ein gewisses Vorverständnis von dem zu erhebenden Phänomen Indikatorisierung und Operationalisierung, sodaß die Meßoperation nicht in jedem Falle gültig ist.

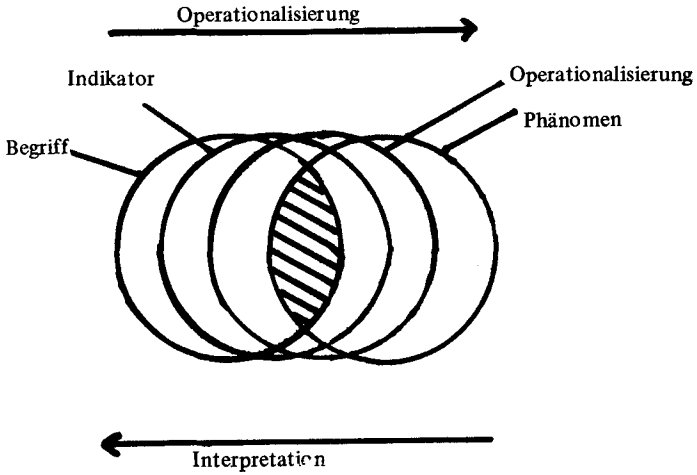
1.3.4 Gültigkeit der Operationalisierung

Ein erheblicher Kritikpunkt an der Operationalisierung betrifft die Beziehung zwischen Begriff und Theorie. Hier wird behauptet, daß die Operationalisierung die wechselseitige Befruchtung von Erfahrung und Theorie unterdrücke. Akzeptiere man nämlich, daß theoretische Begriffe in empirischer Forschung operationalisiert

werden müßten, so würden die theoretischen Überlegungen im Grunde genommen durch Antizipation des empirisch Möglichen auf diese selbst beschränkt, weil die Operationalisierungen schon immer mitgedacht werden. Während anerkannt wird, daß es durch die Operationalisierung gelingt, die Begriffe den Phänomenen kompatibel zu machen, werde der umgekehrte Weg, nämlich der der Abstraktion von der Empirie zur Theorie praktisch verunmöglicht.

Ist der obige Kritikpunkt etwas überspitzt formuliert, so ist er in der Tendenz unbestreitbar. Ebenso unbestreitbar ist, daß es mit den Operationalisierungen nie vollständig gelingen wird, einen theoretischen Begriff mit dem empirischen Phänomen, das dieser bezeichnet, deckungsgleich zu machen. Wenn wir z.B. von einem nur indirekt empirisch feststellbaren Begriff ausgehen, so muß dieser indikatorisiert und operationalisiert werden und in der Operationalisierung das gemeinte Phänomen erfassen. Bezeichnet man Begriff, Indikator, Operationalisierung und Realität jeweils mit einem Kreis, so müßten in der konkreten empirischen Forschung diese vier Kreise „verschwimmen"; sie müßten vollständig deckungsgleich sein. Tatsächlich jedoch, und darauf deutet die untenstehende Abbildung hin, wird es nur gelingen, letztendlich einen Teil der Realität (des gemeinten Phänomens) zu erfassen.

Abb. 3: Gültigkeitsprobleme der Operationalisierung



Vollständige Deckungsgleichheit wäre theoretisch und idealiter zu fordern, damit theoretisch wie empirisch nur das bezeichnete Phänomen und dieses vollständig erfaßt wird. Überschneiden sich die obigen Kreise überhaupt nicht so wird empirisch alles Mögliche erhoben, nur nicht der theoretisch gemeinte Sachverhalt. Beide Extremfälle werden relativ selten sein. Mischformen, wo das theoretisch Gemeinte auch teilweise empirisch erfaßt wird, dürfen die Regel darstellen.

Da man bei der Interpretation der gewonnenen empirischen Resultate eigentlich den rückwärtigen Weg der Operationalisierung zu gehen hat, nämlich das empirisch gewonnene auf die Theorien und Hypothesen zu beziehen, treten aus mangelnder Kongruenz erhebliche Fehlerquellen bei der Interpretation auf. Deswegen ist zu fordern, daß gültige Indikatorisierungen und Operationalisierungen anzustreben sind, auch wenn sie realiter nicht voll erreicht werden können.

1.4 Theorien und Hypothesen

Eine Erklärung bestand bei der deduktiv-nomologischen Ableitung darin, daß aus einer Gesetzesaussage und den Randbedingungen ein Explanandum abgeleitet wurde. Die Gesetze können auch als Theorien bzw. Hypothesen gefaßt werden. *Im kritischen Rationalismus werden Theorien als ein aus Hypothesen zusammengesetztes, durch Ableitbarkeitsbeziehungen verknüpftes, komplexes Netzwerk von nomologischen Aussagen verstanden; Hypothesen sind demnach Elemente von Theorien.* Da Theorien und Hypothesen die erkenntnisleitenden und die empirisch strukturierenden Elemente der Forschung sind, erscheint es notwendig und angebracht, sich mit ihnen und den sie betreffenden methodologischen Gesichtspunkten ein wenig zu beschäftigen.

1.4.1 Deskriptive und relationale Hypothesen

Jeder empirischen Untersuchung gehen Hypothesen voraus, die in unterschiedlichster Form formuliert sein können. Je nach Erkenntnisabsicht und Erkenntnismöglichkeiten unterscheidet man deskriptive von relationalen Hypothesen.

Deskriptive Hypothesen sind zwar nicht erklärungskräftig (genügen also mithin nicht einem der wesentlichsten Ziele der Forschung), sie sind aber doch geeignet, die wissenschaftliche Forschung zu befruchten. *Deskriptive Hypothesen behaupten retrospektiv, gegenwärtig oder prospektiv einen Sachverhalt, der realiter zutreffen soll.* So kann als Hypothese aufgestellt werden, daß 5 % der Frauen in der Bundesrepublik schon mindestens einmal eine Abtreibung an sich haben vornehmen lassen. Diese Hypothese ist, obwohl rein beschreibend, keineswegs unsinnig, weil sie eine Aussage macht, die empirisch erst festgestellt werden muß. Dem einzelnen Beobachter ist es sicher nicht möglich, den in der Hypothese vermuteten Sachverhalt unmittelbar festzustellen. Solche deskriptiven Hypothesen sind also in der Lage (sofern sie empirisch eingelöst werden können), die Umwelt zu strukturieren und auch für den einzelnen erfahrbar zu machen. Sie sind im Regelfalle jedoch nur Vorstufe für solche Hypothesen, die verschiedene Sachverhalte miteinander in Beziehung setzen und aus diesem Beziehungsverhältnis erklärungskräftige Rückschlüsse gewinnen wollen.

So spricht man von Hypothesen häufig erst dann, wenn sie relational gemeint sind, d.h., wenn es sich um solche Aussagen handelt, die mindestens zwischen zwei

(natürlich auch zwischen mehreren) Merkmalen eine Beziehung herzustellen versuchen. In der einfachsten Form geht man davon aus, daß zwei Variablen zueinander in Beziehung gesetzt werden. Solche *bivariate Hypothesen* erweisen sich jedoch zumeist als ungenügend, als zu grob. Im Anfangsstadium der Forschung beginnt man zwar damit, einen Zusammenhang zwischen zwei Phänomenen festzustellen, doch bald merkt man, daß dieser Zusammenhang nicht unbedingt gilt, d.h., daß andere zusätzliche Bedingungen mitzuberücksichtigen sind. Mit zunehmendem Erkenntnisfortschritt gelangt man daher zur Entwicklung von *multivariaten Hypothesen.*

Wenn man weiß oder annimmt, in welcher *Richtung die Beziehung zwischen den in den Hypothesen genannten Phänomenen verläuft, dann kann man das eine Merkmal als unabhängige Variable (Ursache) und die andere als abhängige Variable (Wirkung) bezeichnen. Die unabhängige Variable wird auch als Determinante und die abhängige als Resultante bezeichnet.*

1.4.2 Die Hypothesenstruktur

Determinanten und Resultanten können in den Hypothesen in unterschiedlicher Weise miteinander verknüpft werden. Während die allgemeine Formulierung von Hypothesen als „wenn x, dann y" oder „je größer x, desto größer y" die Verknüpfung zwischen den in den Hypothesen genannten Variablen realtiv unspezifisch beläßt, weist ZETTERBERG mit Recht darauf hin, daß verschiedene logische Verknüpfungen denkbar sind (vgl. ZETTERBERG, H., On Theory and Verification in Sociology, Totawa N.J. 1968). Er stellt insbesondere 5 Alternativen von Verknüpfungen zwischen Determinanten und Resultanten (Ursachen und Wirkungen) vor, aus denen sich unterschiedliche theoretische, wie auch praktische Konsequenzen ableiten lassen.

1. Reversibel versus irreversibel
   Wenn x, dann y und wenn y, dann x (reversibel)
   Wenn x, dann y aber wenn y, dann nicht x (irreversibel)
2. Deterministisch versus stochastisch
   Wenn x, dann immer y (deterministisch)
   Wenn x, dann wahrscheinlich y (stochastisch oder probabilistisch)
3. Aufeinanderfolgend versus koexistent
   Wenn x, dann später y (aufeinanderfolgend)
   Wenn x, dann gleichzeitig x (koexistent)
4. Hinreichend versus bedingt
   Wenn x, dann y ungeachtet alles sonstigen (hinreichend)
   Wenn x, dann y aber nur dann, wenn auch z (bedingt)
5. Notwendig versus substituierbar
   Wenn x, dann und nur dann y (notwendig)
   Wenn x, dann y aber wenn z, dann auch y (substituierbar).

Diese idealtypisch herausgestellten Verknüpfungen von Determinanten und Resultanten in Hypothesen können natürlich untereinander kombiniert werden: So ergeben sich z.B. *interdependente* Beziehungen als reversible, aufeinanderfolgende und bedingte Verknüpfungen der Variablen. *Funktionale* Beziehungen sind reversibel und koexistent, während *kausale* Beziehungen irreversibel, deterministisch aufeinanderfolgend, notwendig und hinreichend sind.

Je nach dem, in welcher Weise die Hypothesen formuliert sind, ergeben sich unterschiedliche Möglichkeiten der Überprüfung, bzw. unterschiedliche Grade der Feststellbarkeit von deren Richtigkeit. So kann eine deterministische Hypothese durch einen einzigen empirischen Fall, der ihr widerspricht, widerlegt werden, während dies bei einer stochastischen Hypothese nicht möglich ist (vgl. hierzu 1.4.4). Von daher ist es grundsätzlich notwendig, seine Hypothesen so spezifisch wie möglich zu formulieren, um klar entscheiden zu können, was sie meinen und in welchen Fällen sie als akzeptiert oder widerlegt zu gelten haben. Die Struktur der Hypothesen zu erkennen heißt, Fehlerquellen bei der empirischen Überprüfung zu vermeiden.

1.4.3 Hypothesenprüfungen

Hypothesen sind zunächst nichts anderes als aus wissenschaftlicher Literatur, aus Primärerfahrung etc. gebildete Vor-Urteile, die ein Forscher an eine bestimmte Fragestellung heranträgt. Die nächste Frage ist daher, wie kann man feststellen, ob diese Vor- oder Vorwegurteile auch Urteile, d.h. gültige Aussagen darstellen. Wie also kann man unterschiedliche theoretische Aussagen – möglichst exakt in Hypothesen gefaßt – überprüfen?

Da Hypothesen im Regelfall niemals isoliert stehen, sondern eingebettet sind in einen theoretisch übergreifenden Rahmen, überprüft man zunächst, ob die Hypothese mit diesem theoretischen Rahmen kompatibel ist. Der in der Hypothese formulierte Satz wird einer *logischen Überprüfung* dahingehend unterzogen, ob er den Regeln der formalen Logik genügt, ob er in sich widerspruchsfrei ist, ob er aus übergeordneten Sätzen richtig abgeleitet ist etc. Gerade wenn man davon ausgeht, daß eine Theorie ein deduktiv axiomatisches System von Hypothesen darstellt, kommt diesem Schritt der Überprüfung mitentscheidende Bedeutung zu. Zur logischen Überprüfung gehört auch die Frage, ob eine Hypothese eventuell nicht tautologisch formuliert ist, denn Tautologien entziehen sich grundsätzlich einer empirischen Überprüfung. Dies impliziert, daß jede Hypothese daraufhin überprüft werden muß, ob sie empirischen Charakter aufweist, ob sie auf dem Erfahrungswege kontrolliert werden kann.

Neben der logischen Überprüfung ist eine *inhaltliche Überprüfung der Hypothesen* erforderlich. Sie dient u.a. dazu, festzustellen, ob die zu prüfende Hypothese mit schon vorliegenden Hypothesen und Theorien kompatibel ist, oder diesen widerspricht. Ist diese Hypothese bereits in anderen Theorien enthalten, oder macht sie grundsätzlich neue Aussagen? Je nachdem wie diese Fragen zu beantworten sind,

ist mit der Hypothese auf unterschiedliche Weise zu verfahren. Auf alle Fälle ist eine solche inhaltliche Überprüfung schon deswegen erforderlich, um nicht dem Irrtum zu verfallen, man habe das „Ei des Kolumbus" gefunden, während dieselbe Hypothese schon längst von anderen Forschern aufgestellt und empirisch überprüft wurde. Da Hypothesen einen Teil wissenschaftlicher Arbeit darstellen, und diese dazu dient, wissenschaftlichen Fortschritt herbeizuführen, können Hypothesen auch inhaltlich daraufhin überprüft werden, ob sie diesem Ziele dienen.

Auf die inhaltliche und logische Überprüfung der Hypothesen als theoretische Arbeit folgt eine zweite Stufe, nämlich die Konfrontation der Hypothese mit dem von ihr bezeichneten Phänomen, d.h. die Frage, inwieweit die Aussage der Hypothese mit den beobachteten Verhältnissen der sozialen Realität übereinstimmen. Dieser *empirischen Überprüfung* kommt entscheidende Bedeutung bei, weil von ihr abhängig gemacht wird, ob eine Hypothese (unabhängig von der logischen und inhaltlichen Überprüfung, die hier als positiv vorausgesetzt werden), als richtig oder als falsch beurteilt wird. Genau diese Fragestellung wird mit Hilfe der einzelwissenschaftlichen Forschungsmethoden unter Zuhilfenahme statistischer Verfahren beantwortet (vgl. 1.6.2 und 1.6.3). Die empirische Hypothesenüberprüfung ist, obgleich ihr höchste Priorität in der wissenschaftlichen Forschung zukommt, nur ein dritter Schritt, der auf die logische und inhaltliche Prüfung folgt.

1.4.4 Probleme der Hypothesenüberprüfung

In diesem Abschnitt können nicht alle methodologischen Probleme der Hypothesenüberprüfung behandelt werden, weil diese den wissenschaftstheoretischen Laien eher überfordern würden als es ihm förderlich wäre. Daher sollen nur die wichtigsten Probleme herausgegriffen und paradigmatisch vorgestellt werden.

Wir haben bereits gesehen, daß der kritische Rationalismus verlangt, daß Hypothesen in deterministischer Form, d.h., als *raum-zeitlich unabhängige Allaussagen* vorliegen sollten. Es wurde ebenfalls schon gezeigt, daß solche Allaussagen grundsätzlich nicht verifiziert werden können, daß sie jedoch prinzipiell falsifizierbar sind. Diese Prämisse der Allaussage trifft jedoch nicht alle wissenschaftlich formulierten Hypothesen. So kann es durchaus sein, daß Hypothesen nicht in Form von Allaussagen, sondern als *singuläre Sätze* formuliert werden, als sog. *Existenzaussagen.* Die Existenzaussage: „Es gibt schwarze Raben" kann sicherlich verifiziert aber nicht grundsätzlich falsifiziert werden, während die Allaussage: „Alle Raben sind schwarz" grundsätzlich nicht verifizierbar, aber jederzeit falsifizierbar erscheint. *Je nachdem also, ob es sich um Allaussagen oder um Existenzaussagen handelt, erscheint eine unterschiedliche Logik in der Hypothesenüberprüfung angebracht zu sein.*

Erfuhr das Falsifikationsprinzip eine Relativierung durch die Form der Hypothesen (Allaussagen versus Existenzaussagen und wie im Kap. 1.6.2 und 1.6.3 noch zu zeigen sein wird durch deterministische oder stochastische Formulierung), so ist weiter zu problematisieren, daß bei der Hypothesenprüfung mit der Anwen-

dung des Falsifikationsprinzips der Erfahrung Priorität vor der Theorie zukommt; denn die Theorie wird allemal durch konträre Beobachtung widerlegt. Durch die Überlegung zur Operationalisierung der Begriffe, die in den Hypothesen aufscheinen, wissen wir jedoch, daß diese nicht notwendigerweise von ihrem gemeinten Inhalt her mit der Realität, wie sie durch die Beobachtung erfaßt wird, übereinstimmen. Mit anderen Worten, *empirische Beobachtungen können durchaus einer Theorie widersprechen, ohne daß die Realität der Theorie entgegenläuft.* Ist nämlich die Operationalisierung (man spricht auch von der Beobachtungstheorie) nicht adäquat gelungen, so kann allein dadurch ein Widerspruch zwischen Beobachtung und Theorie auftreten. Die scheinbare Widerlegung der Theorie stellt sich als Inkonsistenz zwischen Theorie und Empirie heraus, die durch falsche Beobachtungstheorien verursacht wurden. Das strenge Prinzip der Falsifikation wird auf diese Weise erneut relativiert.

Als drittes, exemplarisch herausgegriffenes Problem ist das des *Basissatzes* zu nennen. Das Grundschema empirischer Überprüfung geht davon aus, daß eine Hypothese mit Erfahrungsdaten konfrontiert wird. Tatsächlich jedoch können wir nie Hypothesen unmittelbar mit Sinneswahrnehmungen konfrontieren, sondern nur mit Sätzen über Sinneswahrnehmungen, bzw. Sätzen über die Ergebnisse wissenschaftlich-systematischer Beobachtungen.
Das ist genau der Ansatzpunkt: *Sätze sind stets nur mit Sätzen konfrontierbar. Zum Zwecke der Konfrontation mit den hypothetischen Sätzen müssen Sinneswahrnehmungen immer erst in Sätze übersetzt werden; diese Sätze heißen Basissätze.* Auch in diese Übersetzung können prinzipiell Fehlannahmen und Verzerrungen eingehen. Sind sie fehlerhaft, so sind sie offensichtlich zur Konfrontation mit den hypothetisch-theoretischen Aussagen kaum geeignet, weil dann die eventuell auftretende Inkompatibilität zwischen theoretischen und empirischen Aussagen nicht mehr beurteilbar erscheint. Als Fazit kann daher festgehalten werden, daß eine eindeutige und endgültige Falsifikation von Hypothesen zwar rein logisch als Konfrontation von Sätzen möglich ist, jedoch realiter problematisch erscheinen muß, wenn nicht ausgeschlossen werden kann, daß Fehlerquellen in die Formulierung der Basissätze eingeflossen sind.

Letztlich kann also das Problem der Wahrheitsqualität nur durch Dezisionen der Wissenschaftler über die Annahme oder Ablehnung von Basissätzen auf dem Wege des Konsens entschieden werden. Die so exakte Überprüfungsmöglichkeit theoretischer Aussagen an der Realität entpuppt sich als problembehaftet und scheinbare, die in besonderer Weise dann zutrifft, wenn in den theoretischen Aussagen solche Begriffe verwandt werden, die nicht unmittelbar, direkt empirisch beobachtbar sind.

1.5 Messen und Skalierung

Wenn Wissenschaft es darauf anlegt, die theoretisch formulierten Hypothesen einer empirischen Überprüfung zuzuführen, um damit Aussagen über die Richtigkeit oder Falschheit der Hypothesen vorzunehmen, so muß die Konfrontation von Hypothese und Realität methodologisch faßbar gemacht werden. Die bisherigen Überlegungen zur Begriffsbildung, Hypothesenkonstruktion und Operationalisierung waren Vorstufen für die empirische Überprüfung der Hypothesen. Realitätsgerechte Überprüfung von Hypothesen bedeutet jedoch auch, daß der zu erfassende Objektbereich in seiner wahren Gestalt mit den Hypothesen konfrontiert wird. Sie setzt Regeln voraus, nach denen – mehr oder weniger gut wissenschaftlich abgesichert – eine zuverlässige und gültige Datenerhebung möglich ist. Erfolgt die Datenerhebung nach solchen Regeln, die im weiteren zu spezifizieren sind, so spricht man von Messung. Die Hypothesenüberprüfung ist also unabdingbar mit dem Messen des Objektbereiches verknüpft.

Einige grundsätzliche Elemente des Messens können bereits vor einer exakten Definition des Messens herausgearbeitet werden: Wie aus den bisherigen Ausführungen implizit zu entnehmen war, beginnt Messung keineswegs erst bei der abschließenden Gewinnung von Daten, wie die umgangssprachliche Verwendung des Begriffs suggeriert. Gerade auch in den Sozialwissenschaften ist davon auszugehen, daß Messung bereits bei der Entwicklung eines Meßinstruments und der Entwicklung der Meßgrößen (Variablen, Indikatoren, Merkmalsausprägungen) beginnt, denn dort werden bereits Vorentscheidungen darüber getroffen, wie der Objektbereich ermittelt werden kann.

Mit dem Begriff des Messens wird mehr oder weniger automatisch der des Quantifizierens assoziiert. So ist zweifellos richtig, daß mit zunehmendem Erkenntnisstand und fortschreitender Entwicklung der Wissenschaften sich eine immer stärkere Tendenz auf das Quantifizieren hin entwickelt hat, was nicht allein damit begründet wird, daß Quantifizierung ein Wert per se sei. Vielmehr ist zu sehen, daß es mit der Quantifizierung gelingt, die hier unübersehbare Fülle von Einzelinformationen so zu filtern und zusammenzufassen, daß sie menschlicher Erkenntis überhaupt zugänglich wird. Quantifizierung erfolgt also nicht, weil man davon ausgeht, daß der zu messende Objektbereich grundsätzlich quantitativer Natur sei, sondern weil man hofft, durch Informationsreduktion (vgl. hierzu 1.6) menschliche Erkenntnis im Rahmen der Fülle möglicher Wahrnehmungen zu erleichtern. Quantifizierung so verstanden, läßt sich nicht in Gegensatz zu Qualität bringen. Der Vorwurf der Auflösung von Qualität in Quantität ist genau da unberechtigt, wo durch die Quantifizierung die Qualität nicht verlorengeht. (So könnte man sich vorstellen, daß die „Qualität" des Geschlechtes in den Ausprägungen als männlich oder weiblich auch quantitativ erfaßt werden kann, indem z.B. der Hormonhaushalt durch subtile Meßmethoden quantifiziert und gemessen wird, was keineswegs zu einer Auflösung der Qualität führen würde, sondern bestenfalls zu einer Spezifizierung.).

Das Messen kann ebenso, wie eine fundamentale Kritik an ihm, nicht losgelöst von dem jeweiligen Objektbereich, der erfaßt werden soll, diskutiert werden. Wenn man im Meßprozeß jeden einzelnen Schritt im Meßverfahren hinsichtlich seiner Voraussetzungen, seiner Abstraktionen und seiner Einschränkungen vollständig expliziert, besteht durchaus die Möglichkeit, daß die Zuverlässigkeit dieser Schritte zwar nicht abstrakt und generell, aber auf den Gegenstand bezogen, beurteilt werden kann. Messung kann sich also immer nur am Gegenstand, der gemessen werden soll, erweisen (und zwar nicht am Gegenstand als solchem, sondern hinsichtlich bestimmter Eigenschaften dieses Gegenstandes). Daher sollte man nicht grundsätzlich für oder gegen quantifizierendes Messen plädieren, sondern die Gültigkeit und Zuverlässigkeit einer Messung anhand der konkreten Messung eines bestimmten Gegenstandes von Fall zu Fall beurteilen.

Auch das allgemein vorgebrachte Argument, daß bei quantitativer Messung die intersubjektive Überprüfbarkeit der gewonnenen Aussagen höher ist als beim qualitativen Messen, ist in seiner Allgemeinheit mit Vorsicht zu genießen, denn man kann sich durchaus vorstellen, daß durch das Anlegen eines quantitativen Meßinstrumentes erheblich mehr Fehlerquellen auftreten als bei einer nur qualitativen Zuordnung (man vergleiche z.B. die Einstellungsmessung, die in sieben graduellen Abstufungen vorgenommen wird (damit quantitativ ist) gegenüber jener, die nur qualitativ entscheidet zwischen „pro und kontra"). Zwar ist richtig, daß die Quantifizierung, wie aus dem Beispiel ersichtlich, differenziertere Aussagen ermöglicht; aber ob diese differenzierteren Aussagen gültiger und auch zuverlässiger sind, kann nicht allgemein entschieden werden. Man glaubt aber, daß Qualitäten als oder in Quantitäten in Erscheinung treten oder zumindest operational sinnvoll so konzipiert werden können: „ ... whether we can mesure something depends not on that thing, but on how we have conceptualized it, on our knowledge on it ..." (A. Kaplan, The Conduct of Inquiry 1964, S. 176). Kaplan meint schließlich auch: Zu sagen, etwas könne nicht gemessen werden, komme der Aussage gleich, daß unsere Vorstellungen darüber unbestimmt bleiben müssen, d.h. also,

1. daß hier die Auffassung vertreten wird, daß die quantitative Fassung von Phänomenen unsere unbestimmten Vorstellungen von diesen Sachverhalten zu bestimmten machen,

2. daß Quantifizierung eindeutig mit den theoretischen Ausgangskonzeptionen in Zusammenhang steht und

3. daß darüber die Frage, ob Objekten in der Forschung quantitative Eigenschaften zukommen, in den Hintergrund tritt.

Gleichwohl gilt zu berücksichtigen: „Messung erfaßt die Gegenstände nicht als solche, sondern nur den Begriff, den wir uns davon machen. Genauer gesagt, richtet sie sich auf ausgewählte Merkmale des Gegenstandes, die in der Regel jedoch nicht direkt gemessen werden, sondern über Indikatoren. Die Indikatoren schließlich werden beschrieben mit Hilfe von Zahlen, die als solche einer eigenen Wirklichkeitsschicht entstammen. Die große Präzision, die unsere Messung erreicht, ent-

steht also in einiger Entfernung vom Objekt selbst und kommt zustande erst nach Brückenschlägen, die jeweils mit eigenem Risiko behaftet sind." (HARTMANN,H., Empirische Sozialforschung, München 1970).

1.5.1 Definition des Messens

„Messen besteht im Zuordnen von Symbolen (in der Regel Zahlen, bzw. Ziffern) zu Objekten derart, daß bestimmte Relationen zwischen den empirischen Objekten sich in den Relationen zwischen den Symbolen widerspiegeln." (LAMNEK,S., Methodenkritisches Verständnis, in: RATHGEBER (Hrsg.), Medizinische Psychologie, München 1977, 2. Aufl. S. 2).

Bezeichnet man die zu messende, empirische Objektmenge als empirisches Relativ und die ihr zugeordneten Symbole als numerisches Relativ, so erfordert eine zuverlässige und gültige Zuordnung der Elemente des empirischen Relativs zu denen des numerischen Relativs eine Regel, die diese Zuordnung definiert. Diese Regel wird durch eine entsprechende Zuordnungfunktion φ angegeben. Eine Skala oder Skalierung kann nun als das geordnete Tripel (A, Z, φ) definiert werden, wobei A das empirische Relativ, Z das numerische Relativ und φ die Zuordnungsfunktion ist. *Eine Skalierung ist also nichts anderes, als das Ergebnis eines Meßvorgangs, nämlich die Abbildung einer bestimmten empirischen Struktur in eine entsprechende numerische Struktur.* Das Messen und mithin das Meßergebnis (Skalierung) ist dann korrekt, wenn ein empirisches Objekt A durch die numerische Abbildung φ(A) derart reproduziert wird, daß gleiche Objekte bzw. gleiche Eigenschaften von Objekten (= empirisches Relativ) auch jeweils gleichen Symbolen zugeordnet werden und entsprechen (= numerisches Relativ).

$$a \sim b \longrightarrow \varphi(a) = \varphi(b)$$

Hier zeigt sich allerdings eine Schwierigkeit in der Begriffsbestimmung, die auf die Unschärfe in der Reichweite des Zuordnungsprozesses zwischen empirischem und numerischem Relativ basiert: Wie schon erwähnt, beginnt Messung bereits bei der Entwicklung von Meßgrößen und zwar deshalb, weil diese Entwicklung bereits im Hinblick auf die späteren Zuordnungen zu Zahlensymbolen bzw. numerischen Relationen erfolgen muß. Es handelt sich dabei um die Skalierungsverfahren, die sich der doppelten Forderung nach theoretischer Adäquanz und mathematisch-operationaler Verwendbarkeit gegenübersehen. *Skalierung ist nicht nur Voraussetzung für die Erhebung von Daten, sondern erzeugt diese Daten in ihrer Anwendung durch das Meßinstrument.* Zwar gilt das nicht so absolut und generell, wie hier formuliert, doch im allgemeinen möchte man den Prozeß der Datenaufbereitung, der sich anschließt, weitestgehend von theoretischen Grundproblemen freihalten. Theoretische Aussagen über ein empirisches Relativ präjudizieren die Meßoperationen, wie diese die Erkenntnismöglichkeiten determinieren. Dementsprechend sind Hypothesenbildung, Operationalisierung und Messung nicht voneinander trennbare Bereiche, sondern ineinander verschränkt.

1.5.2 Voraussetzungen des Messens

Bevor die Absicht des Messens realisiert werden kann, müssen einige grundlegende Voraussetzungen gewährleistet sein, die für die empirischen Wissenschaften allgemein gelten (vgl. hierzu KRAPP/PRELL. Studienhefte zur Erziehungswissenschaft, Heft 5: Empirische Forschungsmethoden, München 1975, S. 12f):

1. *Die zu beobachtenden Tatsachen müssen eine bestimmte Konstanz* aufweisen, um sie empirisch zuverlässig und intersubjektiv überprüfbar ermitteln zu können. Das rein zufällige Auftreten würde die empirische Vorgehensweise entscheidend einengen.

2. Wie schon ausgeführt, müssen die in den Hypothesen aufscheinenden *theoretischen Begriffe einen direkten oder indirekten empirischen Bezug* aufweisen, weil sonst deren Konfrontation mit der Realität nicht möglich erscheint.

3. Wissenschaftliche Aussagen müssen so formuliert sein, daß sie *intersubjektiv nachprüfbar und empirisch kritisierbar* (falsifizierbar oder verifizierbar) sind.

Neben diesen allgemeinen Kriterien für empirische Wissenschaften, die ihre Erkenntnisse durch das Messen bestimmter Phänomene überprüfen wollen, sind weitere Voraussetzungen im engeren Sinne des Messens für dessen Realisierung notwendig:

1. *Messen setzt grundsätzlich voraus, daß eine empirische Struktur der Objektmenge vorliegt.* Wäre der Objektbereich nicht durch gewisse Ordnungsprinzipien (welcher Art diese auch immer sein mögen) strukturiert, würde sich Messen ausschließen, weil nicht entscheidbar wäre, was, wo, wie gemessen werden soll. Da sowohl die Naturwissenschaften (z.B. durch physikalische Gesetze), wie auch die Sozialwissenschaften (z.B. durch Normen, Werte etc.) eine solche empirische Struktur aufweisen, kann im Regelfall davon ausgegangen werden, daß diese Voraussetzung erfüllt ist.

2. Das Messen war definiert worden als die Zuordnung eines numerischen Relativs zu einem empirischen Relativ, gemäß bestimmten Regeln (Zuordnungsfunktion). Damit diese Definition und in deren Anwendung das Messen gültig ist, bedarf die Zuordnungsfunktion einer Spezifizierung dahingehend, daß numerisches und empirisches Relativ strukturidentisch sein müssen, oder anders formuliert: *das numerische Relativ muß eine strukturtreue Abbildung des empirischen Relativs sein.* Die Struktur zeigt sich dabei darin, daß Relationen im numerischen Relativ die Relationen des empirischen Relativs widerspiegeln. *Solche strukturtreue Abbildungen nennt man isomorph (eineindeutig) oder homomorph (eindeutig).*

3. Liegt eine solche strukturtreue Abbildung des empirischen Relativs im numerischen Relativ vor, so muß es möglich sein, *aufgrund der numerisch festgestellten Relationen auf ebensolche im empirischen Objektbereich zu schließen und vice versa.* Diese Voraussetzung des Messens ist als konsequente Folge der strukturtreuen Abbildung zu gewinnen. Liegt Strukturtreue vor, so sind Relationen im numerischen Relativ als solche des empirischen Relativs interpretierbar.

4. Während der Objektbereich nicht beliebig manipuliert werden kann, sind im numerischen Relativ durchaus mathematisch-statistische Operationen und Transformationen möglich (wie im Abschn. 1.6 noch zu zeigen sein wird). Die mathematisch-statistischen Transformationen des numerischen Relativs sind jedoch so vielfältig und gehorchen unterschiedlichen mathematischen Axiomen, daß darauf geachtet werden muß, daß *Transformationen im numerischen Relativ dessen Strukturtreue zum empirischen Relativ nicht verändern* (vgl. hierzu den folgenden Abschnitt 1.5.3). So könnte man sich z.B. vorstellen, daß Additionen von Meßwerten die Strukturtreue zum empirischen Relativ nicht verändern, wohingegen Divisionen von Meßwerten zu falschen Rückschlüssen auf das empirische Relativ führen.

1.5.3 Meßniveaus

Messen besteht im Zuordnen eines numerischen Relativs zu einem empirischen Relativ derart, daß Strukturtreue zwischen beiden gewährleistet ist. An einem einfachen Beispiel kann man sich nun verdeutlichen, daß offensichtlich einem bestimmten empirischen Relativ mehrere, gleichwohl strukturtreue numerische Relative zugeordnet werden können.

Abb. 4: Strukturidentische numerische Relative

empirisches Relativ	numerische Relative
männlich	♂, m, 0, +
weiblich	♀, w, 1, –

Wenn es aber eine Vielzahl von numerischen Relativen gibt, die jeweils strukturtreu zu einem bestimmten empirischen Relativ sind, so bedeutet dies, daß numerische Relative offensichtlich in andere numerische Relative transformiert werden können. *Je nach dem, welche Transformationen im numerischen Relativ durchführbar sind, ohne dessen Strukturtreue zum empirischen Relativ zu verändern, unterscheidet man verschiedene Meßniveaus oder Skalentypen.*

Der Zuordnungsprozeß von Merkmalen oder Merkmalsausprägungen eines Objektbereiches zu Symbolen (im Regelfalle zu Ziffern, weil von Zahlen im numerischen Sinne nicht bzw. noch nicht gesprochen werden kann), wird zwar insgesamt Messung genannt, weist jedoch sehr heterogene Formen mit unterschiedlichen Implikationen auf, sodaß die Differenzierung nach Meßniveaus bzw. Skalentypen notwendig erscheint. In einer ersten Grobgliederung unterscheidet man *topologische Skalen* (Zählen), zu denen die Nominalskala und die Ordinalskala gehören, von *metrischen Skalen* (Messen), zu denen die Intervall- und Ratioskala gehören.

Topologische Skalen beruhen im Grunde genommen auf dem Zählen. Es erfolgt eine im Prinzip willkürliche Zuordnung von Merkmalen bzw. Merkmalsausprägungen zu Symbolen (Ziffern), die den Bedingungen des Messens, wie sie vorher ent-

wickelt wurden, genügen. Die Symbole des numerischen Relativs müssen dabei den Kriterien von Trennschärfe und Vollständigkeit genügen, d.h. daß jeder Meßwert des interessierenden Objektbereiches einem und nur einem der Symbole des numerischen Relativs zugeordnet wird. So werden z.B. alle männlichen Untersuchungspersonen mit dem Symbol 0 bezeichnet, alle weiblichen mit dem Symbol 1. Somit stellen wir nur Gleichheits- und Ungleichheitsrelationen fest, d.h. Männer = Männer und Frauen = Frauen (0 = 0 und 1 = 1) bzw. Männer ≠ Frauen (0 ≠ 1). Diese „primitive" Form des Messens wird für klassifikatorische Merkmale angewandt.

Ein *klassifikatorischer* Begriff ermöglicht die Konstruktion einer Nominalskala, d.h. nominales Messen. Man betrachtet nur die Relationen „gleich" und „ungleich"; die Zahlen haben eigentlich nur eine namengebende Aufgabe.

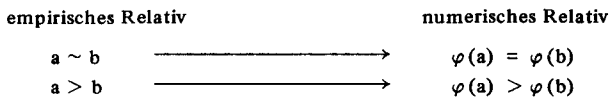
empirisches Relativ		numerisches Relativ
$a \sim b$	⟶	$\varphi(a) = \varphi(b)$
$a \not\sim b$	⟶	$\varphi(b) \neq \varphi(b)$

Sofern tatsächlich ein klassifikatorischer Begriff gegeben ist, d.h. ein Begriff, dessen Extension eine Klasseneinteilung darstellt, ist die so definierte Abbildung strukturtreu, also eine Skala. Alle ein-eindeutigen Transformationen der Skalenwerte liefern wieder eine strukturtreue Abbildung und sind daher legitim. Skalen, die durch ein-eindeutige Transformationen wieder in strukturidentische Abbildungen übergeführt werden, nennt man *Nominalskalen.* Bei Nominalskalen sind also alle Aussagen über die Relationen zwischen Skalenwerten empirisch sinnvoll, deren Wahrheitswert sich nicht durch ein-eindeutige Transformationen verändert.

Nominalskalen quantifizieren nicht, sie klassifizieren. Demgemäß haben die Ziffern im numerischen Relativ keine Zahlenfunktionen; sie ordnen sich nur empirischen Tatsachen zu. Da diese Zuordnung geradezu beliebig erfolgen kann, sofern die Gleich- und Ungleichheitsrelation nicht verletzt wird und keine weiteren zusätzlichen Regeln zu beachten sind, kann auf dem Niveau einer Nominalskala eigentlich immer gemessen werden, soweit der zu messende Objektbereich eine feststellbare Struktur und mithin eine Zuordnungsmöglichkeit aufweist. *Da Nominalskalen die geringsten Voraussetzungen an das Messen stellen, spricht man auch von dem niedrigsten Meßniveau.*

Beispiele für nominales Meßniveau: Die Aufteilung der Fachärzte in deren Spezialdisziplinen; die Beurteilung des Patienten, ob dieser gesund oder krank ist; das Geschlecht; der Familienstand; die Religionszugehörigkeit etc.

Ein *komparativer* Begriff ermöglicht die Konstruktion einer Ordinalskala, d.h. ordinales Messen. Zusätzlich zu den Relationen der Nominalskala werden bei der Ordinalskala die Relationen „kleiner" und „größer" aufgenommen, d.h. die Zahlen haben zusätzlich ordnende Aufgaben.



Sofern tatsächlich ein komparativer Begriff gegeben ist, also ein Begriff, dessen Extension eine Quasireihe darstellt, ist die so definierte Darstellung strukturtreu.

Die Ordinalskala unterscheidet sich von der Nominalskala also dadurch, daß sich die Ausprägungen des zu messenden Merkmals zusätzlich zur Gleich- und Ungleichheitsrelation ordnen lassen. Alle numerischen Relative müssen – wollen sie strukturtreu sein – diese Ordnung widerspiegeln. Da sich solche Ordnungen nur innerhalb einer Dimension vornehmen lassen, sind Ordinalskalen eindimensional; sie beziehen sich auf ein Merkmal mit mehreren abgestuften Ausprägungen, wobei entscheidbar sein muß, welche Ausprägung anderen Werten vor- oder nachgeordnet ist. Ordinalskalen bieten gegenüber den Nominalskalen einen Informationsgewinn, weil nicht nur klassifiziert sondern zusätzlich eine komparative Rangordnung aufgestellt werden kann. (So kann man z.B. die einzelnen Facharztausbildungen bestenfalls aufzählen, nicht jedoch in eine Ordnung bringen; mißt man aber die Variable des Einkommens für die Fachärzte, so wird man etwa feststellen können, daß der Internist mehr verdient als der Anaesthesist und dieser mehr als der Augenfacharzt.)

Beispiele für ordinales Meßniveau: Die Entscheidung des Arztes bei Bettennot im Krankenhaus, welche Patienten eher entlassen werden können als andere; die Prioritätenliste von Maßnahmen bei einer Erkrankung (z.B. Steigerung der Dosis eines Medikaments); die fachliche Beurteilung des Krankenhauspersonals durch den Chefarzt; Zeugnisnoten u.a.m.

Die bisherigen Strukturmerkmale des empirischen und des zugeordneten numerischen Relativs betrafen Axiome, die durch die Forderungen nach Trennschärfe, Vollständigkeit und bei der Ordinalskala zusätzlich nach eindeutiger Reihenfolge bestimmt waren. Als zusätzliches Strukturmerkmal tritt bei der Intervallskala die Definition des Abstandes zweier Elemente des Relativs hinzu. Diese Feststellung ist wesentlich, weil damit angesprochen ist, daß die möglichen – nun schon mathematischen – Operationen innerhalb des numerischen Relativs nur dann sinnvoll sind, wenn dem empirischen Relativ strikte Eindimensionalität unterlegt werden kann. Eine weitere wichtige Bedingung ist, daß Abstände sinnvoll nur innerhalb eines Kontinuums definiert werden können. Jedes Intervall der reellen Zahlen und das System der reellen Zahlen insgesamt, erfüllen diese Forderungen. Das empirische Relativ muß diese Bedingungen ebenfalls erfüllen, auch wenn in der Praxis z.B. nur bestimmte Punkte dieses Kontinuums (Intervallpunkte) von Interesse sind.

In der Praxis hat man es im allgemeinen mit geschlossenen Intervallen aus der Menge der reellen Zahlen zu tun. Die Randpunkte eines solchen numerischen Relativs entsprechen im empirischen Relativ den Grenzen, innerhalb deren sich

Ausprägungen des Merkmals bewegen. Diese Randpunkte oder Grenzwerte definieren einen Abstand, der in *gleiche* Teile (Intervalle) geteilt wird. Die Anzahl der Teilungspunkte ist bei der Frage von Interesse, ob die Intervalle empirisch und theoretisch relevanten Differenzierungen entsprechen. *Der Abstand der beiden Randpunkte definiert in diesem Zusammenhang einen Maßstab, die Anzahl der Teilungspunkte eine Maßeinheit.* Der Grundsachverhalt, der hier vorliegt, ist also der, daß bestimmte Merkmalsausprägungen in der Form gemessen werden, daß sie auf die Messungen von Längen oder Entfernungen zurückgeführt werden.

Zusätzlich zu den Relationen der Ordinalskala werden bei der Intervallskala die Relationen betrachtet, die Aussagen über die Differenzen zwischen Skalenwerten zulassen. Die Intervalle zwischen den Skalenwerten haben also hier (im Gegensatz zur Ordinalskala) einen empirischen Sinn.

empirisches Relativ		numerisches Realtiv
$a \sim b$	$\longrightarrow$	$\varphi(a) = \varphi(b)$
$a > b$	$\longrightarrow$	$\varphi(a) > \varphi(b)$
$(b,a) \sim (d,c)$	$\longrightarrow$	$\varphi(a) - \varphi(b) = \varphi(c) - \varphi(d)$
$(b,a) > (d,c)$	$\longrightarrow$	$\varphi(a) - \varphi(b) > \varphi(c) - \varphi(d)$

Alle *affinen* Transformationen $\varphi'(a) = \alpha\,\varphi(a) + \beta$ mit $\alpha > 0$ liefern wieder eine strukturtreue Abbildung und sind daher legitim. Skalen, die durch affine Transformationen wieder in strukturtreue Abbildungen überführt werden, nennt man *Intervallskalen.*

Affine Transformationen enthalten zwei freie Parameter, die Koeffizienten α und β. Man sagt daher auch, daß bei Intervallskalen der Nullpunkt und die Meßeinheit beliebig festgelegt werden können. Intervallskalen erfüllen die Bedingungen von Nominal- und Ordinalskalen und weisen die zusätzliche Qualität auf, daß die Abstände zwischen den Meßwerten (Merkmalsausprägungen) einen empirischen Sinn haben. Somit bietet die Intervallskala gegenüber den topologischen Skalen einen weiteren Informationsgewinn.
Beispiele für Intervallskalen: Blutdruckmessung, Körpertemperatur, Intelligenzquotient usw.

Die Ausführungen zur Intervallskala gelten auch für das höchste Skalenniveau, die Ratioskalen. Die Ratioskala hat zusätzlich einen natürlichen 0-Punkt. Im empirischen Relativ entspricht dieser Sachverhalt der Aussage, daß das betreffende Merkmal die Intensität = 0 annehmen kann. (Es ist zu betonen, daß die Aussage, das betreffende Merkmal tritt nicht auf, notwendige, aber nicht hinreichende Voraussetzung ist).

Zusätzlich zu den bisher betrachteten Relationen kommen bei der Verhältnisskala solche hinzu, die Aussagen über das Verhältnis zweier Skalenwerte erlauben. Im Gegensatz zu allen bisher genannten Skalen haben hier Quotienten zwischen Skalenwerten einen empirischen Sinn. (Unterschied zur Intervallskala: natürlicher Nullpunkt!).

empirisches Relativ		numerisches Relativ
$a \sim b$	$\longrightarrow$	$\varphi(a) = \varphi(b)$
$a > b$	$\longrightarrow$	$\varphi(a) > \varphi(b)$
$(a \circ b)$	$\longrightarrow$	$\varphi(a) + \varphi(b)$

Alle *linearen* Transformationen $\varphi(a) = \alpha' \varphi(a)$ mit $\alpha > 0$ liefern wieder strukturtreue Abbildungen und sind daher legitim – Skalen, die durch lineare Transformationen wieder in strukturtreue Abbildungen überführt werden können, nennt man *Ratioskalen.* Lineare Transformationen enthalten einen freien Parameter: den positiven Koeffizienten, der beliebig gewählt werden kann. Man sagt daher: Bei Ratioskalen kann die Maßeinheit beliebig festgesetzt werden. Ratioskalen liegen auf dem höchsten Meßniveau. Aus der Tatsache, daß es nur wenige sozialwissenschaftlich relevante Merkmale gibt, die einen natürlichen Nullpunkt haben, ist zu entnehmen, daß sie nur eine untergeordnete Rolle spielen. Dies gilt auch deswegen, weil die Differenzen zwischen Intervall- und Verhältnisskalen bei den statistischen Berechnungen – wo die Skalenqualität ein entscheidendes Kriterium für deren Anwendung ist – praktisch nicht mehr relevant sind.

Beispiele für Ratioskalen: Einkommen, Alter, Länge, Gewicht etc. Die doch relativ abstrakt gehaltenen Ausführungen zu den Meßniveaus sollen abschließend durch ein Beispiel illustriert werden, um das Verständnis beim Leser zu erleichtern (nach J. KRIZ, Statistik in den Sozialwissenschaften, Reinbek 1973):

Drei 5000-m-Läufern, A, B und C, werden Startnummern auf den Rücken geheftet: dem A die 1, dem B die 2 und C die 3. Das empirische Relativ besteht also aus den Elementen A, B und C, das numerische aus 1, 2 und 3. Da den empirischen Elementen Zahlen zugeordnet wurden, handelt es sich offenbar um einen Meßvorgang gemäß der oben gegebenen Definition.

Welche Transformationen des numerischen Relativs sind nun möglich, ohne die Struktur des empirischen Relativs zu verändern?

a) Nehmen wir an, die Nummernvergabe erfolgte zu dem Zwecke, die Läufer für den Zuschauer identifizierbar zu machen, so kommt es doch nur darauf an, jedem Läufer eine andere Nummer zuzuordnen, z.B. dem Läufer A 712, dem B die 311, dem C die 713. Es ist also jede Transformation des numerischen Relativs möglich, die die Läufer unterscheidbar beläßt. Solche Transformationen nennt man *ein-eindeutig.*

empirisches Relativ	numerisches Relativ	(legitime Transformationen)
A	$\varphi(A) = 1$	$\varphi'(A) = 7$ oder $\varphi''(A) = 122$
B	$\varphi(B) = 2$	$\varphi'(B) = 25$ oder $\varphi''(B) = 213$
C	$\varphi(C) = 3$	$\varphi'(C) = 11$ oder $\varphi''(C) = 170$

b) Sollte mit der Vergabe der Nummern an die drei Läufer jedoch zum Ausdruck kommen, daß Läufer A die größten Aussichten hat, das Rennen zu gewinnen, B geringere und C die geringsten Chancen hat, also mit der Nummernvergabe die Reihenfolge des Einlaufs prognostiziert werden sollte, so handelt es sich um eine Ordnungsrelation zwischen den empirischen Objekten, die sich im numerischen Relativ niederschlagen muß. Die unter a) vorgenommene Transformation wäre offenkundig falsch, weil die Ordnungsrelation $A < B < C$ durch $7 < 25 > 11$ und damit die Strukturtreue der Abbildung verletzt würde. Es sind also nur Transformationen zulässig, die die Reihenfolge (Ordnung) der Läufer invariant lassen. Solche Transformationen nennt man *monoton.*

empirisches Relativ	numerisches Relativ	(legitime) Transformationen	
A	$\varphi(A) = 1$	$\varphi'(A) = 7$	$\varphi''(A) = 122$
B	$\varphi(B) = 2$	$\varphi'(B) = 11$	$\varphi''(B) = 170$
C	$\varphi(C) = 3$	$\varphi'(C) = 25$	$\varphi''(C) = 213$

c) Wollte man mit der Nummernvergabe nicht nur eine Aussage über die Reihenfolge des Zieleinlaufs vornehmen, sondern auch Angaben über den Leistungsunterschied zwischen den drei Läufern machen, so kommt es im empirischen Relativ auf die Differenzen zwischen A, B und C an und nicht mehr auf die Reihenfolge allein: A läuft genausoviel schneller als B, wie B schneller als C läuft. Wiederum sind die unter a) bzw. b) gemachten Transformationen unzulässig, weil sie die empirische Struktur verändern. Die Intervalle zwischen A und B sowie zwischen B und C sind nicht mehr gleich. Transformationen, die diese Intervalle gleich belassen, nennt man *affin*.

empirisches Relativ	numerisches Relativ	(legitime) Transformationen		
A	$\varphi(A) = 1$	$\varphi'(A) = 7$	oder	$\varphi''(A) = 5$
B	$\varphi(B) = 2$	$\varphi'(B) = 11$	oder	$\varphi''(B) = 8$
C	$\varphi(C) = 3$	$\varphi'(C) = 15$	oder	$\varphi''(C) = 11$

d) Wollte man mit der Nummernvergabe jedoch etwas über das Verhältnis zwischen A, B und C etwa derart sagen, daß A doppelt so schnell laufe wie B und dreimal so schnell laufe wie C, dann sind die in a), b) und c) vorgenommenen Transformationen falsch, denn sie verändern das Verhältnis im empirischen Relativ, das realiter ja invariant ist. Alle *linearen* Transformationen behalten jedoch die Strukturtreue des numerischen Relativs bei und sind daher legitim.

empirisches Relativ	numerisches Relativ	(legitime) Transformationen	
A	$\varphi(A) = 1$	$\varphi'(A) = 6$	$\varphi''(A) = 11$
B	$\varphi(B) = 2$	$\varphi'(B) = 12$	$\varphi''(B) = 22$
C	$\varphi(C) = 3$	$\varphi'(C) = 18$	$\varphi''(C) = 33$
		(Faktor 6)	(Faktor 11)

Man bezeichnet solche Merkmale, die nur auf nominalem bzw. ordinalem Niveau gemessen werden können, auch als *qualitativ*, jene, die auf mindestens Intervallniveau gemessen werden können, als *quantitativ*.

In der Literatur erfolgt die Einteilung der Skalen häufig anhand der Transformationen, die als zulässig angesehen werden, die also die empirischen Elemente und Strukturen invariant belassen. Es läßt sich jedoch zeigen, daß eine solche, axiomatisch-mathematische Definition zu Fehlern führt, wenn nicht zusätzliche Kriterien, die im empirischen Relativ liegen, hinzugezogen werden. Solche Kriterien, die Entscheidungshilfen für die Feststellung der Skalentypen liefern, werden als *Sinnkriterien* bezeichnet. Die Entscheidung über die Skalen bzw. Meßniveaus kann nur über Kenntnisse oder Vermutungen über das empirische Relativ erfolgen, weil dem numerischen Relativ nicht in jedem Falle anzusehen ist, welche Transformationen die empirische Struktur unverändert lassen. Dies bedeutet, daß das *Sinnkriterium ein empirisches* ist, mindestens aber nicht allein aufgrund des numerischen Relativs die Skalenqualität definiert werden kann.

1.5.4 Konsequenzen aus den Meßniveaus

Da Transformationen des numerischen Relativs (die z.B. bei den statistischen Operationen und Berechnung von Maßzahlen notwendigerweise Platz greifen) die

Tabelle 1: Meßniveaus

Grobe Skalentypen	Differenziertere Skalentypen	Allgemeine Eigenschaften	sinnvolle Aussagen	zulässige Transformationen	zulässige arithmetische Operationen
Topologische Skalen (extensives Messen)	Nominalskala	Die hier verwendeten Symbole sind noch nicht als Zahlen, sondern nur als Ziffern anzusehen, also gewissermaßen als Kürzel für begrifflich gefaßte Merkmale bzw. Merkmalsausprägungen. Rechenoperationen an diesen Ziffern haben im empirischen und numerischen Relativ keinen Sinn.	Gleichheit/ Verschiedenheit	alle, die die Gleichheits- und Ungleichheitsrelationen nicht verletzen, das sind *eindeutige* Transformationen.	absolute und relative Häufigkeiten
dto.	Ordinalskalen	Hier können bereits Zahlen als Symbole auftreten, denn für jede Teilmenge eines der Zahlensysteme gilt eine monotone Rangordnungsrelation. Die darüber hinaus dem Zahlensystem zugehörigen Eigenschaften finden jedoch keine Anwendung.	Größer/ kleiner	alle, die die Rangfolge der Elemente nicht verletzen, das sind *monotone* Transformationen	absolute und relative Häufigkeiten
Metrische Skalen	Intervallskala	Man hat es hier mit dichten und zusammenhängenden Punktmengen zu tun, die es ermöglichen, Abstände zwischen je 2 Punkten zu messen. Das System der reellen Zahlen entspricht dieser Forderung, es kann vollständig auf eine Gerade im Raum (Zahlengerade) abgebildet werden.	Gleichheit von Intervallen/ Unterschiede.	alle Transformationen vom Typ $x^* = p \cdot x + q$ das sind *affine* Transformationen	zusätzlich: Addition und Subtraktion
(intensives Messen)	Ratioskala	Hier kommen sämtliche Teilmengen der reellen Zahlen infrage, die wenigstens nach einer Seite hin mit der reellen Zahl 0 abschließen. Im allgemeinen verwendet man den positiven Teil der Zahlengerade	Gleichheit von Summen, Quotienten und Vielfachen	wie bei Intervall, wenn gilt: $q = 0$ (Streichungen) das sind *lineare* Transformationen	zusätzlich: Multiplikation und Division

Strukturtreue zum empirischen Relativ verändern können, ergibt sich bei Nichtberücksichtigung der jeweiligen Skalenqualität und somit falschen Transformationen die notwendige Folge einer nicht korrekten wissenschaftlichen Aussage. Sie tritt immer dann auf, wenn bei den statistischen Verfahren solche angewandt werden, die höhere Meßniveaus erfordern, als die Daten sie bieten. Im umgekehrten Falle, wo statistische Verfahren angewandt werden, die ein niedrigeres Niveau haben als die Daten, begeht man zwar keinen Fehler, man nimmt jedoch *einen Informationsverlust hin, der durch die Reduzierung der Meßniveaus entsteht.* Um den Zusammenhang aufzuzeigen, der zwischen den Meßniveaus als hierarchischer Struktur auffindbar ist, sei das folgende Beispiel entwickelt:

Das Einkommen ist eine Variable, die auf Rationiveau gemessen werden kann. Erhebt man z.B. das Einkommen in DM, so gibt es einen natürlichen Nullpunkt (nämlich kein Einkommen), und die Intervalle zwischen den einzelnen Meßwerten sind gleich und interpretierbar (z.B. jeweils 1 DM). Wird also das Einkommen in einem Fragebogen explizit erhoben, so läge Ratioskalierung vor.

Dieses höchste Meßniveau kann auch auf Intervallmeßniveau reduziert werden, wobei wir jedoch einen Informationsverlust in Kauf nehmen. So könnte man danach fragen, in welcher der vorgegebenen Einkommensklassen das Einkommen des Probanden fällt, wobei die Klassen jeweils von 1 – 100, 101 – 200, 201 – 300 usw. reichen. Jetzt ist der natürliche Nullpunkt weggefallen; es liegen bei den Meßwerten aber auch gleiche Abstände (gleiche Differenzen) vor, womit eine Intervallskalierung gegeben ist.

Die Intervallskala kann jedoch ebenfalls wieder auf ein niedrigeres Niveau reduziert werden. So könnte man in etwa die folgenden Klassen als Antwortalternativen in einem Fragebogen vorgegeben habe (s. Tab. 2).

Tabelle 2: Der Zusammenhang der Meßniveaus am Beispiel des Einkommens

Ratioskala	Intervallskala	Ordinalskala	Nominalskala
0 DM			
1 DM	1 – 100 DM	1000 – 1500 DM	mehr als 1500 DM
2 DM	101 – 200 DM	1501 – 1750 DM	ja/nein
3 DM	201 – 300 DM	1751 – 2000 DM	
.	.	2001 – 2250 DM	
.	.	2251 – 2500 DM	
.	.	2501 – 3000 DM	
3000 DM	3001 – 3100 DM	3001 – 4000 DM	
3001 DM	3101 – 3200 DM	4001 DM und mehr	
3002 DM	.		
.	.		
.	.		
.	.		

Da die jeweiligen Klassen eine unterschiedliche Spannweite aufweisen, also die Abstände zwischen den Klassen nicht gleich groß sind, liegt offensichtlich keine Intervallskala mehr vor. Eindeutig feststellbar ist jedoch nach wie vor eine Ordnungsrelation derart, daß die erste Einkommensklasse kleiner als die zweite ist, die zweite kleiner als die dritte, usw. Diese Ordnungsrelation ohne Abstandsbetrachtung macht die Ordinalskala aus.

Den weitestgehenden Informationsverlust nehmen wir dann hin, wenn wir das Einkommen nur nominal erfassen, indem wir z.B. eine bestimmte Grenze, z.B. 1500.-, vorgeben. Um den nominalen Charakter deutlich zu machen, könnte die Frage formuliert sein: „Verdienen Sie mehr als 1500.- DM?" Als Antwortalternativen sind vorgegeben „ja" oder „nein". Damit wäre das niedrigste Meßniveau der Nominalskala erreicht.

Daten auf hohem Meßniveau können also, wie das ausführliche Beispiel zeigen sollte, auf niedrigere Meßniveaus reduziert werden, wobei allerdings ein Informationsverlust in Kauf genommen wird. Der umgekehrte Weg jedoch – auch dieses kann aus dem obigen Beispiel abgeleitet werden – ist nicht möglich. Hat man nämlich Daten auf nominalem Meßniveau erhoben, so ist eine nachträgliche Transformation auf höhere Meßniveaus ausgeschlossen. (So kann man aus der Beantwortung der Frage, ob jemand mehr oder weniger als 1500.- DM verdient, nicht mehr auf die exakte Einkommenshöhe schließen.)

Gleiches gilt nun auch für die Transformationen des numerischen Relativs: Liegen die Daten z.B. nur auf nominalem Meßniveau und werden solche Transformationen angewandt, die mindestens intervallskalierte Daten voraussetzen, so ist eine Datenstruktur unterstellt, die realiter nicht vorhanden ist. Das angewandte statistische Modell „paßt nicht", man wird in seiner Anwendung Fehler begehen. Daher sollte man sich bei der Anwendung von statistischen Verfahren, wie auch bei deren Interpretation zunächst einmal darüber klar werden, auf welchem Meßniveau die erhobenen Daten eigentlich liegen.

1.6 Statistische Grundlagen

Keine empirische Wissenschaft kann heute auf die Statistik verzichten; insbesondere dann, wenn der Forschung daran gelegen ist, ihre empirisch gewonnenen Informationen auf theoretische Aussagen zu beziehen, muß man die auf der Wahrscheinlichkeitstheorie beruhenden Modelle der Statistik heranziehen, um gültige Schlußfolgerungen über das Kompatibilitätsverhältnis von Theorie und Empirie ziehen zu können. Aber auch im Sinne deskriptiver Hypothesen ist die Statistik heute ein unverzichtbarer Bestandteil der empirischen Wissenschaften geworden. Deshalb scheint es angebracht, einige der wesentlichsten Grundlagen statistischer Analyse knapp zu behandeln.

Die Frage nach der Funktion der Statistik in empirischer Forschung kann im Grunde genommen mit dem lapidaren Satz beantwortet werden, daß *die Statistik dazu dient, die Fülle empirischer Informationen, die durch entsprechende Untersuchungsverfahren gewonnen wurden, auf ein verständliches und interpretierbares Maß zu reduzieren.* So hätte es z.B. überhaupt keinen Sinn, 1200 repräsentativ ausgewählter Bundesbürger im Hinblick auf eine Variable zu befragen und die Befragungsergebnisse in einem Berichtsband dem Leser mitzuteilen. Seine Kapazität würde auf keinen Fall ausreichen, alle Ergebnisse behalten, verstehen und adäquat interpretieren zu können. Nun kann man aber die Fülle empirischer Informationen in numerische Informationen durch den Vorgang des Messens überführen (z.B. auch durch Codierung), so daß mit den numerischen Informationen bestimmte Transformationen vorgenommen werden können, die das empirische Relativ unverändert belassen, andererseits jedoch den Wust an Informationen z.B. auf eine einzige Maßzahl reduzieren. Zwar ist eine solche Informationsreduktion mit einem

Informationsverlust verbunden, doch kann dieser in Kauf genommen werden, wenn die Informationsreduktion dazu angetan ist, die wesentlichere und angestrebte Information als Erkenntnisgewinn zu erhalten. So kann man die Einkommensangaben der 1200 Personen auf eine spezifische Maßzahl, z.B. das arithmetische Mittel reduzieren (vgl. 1.6.1), womit man 1200 Zahlen in eine allgemein versteh- und leicht faßbare Maßzahl transformiert hat.

Der Einsatz von statistischen Verfahren (oder besser Modellen) ist also nicht durch die mathematischen Verfahren als solche bestimmt, sondern durch den theoretisch relevanten Erkenntisgewinn, der durch Informationsreduktion entsteht. Die Statistik legitimiert sich also nicht per se, sondern ihr Einsatz kann nur durch ihre Fruchtbarkeit für die theoretischen Überlegungen begründet werden.

1.6.1 Maßzahlen für deskriptive Hypothesen

Bei den theoretischen Überlegungen waren wir davon ausgegangen, daß deskriptive Hypothesen als Vorstufe relationaler Hypothesen durchaus sinnvoll und notwendig sein können. Auch die deskriptiven Hypothesen bedürfen einer empirischen Überprüfung auf ihren Wahrheitsgehalt hin. Da deskriptive Hypothesen – im Gegensatz zu den relationalen – nur Aussagen über jeweils einzelne, isolierte Variablen machen, in ihrer Aussage also nur eine Dimension auftritt, bezeichnet man sie auch als *eindimensionale (auch monovariate) Hypothesen.*

Lageparameter:

Zur Überprüfung solcher eindimensionaler, deskriptiver Hypothesen bieten sich nun, je nach dem, wie die Hypothese selbst formuliert war und auf welchem Meßniveau die zu erfassende Variable lag, unterschiedliche statistische Maßzahlen an. So kann man die Hypothese, daß das durchschnittliche monatliche Nettoeinkommen einer bestimmten Bevölkerungsgruppe bei 1500 DM liegt, nicht unmittelbar an den ausgewählten Probanden beobachten; vielmehr müssen die einzelnen Informationen so zusammengefaßt werden, daß der mit dem Begriff „durchschnittlich" gemeinte Sachverhalt auch in den empirischen Daten enthalten ist.

Als *Lageparameter für die Charakterisierung eindimensionaler Stichproben im Sinne der obigen Hypothese bietet sich das arithmetische Mittel an.* Das arithmetische Mittel als *ein Maß der zentralen Tendenz* erleichtert den Vergleich zweier empirisch erhobener Verteilungen, weil die Verteilungen nur durch jeweils eine Maßzahl repräsentiert werden. So könnte man feststellen, daß das durchschnittliche Einkommen der Männer bei 1500 DM monatlich läge, während das der Frauen bei 1300 DM ermittelt worden wäre. Diese beiden Maßzahlen sind (obgleich noch weitere Schwierigkeiten zu berücksichtigen wären, vgl. hierzu die Streuungsparameter), leichter miteinander vergleichbar und interpretierbar, als z.B. die Gegenüberstellung von jeweils 500 Meßwerten der Variablen des Einkommens.

Die Berechnung des arithmetischen Mittels einer Verteilung erfolgt mit Hilfe der Formel:

$$\bar{x} = \frac{\Sigma x_i}{n} \quad \text{oder für Häufigkeiten} \quad \bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

Man addiert also alle Merkmalsausprägungen (x_i) und dividiert diese durch die Anzahl der überhaupt gemessenen Werte ($n = \Sigma f_i$) und erhält somit das arithmetische Mittel $\bar{x}$. Dieses arithmetische Mittel stellt eine Transformation des numerischen Relativs dar und kann (vgl. hierzu 1.5.3) nur für solche Daten vorgenommen werden, die mindestens auf Intervallniveau liegen, weil als durchzuführende mathematische Operation eine Division auftritt. Das arithmetische Mittel einer Verteilung von Meßwerten ist also nur dann sinnvoll, wenn die Daten quantitativer Natur sind.

Ein anderer Parameter der zentralen Tendenz (Lageparameter) ist der *Median. Er teilt eine Verteilung in zwei gleich große Hälften,* so daß links und rechts von ihm eine gleiche Anzahl von Meßwerten liegt. Dies setzt voraus, daß man alle Meßwerte in eine Rangfolge bringt, eine Ordnung schafft. Ordnet man also die Werte einer Stichprobe der Größe nach und handelt es sich bei der Stichprobengröße um eine ungerade Zahl, so gibt der Meßwert, der genau in der Mitte dieser Anordnung steht, den Median an. Handelt es sich um eine gerade Zahl, so ergibt sich der Median aus dem arithmetischen Mittel zwischen den beiden Stichprobenwerten, die links und rechts von dem theoretischen Median liegen.

Der Median wird etwas weniger häufig als das arithmetische Mittel berechnet, weil er doch weniger aussagekräftig ist. Der Median ist eine Maßzahl, die durchaus auf dem Meßniveau einer Intervallskala liegen kann, er begnügt sich jedoch auch mit dem Niveau einer Ordinalskala. Da für Ordinalskalen das arithmetische Mittel nicht sinnvoll berechnet werden kann, sollte man in diesem Falle den Median heranziehen. Der Median ist jedoch kein Lageparameter für nominalskalierte Daten, da ja diese in keine Ordnungsrelation zueinander gebracht werden können.

Der einfachste Lageparameter, der auch für nominale Daten angegeben werden kann, ist der *Modus oder Modalwert. Er ist jener Wert, der in der Verteilung am häufigsten auftritt.* Da er auf niedrigstem Niveau mißt, ist sein Informationsgehalt relativ geringer, als der des Median und der des arithmetischen Mittels. Aus gleichem Grund können selbstverständlich auch solche Daten angewandt werden, die einem höheren Meßniveau angehören.

Das folgende Beispiel möge die Berechnung und Interpretationsmöglichkeiten dieser drei Lageparameter verdeutlichen:

Tab. 3: Beispiel für die Berechnung von Lageparametern

x_i = Einkommen	f_i	$x_i f_i$	$f_{i\,cum}$
1200.- DM	10	12 000.-	10
1250.- DM	15	18 750.-	25
1280.- DM	20	25 600.-	45
1320.- DM	30	39 600.-	75
1330.- DM	25	33 250.-	100
1350.- DM	20	27 000.-	120
1380.- DM	15	20 700.-	135
1400.- DM	15	21 000.-	150
1500.- DM	5	7 500.-	155
	$\Sigma f_i = 155$	$\Sigma x_i f_i =$ 205 400.-	

Arithmetisches Mittel $\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i} = \frac{205\,400}{155} \approx 1325.\text{- DM}$

Der Median liegt beim 78. Meßwert $\frac{\Sigma f_i}{2} = \frac{155}{2}$; diesem entspricht ein x_i von 1330.- DM, vgl. Spalte $f_{i\,min}$. Für die genaue Berechnung sei auf die Formeln der einschlägigen Statistikbücher verwiesen.

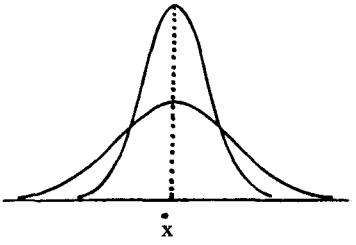
Der Modus liegt bei 1320.- DM.

Hat man eine eingipflige Verteilung, die einigermaßen symmetrisch ist, so fallen Mittelwert, Median und Modus zusammen (vgl. Gauß'sche Normalverteilung in Abb. 5). Da solche Verteilungen in der Realität relativ selten auftreten und eher mit „schiefen" Verteilungen zu rechnen ist, sollte man die Aussagekraft der einzelnen Mittelwerte kennen. Würde man in dem oben angegebenen Beispiel der Tabelle 3 noch einen weiteren Meßwert hinzunehmen, der bei 15 000,- DM läge, so würde sich der Modus überhaupt nicht verändern, der Median bei exakter Berechnung nur minimal, das arithmetische Mittel hingegen erheblich ($\bar{x}$ = 1412.- DM). Dies bedeutet, daß das arithmetische Mittel besonders empfindlich gegenüber extremen Meßwerten reagiert, während der Median etwas stärker die gesamte Verteilung von der Häufigkeitsstruktur her berücksichtigt und der Modus nur auf den häufigsten Wert der Verteilung überhaupt abstellt.

Streuungsparameter:

Vergleicht man die Verteilungen der Abbildung 5, die beide denselben Mittelwert aufweisen, so spiegelt der gleiche Mittelwert eine Ähnlichkeit vor, die realiter nicht gegeben ist. Man wird bei der Inspektion der Abbildung 5 unschwer feststellen können, *daß die Merkmalsausprägungen in der ersten Stichprobe viel enger beieinanderliegen als in der zweiten Stichprobe. Die zweite Verteilung streut stärker als die erste.* Da das arithmetische Mittel offensichtlich solche Faktoren der Streuung unberücksichtigt beläßt, kann seine Interpretation zu Verzerrungen führen.

Abb. 5. Unterschiedliche Streuungen bei gleichen arithmetischem Mittelwert



So macht es z.B. einen Unterschied aus, ob jedermann in einer Stichprobe ein Einkommen von 1500 DM hat (was einem $\bar{x}$ = 1500.- DM entspricht), oder ob dieses Durchschnittseinkommen dadurch zustandekommt, daß im Extremfall die Hälfte aller Personen 3000 DM und die andere Hälfte gar nichts verdient.

Um solche eventuell auftretenden verzerrenden Interpretationen zu vermeiden, kennzeichnet man die Verteilungen von Merkmalsausprägungen auch durch *Streuungsparameter.* Die drei wichtigsten davon seien hier besprochen. Solche Streuungsparameter sind *Dispersionsmaße* und geben an, wie stark die Meßwerte um den Mittelwert streuen, ob sie also sehr eng um ihn herum angesiedelt sind, oder ob es sehr extreme Werte gibt. Ein solches Dispersionsmaß ist die *Varianz* (s^2). Sie wird berechnet als die Summe der quadrierten Abweichungen (Differenzen) der jeweiligen Meßwerte von dem berechneten arithmetischen Mittel durch die Anzahl der Meßwerte dividiert.

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

Die Einzelabweichungen der Meßwerte von dem arithmetischen Mittel werden quadriert, weil ohne Quadrierung sich die Differenzen eventuell aufheben würden. (Gelegentlich findet man im Nenner der Formel (n − 1), was mathematisch-statistische Gründe hat, die hier vernachlässigt werden können. Für große n gilt sowieso, daß die Subtraktion von 1 nicht weiter ins Gewicht fällt.)

Die Wurzel aus der Varianz bezeichnet man als *Standardabweichung.* Die Standardabweichung hat im Gegensatz zur Varianz dieselbe Benennung wie die Stichprobenwerte als solche (z.B. DM), während die Varianz jeweils die quadrierten Benennungen enthält (also DM^2), womit die Standardabweichung ein leichter verständliches Maß für die Streuung der Stichprobe abgibt.

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

Ein sehr einfaches Maß für die Streuung einer Stichprobe ist die *Spannweite (Variationsweite oder Range).* Die Spannweite (V) ergibt sich als Differenz aus dem

maximalen und dem minimalen Wert eines Merkmals, die bei einer empirischen Erhebung festgestellt wurden. Zwar ist ein solches Dispersionsmaß relativ unspezifisch, doch kann sein Einsatz (z.B. in Kombination mit dem arithmetischen Mittel) durchaus zusätzliche Erkenntnisse liefern, sodaß nicht von vornherein auf diese Information verzichtet werden sollte. (So macht es einen Unterschied, ob bei einem gleichen arithmetischen Mittel die Spannweite der Meßwerte bei 1 000 oder bei 10 000 liegt.)

Tab. 4: Beispiel für die Berechnung von Streuungsparametern*

V_p	x_i = Intelligenzquotient	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
A	140	+ 30	900
B	135	+ 25	625
C	100	− 10	100
D	95	− 15	225
E	120	+ 10	100
F	115	+ 5	25
G	90	− 20	400
H	100	− 10	100
I	95	− 15	225
$\Sigma f_i = n = 9$	$\Sigma x_i = 990$		2700

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{990}{9} = 110$$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{2700}{9} = 300$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = 300 = 17{,}3$$

$$V = x_{i\,max} - x_{i\,min} = 140 - 90 = 50$$

* Hierzu wird auf Häufigkeitsdaten (also gleiche Merkmalsausprägungen für mehrere Versuchspersonen) aus Vereinfachungsgründen verzichtet.

Die hier vorgestellten Parameter für eindimensionale Verteilungen (deskriptive Hypothesen) sind also geeignet, eine Fülle von Informationen auf eine Maßzahl zu reduzieren. Sie beschreiben empirisch gewonnene Meßwertverteilungen durch knappe Maßzahlen und sind damit auch in der Lage, die Richtigkeit oder Falschheit einer deskriptiven Hypothese (– je nach deren Formulierung – diese allein, oder im Zusammenhang mit anderen Maßzahlen), zu belegen. Da Varianz und Standardabweichung das arithmetische Mittel in ihrer Berechnung benötigen, versteht sich von selbst, daß das Meßniveau der Daten bei deren Anwendung mindestens intervallskaliert sein muß.

1.6.2 Die Logik der Prüfung relationaler Hypothesen

Theoretische, relationale Hypothesen behaupten einen Zusammenhang zwischen mindestens zwei Variablen. Dieser theoretisch vermutete Zusammenhang soll auf seine empirische Richtigkeit hin überprüft werden, was mit Hilfe der statistischen Signifikanztests erfolgt. Dabei wird davon ausgegangen, daß der theoretisch formulierten Hypothese eine konträre Hypothese gegenübergestellt wird, die den „konservativen" Standpunkt vertritt, es gäbe keinerlei Abhängigkeit zwischen den in der theoretischen Hypothese zueinander in Beziehung gesetzten Variablen. Man

bezeichnet letztere Hypothese als *Nullhypothese,* während die theoretisch aufgestellte Hypothese als *Alternativhypothese* angesehen wird. Die Prüfung einer relationalen Hypothese beginnt also damit, daß die theoretische Alternativhypothese einer statistischen Nullhypothese gegenübergestellt wird und danach gefragt wird: Wie wahrscheinlich ist es, daß die empirisch gewonnenen Informationen der Alternativhypothese oder der Nullhypothese eher entsprechen (zunächst noch vorläufig formuliert).

Es werden also nicht die theoretischen Aussagen unmittelbar mit der empirischen Realität konfrontiert, sondern es wird die statistische Nullhypothese in dem Sinne angewandt, daß geprüft wird, wie wahrscheinlich es ist, daß eine in der Realität gewonnene empirische Verteilung von der in der Nullhypothese behaupteten, zufälligen Verteilung abweicht. Diese Überlegung muß noch verdeutlicht werden: Da wir es im Regelfalle nicht mit Vollerhebungen, sondern mit Stichproben zu tun haben, ist normalerweise schon allein durch die Stichprobenerhebung zu erwarten, daß in den so ermittelten empirischen Daten eine Variation auftritt, die aber nicht als systematische Variation innerhalb der Daten zu deuten, sondern auf die Tatsache der Teilerhebung zurückzuführen ist.

So wird, wenn man einen bevölkerungsrepräsentativen Querschnitt von 1 200 Personen nach der Häufigkeit der Arztbesuche fragt, z.B. ein Wert von 0,6 pro Jahr gewonnen. Der tatsächliche Wert für die gesamte Bevölkerung der Bundesrepublik Deutschland mag durchaus bei 0,65 oder 0,7 oder auch bei 0,55 oder 0,5 liegen. Die Variation (die Abweichung), die wir aufgrund unserer Teilerhebung erhalten haben, ist aber keineswegs eine systematisch zu nennende, sondern auf die Stichprobenerhebung zurückzuführen.

Die Frage, die nun zu stellen und zu beantworten ist: Wie groß darf eine durch die Stichprobe provozierte Variation sein, damit diese als rein zufällig entstanden und nicht systematisch durch irgendwelche anderen Variablen verursacht gelten kann? Die Beantwortung dieser Frage erfolgt über eine Entscheidung über Zufälligkeit oder Systematik der Variation, wobei eine solche Entscheidung fehlerbehaftet sein kann. „Die induktive Statistik stellt mit dem Signifikanztest ein Verfahren zur Entscheidung zwischen alternativen Hypothesen aufgrund von Untersuchungsergebnissen zur Verfügung, welche ermöglicht, anzugeben, wie sicher die Entscheidung ist. Die bei einer solchen Entscheidung immer auftretende Unsicherheit über die Richtigkeit der Entscheidung wird durch Wahrscheinlichkeit für den Fehler erster Art quantifiziert." (Ulrich G. OPPEL, Bernhard RÜGER, Biomathematik, medizinische Statistik und Dokumentation, München 1975, S. 117). *Fehler erster Art* heißt hierbei: Da man ja nie mit Sicherheit ausschließen kann, daß eine tatsächlich gültige Nullhypothese (nämlich, daß keine Beziehung zwischen den Variablen besteht) aufgrund des statistischen Tests nicht doch verworfen wird, begeht man in der statistischen Entscheidung Fehler. Die Höhe der Wahrscheinlichkeit für diesen Fehler erster Art, die man bereit ist, noch zu akzeptieren, wird gemäß Konvention festgesetzt. So geht man in der Psychologie und Soziologie von einer *Irrtumswahrscheinlichkeit* (= Fehler 1. Art) von 0,05 oder 0,01 aus, d.h., daß man sich unter 100 Fällen nur fünf- bzw. einmal derart irrt, daß man eine Nullhypothese ablehnt, obwohl sie tatsächlich richtig ist.

In den Naturwissenschaften wird man ein solches Fehlerrisiko nicht akzeptieren können. So wird ein Mediziner nicht hinnehmen wollen, daß die Verabreichung eines bestimmten Medikamentes in 95 von 100 Fällen eine positive Wirkung zeigt, jedoch in 5 von 100 Fällen zu letalen Nebenwirkungen führt. Dieses Risiko wäre zu groß; deshalb wird man in den Naturwissenschaften (z.B. Aufzugbau, Brückenbau, Medizin etc.) eine weitaus kleinere Irrtumswahrscheinlichkeit fordern müssen.

Neben dem Fehler erster Art kennt man den *Fehler zweiter Art,* der sich darauf bezieht, daß die Alternativhypothese richtig ist, obgleich man die Nullhypothese aufgrund des statistischen Tests akzeptiert. (Der Fehler zweiter Art ist für den Signifikanztest selbst von untergeordneter Bedeutung.) Das folgende Schema stellt die beiden Fehlerarten tabellarisch gegenüber:

Abb. 6: Statistische Fehlerarten

Tatsächlich gilt Entscheidung für	H_0	H_1
H_0	$1 - \alpha$	Fehler 2. Art (β)
H_1	Fehler 1. Art (α)	$1 - \beta$

Fehler 1. Art: Eine richtige Nullhypothese wird verworfen.
Fehler 2. Art: Eine falsche Nullhypothese wird beibehalten.

Hat man das *Signifikanzniveau* α (die Irrtumswahrscheinlichkeit für den Fehler erster Art) gewählt, so geht man daran, die empirisch gewonnene Meßwertverteilung mit jener hypothetisch-theoretischen Meßwertverteilung zu vergleichen, die unter der Voraussetzung entstanden wäre, daß die Nullhypothese Gültigkeit besäße (also Unabhängigkeit zwischen den beiden Variablen) und bezieht die Differenzen zwischen den beiden Verteilungen auf die festgesetzte Irrtumswahrscheinlichkeit. Stellt sich nun heraus, daß die Irrtumswahrscheinlichkeit größer als die vorgegebene Irrtumswahrscheinlichkeit ist (statt 0,05 z.B. 0,1), wird die Nullhypothese beibehalten; man geht also weiter davon aus, daß keine Beziehung zwischen den beiden Variablen besteht, sodaß die Alternativhypothese als abgelehnt gelten muß. Erreicht man ein Signifikanzniveau, das kleiner ist als die vorgegebene Irrtumswahrscheinlichkeit, so wird mit dieser Irrtumswahrscheinlichkeit die Nullhypothese zugunsten der Alternativhypothese verworfen.

Um das bisher abstrakt Ausgeführte an einem Beispiel zu verdeutlichen, sei auf eines zurückgegriffen, das von dem berühmten englischen Statistiker FISHER stammt:

Bei einer Gesellschaft behaupte eine Dame, wenn man ihr eine Tasse Tee vorsetze, der Milch beigegeben wurde, so könne sie im allgemeinen einwandfrei schmecken, ob zuerst Tee oder zuerst Milch in die Tasse eingegossen worden sei. Wie prüft man nun diese Behauptung? Sicherlich nicht so, daß man zwei äußerlich völlig gleiche Tassen der Dame vorsetzt, wobei die eine zuerst mit Milch und dann mit Tee (MT) und die zweite zuerst mit Tee und dann mit Milch (TM) gefüllt wurden. Würde man jetzt die Dame wählen lassen, so hätte sie ja offenbar eine Chance von 50%, die richtige Antwort zu geben, auch wenn ihre Behauptung (Hypothese), daß sie die genannte Sonderbegabung hätte, falsch wäre! Richtig wäre hingegen z.B., man setzte der Dame 8 äußerlich gleiche Tassen vor, vier in der Reihenfolge MT und vier in der Folge TM. Man verteile die Tassen zufällig über den Tisch, rufe dann die Dame herbei und teile ihr mit, daß von den Tassen je vier von dem Typ MT bzw. TM sind. Ihre Aufgabe wäre nun, die vier Tassen herauszufinden. Jetzt ist die Wahrscheinlichkeit, ohne eine Sonderbegabung die richtige Auswahl zu treffen, sehr gering geworden. Aus 8 Tassen kann man nämlich auf $\frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 70$ Arten vier auswählen, d.h.,es gibt 70 Kombinationen; nur eine dieser 70 Kombinationen ist jedoch richtig. Die Wahrscheinlichkeit, ohne Sonderbegabung, also zufällig die richtige Auswahl zu treffen, ist daher mit 1/70 = 0,014 % sehr gering (Irrtumswahrscheinlichkeit). Wählt die Dame wirklich die vier richtigen Tassen, so werden wir die Nullhypothese, daß die Dame die von ihr vorgegebene Sonderbegabung nicht habe, fallenlassen und ihr diese besondere Fähigkeit zuerkennen. Dabei nehmen wir eine Irrtumswahrscheinlichkeit von 1,4 % (α = 0,014) in Kauf. (Natürlich könnte man die Irrtumswahrscheinlichkeit noch weiter verringern, indem die Zahl der Tassen erhöht wird.)

Charakteristisch für das Vorgehen, das für alle Signifikanztests gilt, war: Wir stellen zunächst die Nullhypothese auf (keine Sonderbegabung bzw. keinerlei Beziehung zwischen den Variablen) und verwerfen sie genau dann, wenn sich ein Ereignis einstellt, das bei Gültigkeit der Nullhypothese unwahrscheinlich ist.

1.6.3 Signifikanztests

Ausgangspunkt für die Signifikanztests sind eine Alternativhypothese und eine Nullhypothese. Die Überprüfung der Alternativhypothese erfolgt durch die Operationalisierung der in ihr enthaltenen Begriffe und deren Beobachtung in der Realität. Die Kombination der in einer bivariaten Hypothese enthaltenen Variablen bzw. die Kombination von deren Meßwerten, kann in einer *Kreuztabelle = Korrelationsmatrix* festgehalten werden. Diese Tabelle dient dann als Grundlage für die Durchführung des Signifikanztests. Hierzu folgendes Beispiel:

H_0: Es besteht keinerlei signifikante Beziehung zwischen Geschlecht und Rauchen von Zigaretten.

H_1: Männer rauchen häufiger als Frauen.

Bei der empirischen Erhebung habe man folgende Daten erhalten (alle anderen relevanten Variablen wie z.B. Alter etc. werden als konstant vorausgesetzt):

	♂	♀	Σ
Raucher	180	20	200
Nichtraucher	420	580	1000
Σ	600	600	1200

Tab. 5. Signifikanztest mit Hilfe der Kreuztabelle

Diese Kreuztabelle ist die einfachste Form einer Tabelle (Kontingenztabelle); sie enthält 4 Zellen; selbstverständlich kann sie beliebig erweitert werden. Die Daten dieser Tabelle liegen auf nominalem Meßniveau, was ebenfalls eine Vereinfachung darstellt. Selbstverständlich können alle empirischen Daten auch auf höchstem Meßniveau in einer Kreuztabelle zusammengestellt werden. Dies bedeutet, daß der hier zu berechnende Signifikanztest des Chi^2 – da er für das niedrigste Meßniveau anwendbar ist – auch für alle anderen höheren Meßniveaus gültig ist. Mit Hilfe des Chi^2-Testes überprüfen wir also die obige Tabelle im Hinblick darauf, ob die empirisch festgestellte Verteilung der Meßwerte zufällig oder überzufällig von jener Verteilung abweicht, die wir hätten erhalten müssen, wenn die Nullhypothese gültig wäre (also kein Zusammenhang zwischen Geschlecht und Rauchen).

In einem ersten Schritt berechnet man jene Zellenwerte f_e (Erwartungswerte, die unter der Voraussetzung der Gültigkeit der Nullhypothese zu erwarten gewesen wären). Diese Erwartungswerte sind nichts anderes als die bedingten Wahrscheinlichkeiten, die wie folgt berechnet werden: Die zu der jeweils ausgewählten Zelle gehörende Spaltensumme wird mit der Zeilensumme (Randverteilung) multipliziert und durch die Gesamtsumme (= Anzahl aller Personen) dividiert. Für die erste Zelle also: 600 x 200 : 1200 = 100 und für die letzte Zelle z.B.: 600 x 1000 : 1200 = 500.

In einem zweiten Schritt vergleicht man diese Erwartungswerte mit den tatsächlich beobachteten Zellenwerten (f_0), indem man die Differenz aus beiden bildet ($f_0 - f_e$). Diese Differenzen aus den einzelnen Zellen der Matrix müssen summiert werden, denn wir wollen Aussagen über die gesamte Verteilung machen (also über die gesamte Tabelle und nicht über einzelne Zellen: $\Sigma(f_0 - f_e)$.

Da die Aufsummierung der reinen Differenzen zwischen Erwartungs- und Beobachtungswerten (wie die obige Tabelle sehr deutlich zeigt), sich gegenseitig aufheben können, genügt die Aufsummierung der reinen Differenzen nicht. In einem dritten Schritt werden daher die Differenzen zwischen Erwartungs- und Beobachtungswerten quadriert, womit alle Werte positiv werden: $\Sigma(f_0 - f_e)^2$.

Man kann sich auch vorstellen, daß mit zunehmender Zahl der Probanden die quadrierten Differenzen zwischen Beobachtungs- und Erwartungswerten immer größer werden können und damit mit kleineren Werten nicht mehr vergleichbar sind. Deshalb geht man in einem 4. Schritt dazu über, die quadrierten Differenzen auf die jeweiligen Erwartungswerte in den Zellen zu beziehen, indem man diese durch die Erwartungswerte dividiert, um so einen besseren Vergleichsmaßstab zu erhalten. $\Sigma \frac{(f_0 - f_e)^2}{f_e}$. Die Summe dieser so berechneten Werte stellt die *Prüfgröße* dar.

In einem 5. Schritt müssen die sog. *Freiheitsgrade* der Tabelle berechnet werden. Die Chi^2-Verteilung als Prüfverteilung ist nämlich keine konstante Verteilung, sondern sie weist je nach Zahl der Freiheitsgrade eine unterschiedliche Form auf. Die Entscheidung, ob eine beobachtete Verteilung von einer theoretischen systematisch abweicht, ist daher in Abhängigkeit von den Freiheitsgraden zu sehen. Die

Freiheitsgrade einer Tabelle berechnen sich als die (Zahl der Spalten – 1) x (Zahl der Zeilen – 1).

Am Beispiel der obigen Tabelle kann man sich den Sachverhalt der Freiheitsgrade deutlich machen: Die *Randverteilung* als deskriptive Verteilung der einzelnen Variablen ist fix. (600 Frauen und 600 Männer, sowie 200 Raucher und 1000 Nichtraucher). Über die Besetzung der einzelnen Zellen ist jedoch durch die Randverteilungen allein noch nichts ausgesagt. Wählt man jedoch in einer Vierfelder-Tafel – sozusagen völlig frei und beliebig – irgendeinen passenden Wert für eine Zelle aus, so werden damit und durch die Randverteilung alle anderen Zellenwerte determiniert. Es war also nur ein Zellenwert frei bestimmbar; man hat einen Freiheitsgrad.

(Der Leser konstruiere sich z.B. eine 2x3 oder 3x3 Tabelle, gebe sich fiktiv irgendwelche Randverteilungen vor und stelle fest, wieviele Zellen frei wählbar sind, bevor alle anderen determiniert sind, und überprüfe die Richtigkeit seiner Überlegung, indem er die Zahl der Freiheitsgrade als (Zeilen – 1) x (Spalten – 1) berechnet.)

Der 6. Schritt besteht nun darin, ein Signifikanzniveau so zu wählen, daß es den Erfordernissen der jeweiligen Einzeldisziplinen genügt.

In einem 7. Schritt wird die berechnete Prüfgröße mit dem ihr korrespondierenden Chi^2-Wert der Chi^2-Verteilung für die berechneten Freiheitsgrade und das gewählte Signifikanzniveau verglichen, woraus sich Rückschlüsse auf eine zufällige oder überzufällige Abweichung der Beobachtungswerte der obigen Tabelle von den theoretischen Erwartungswerten unter Gültigkeit der Nullhypothese ergeben.

Für unser Beispiel sehen wir eine Irrtumswahrscheinlichkeit von 0,001 als ausreichend an. Für dieses Signifikanzniveau und die Zahl der Freiheitsgrade (df = 1) ist, wie wir aus einer Chi^2-Tabelle (Grenzwerte der Chi^2-Verteilung) entnehmen können, eine berechnete Prüfgröße von mindestens 10,827 erforderlich, um die von den theoretischen Erwartungswerten – unter der Voraussetzung der Gültigkeit der Nullhypothese – abweichenden Beobachtungswerte als überzufällig und systematisch entstanden, bezeichnen zu können.

Wir führen für die obige Tabelle die einzelnen angegebenen Schritte durch:

Tab. 6. Beobachtungs- und Erwartungswerte einer Kreuztabelle

	♂		♀		
	f_o	f_e	f_o	f_e	Σ
Raucher	180	100	20	100	200
Nichtraucher	420	500	580	500	1000
Σ	600	600	600	600	1200

$$\chi^2 = \Sigma\frac{(f_0 - f_e)^2}{f_e} = \frac{(180 - 100)^2}{100} + \frac{(420 - 500)^2}{500} + \frac{(20 - 100)^2}{100} + \frac{(580 - 500)^2}{500} =$$

$\chi^2 = \mathbf{153{,}6}$

$df = 1$

$\chi^2_{df = 1;\ \alpha = 0{,}001} = 10{,}827 < 153{,}6$ daher Ablehnung der Nullhypothese, daß kein Zusammenhang zwischen Geschlecht und Rauchen bestehe.

Wir stellen fest, daß unser berechneter Chi²-Wert größer ist als der in der Tabelle ausgewiesene Grenzwert der Chi²-Verteilung, sodaß wir davon ausgehen können, daß wir mit einer Irrtumswahrscheinlichkeit von 0,1 % die oben formulierte Nullhypothese zurückweisen können. Es bestätigt sich also der theoretisch vermutete Sachverhalt, daß Männer häufiger rauchen als Frauen.

Der hier durchgeführte Chi²-Test ist ein universell einsetzbarer Signifikanztest, da er sowohl für eindimensionale wie auch (wie das obige Beispiel gezeigt hat) für zweidimensionale Häufigkeiten angewandt werden kann, weil er zudem zuläßt, daß die Variablen beliebig viele Merkmalsausprägungen haben und weil er auf dem niedrigsten Meßniveau liegt und somit für alle anderen Meßniveaus richtigerweise angewandt werden kann. Zwar sind für seine Anwendung bestimmte Restriktionen durchaus zu beachten (z.B. sollten die berechneten f_e-Werte für die einzelnen Zellen niemals kleiner als 5 sein), doch werden die Restriktionen hier nicht behandelt, weil der Praktiker in jedem Falle vor Anwendung des Tests sich zusätzlich anhand der Spezialliteratur weitergehend informieren sollte.

Hat man Meßwerte, die auf höherem Meßniveau liegen (z.B. Intervalldaten), so wird man oft nicht die gesamte Verteilung der jeweiligen Meßwerte überprüfen wollen, sondern z.B. nur das aus ihr berechnete arithmetische Mittel. Für solche Zwecke können der *t-Test* oder *z-Test* eingesetzt werden. (Der t-Test wird berechnet für eine Populationsgröße von $n \leqslant 30$, während der z-Test für Stichprobengrößen über 30 gilt.) Die Logik der beiden Tests ist jedoch gleich.

Nehmen wir an, wir hätten zwei Meßwertreihen, wovon sich die eine auf Personen bezieht, die unter 30 Jahre alt sind, und die andere auf solche, die älter sind. Gemessen wurde der Blutdruck.

H_0: Es zeige sich kein Unterschied im durchschnittlichen Blutdruck zwischen jüngeren und älteren Personen.

H_2: Es bestehe eine systematische Beziehung zwischen Alter und Blutdruck. (Man kann die Hypothese auch einseitig spezifizieren und einen einseitigen Test anwenden, indem man formulieren würde: Ältere haben einen höheren Blutdruck als Jüngere).

Als durchschnittlichen Blutdruck bei den Jüngeren hätten wir $\bar{x}_1 = 125$ und bei

den Älteren $\bar{x}_2 = 150$ ermittelt. Wir prüfen also nun die Frage, ob die offensichtlich vorhandene Differenz der beiden arithmetischen Mittel auch rein zufällig entstanden sein könnte. Es ist plausibel, daß sie desto weniger zufällig entstanden ist, je größer die Personenzahl war, an denen diese Meßwerte gewonnen wurden. Wir müssen daher in unsere Berechnungen die jeweiligen Stichprobengrößen miteinbeziehen. Nehmen wir daher an, wir hätten 100 jüngere Personen ($n_1 = 100$) und 80 ältere Personen ($n_2 = 80$) einer Blutdruckmessung unterzogen.

Die aufgrund der Meßwertverteilung zu berechnende Prüfgröße, die wiederum mit einer theoretischen Verteilung unter der Voraussetzung der Gültigkeit der Nullhypothese konfrontiert wird, erfolgt nach folgender Formel:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s} \cdot \frac{n_1 \cdot n_2}{n_1 + n_2}$$

(vgl. hierzu z.B. CLAUSS, G. und EBNER, H.: Grundlagen der Statistik, Thun und Frankfurt 1977, S. 206). Bevor die Prüfgröße selbst berechnet werden kann, muß die Standardabweichung s für beide Stichproben ermittelt werden, was nach der folgenden Formel geschieht:

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x}_1)^2 + \Sigma(x_i - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

Die nun gemäß der obigen Formel berechnete Prüfgröße wird mit einer aus der Normalverteilungs-Tabelle für ein vorgegebenes Signifikanzniveau (hier z.B. $\alpha = 0{,}05$) verglichen. Ist unser Prüfwert größer als der Tabellenwert, so können wir die Nullhypothese bei der vorgegebenen Irrtumswahrscheinlichkeit verwerfen. Ist der Wert kleiner, müssen wir die Nullhypothese beibehalten. Unsere Berechnung hat einen Prüfwert erbracht, der größer als der Tabellenwert ist, weswegen wir die Nullhypothese zurückweisen.

Da unsere Stichprobe größer als 30 war, konnten wir als Prüfverteilung die Normalverteilung verwenden. Hätten wir eine Stichprobe kleiner als 30 gehabt, so hätten wir die t-Verteilung heranziehen müssen. Die Berechnung der Prüfgröße selbst hätte sich nicht verändert; allerdings gilt für die t-Verteilung, was für die Chi^2-Verteilung schon gesagt wurde, nämlich daß sie nicht konstant ist, sondern von der Zahl der Freiheitsgrade abhängt. Für unterschiedliche Freiheitsgrade und Signifikanzniveaus gibt es also unterschiedliche Verteilungen und unterschiedliche Tabellen der kritischen Grenzwerte. Die Zahl der Freiheitsgrade berechnet sich aus der Summe der beiden Stichprobengrößen minus 2, also ($df = n_1 + n_2 - 2$). Die formal gleiche Prüfgröße für die z-Verteilung (Normalverteilung) und t-Verteilung ergibt sich daraus, daß mit zunehmender Stichprobengröße sich die t-Verteilung der Normalverteilung immer stärker annähert. Ab $n > 30$ kann man davon ausgehen, daß t-Verteilung und Normalverteilung fast deckungsgleich sind, sodaß auf die Berechnung der Freiheitsgrade verzichtet werden kann.

Für ein Signifikanzniveau von $\alpha = 0{,}05$ ergibt sich ein kritischer Tabellenwert von $z = 1{,}96$. Berechnen wir nun für unser Beispiel die Prüfgröße, wenn wir davon ausgehen können, daß $s = 25$. (Auf die Berechnung von s wird hier verzichtet, weil sonst die gesamte Meßwertverteilung hätte angegeben werden müssen, was für den Test selbst jedoch unerheblich ist).

$$
\begin{aligned}
n_1 &= 100 \\
n_2 &= 80 \\
\bar{x}_1 &= 125 \\
\bar{x}_2 &= 150 \\
s &= 25
\end{aligned}
$$

$$
\begin{aligned}
z &= \frac{\bar{x}_1 - \bar{x}_2}{s} \quad \frac{n_1 \cdot n_2}{n_1 + n_2} \\
&= \frac{125 - 150}{25} \quad \frac{100 \cdot 80}{180} \\
&= 44{,}4
\end{aligned}
$$

Da unsere Prüfgröße weit größer als der kritische Tabellenwert ist, können wir die Nullhypothese mit einer Irrtumswahrscheinlichkeit von 5% ablehnen.
(Vgl. zu den Signifikanztests auch die noch folgenden Beispiele zu den Kapiteln 2, 3 und 4).

1.6.4 Korrelationsmaße

Die im vorigen Abschnitt angestellten Überlegungen und Berechnungen zur Signifikanz sollten die Frage beantworten, ob zwischen zwei Variablen eine Beziehung besteht und ob diese Beziehung zufällig oder signifikant ist. Dabei war völlig außer Betracht geblieben, wie stark die Beziehung zwischen den beiden, infragestehenden Variablen ist. Diese Frage kann auch nicht mit Signifikanztests beantwortet werden, sondern hierfür sind spezielle Korrelationsmaße entwickelt worden. Diese Korrelationsmaße haben, bzw. sollten im Regelfalle die folgenden Eigenschaften aufweisen:

1. Besteht kein Zusammenhang zwischen den Variablen, so sollten sie den Wert 0 annehmen.

2. Besteht ein absoluter Zusammenhang, also eine perfekte Korrelation, so sollen sie den Wert |1| annehmen.

3. Liegt mindestens ordinales Meßniveau der Variablen vor, so sollte es möglich sein, mit Hilfe des Vorzeichens des Korrelationskoeffizienten eine negative oder positive Korrelation zum Ausdruck zu bringen. (Positive Korrelation wäre dann gegeben, wenn zunehmende Größe der Merkmalsausprägungen der einen Variablen mit zunehmender Größe der anderen Variablen einhergeht, bzw. allgemein gesagt, ein gleichsinniges Verhältnis besteht. Nimmt die eine Variable zu, während die andere Variable abnimmt (gegensinniges Verhältnis), so sollte dies durch einen negativen Korrelationskoeffizienten als negativer Zusammenhang gedeutet werden.)

4. Genügen die Korrelationskoeffizienten diesen Kriterien, dann sind auch unterschiedliche Koeffizienten im Hinblick auf deren Höhe miteinander vergleichbar (hierzu wären allerdings noch weitere Ausführungen zu machen, denn die Logik der Verfahren bzw. Modelle der einzelnen Korrelationskoeffizienten kann sehr unterschiedlich sein, obgleich sie den formal gleichen Bedingungen im Hinblick auf deren Extremwerte genügen können).

Die Fülle der in der Literatur entwickelten Korrelationskoeffizienten kann natürlich nicht in dieser allgemeinen Einführung behandelt werden. Wir werden uns vielmehr auf 4 Korrelationskoeffizienten beschränken. Die ersten beiden sind solche, die auf nominalem Meßniveau liegen, der dritte mißt auf ordinalem Niveau, während der vierte ein Intervallniveau der Meßdaten voraussetzt. Wie bereits im Abschnitt 1.5 ausgeführt, können solche Korrelationskoeffizienten, die auf niedrigem Meßniveau liegen, auch auf Daten angewandt werden, die auf höherem Meßniveau liegen, jedoch nicht umgekehrt. Von daher kommt natürlich Koeffizienten auf den unteren Meßniveaus für eine allgemeine Einführung besondere Bedeutung bei.

Wir greifen auf das obige Beispiel zurück, bei dem die Frage geprüft wurde, ob Männer häufiger Raucher sind als Frauen. Beide Variablen wurden nominal gemessen, und wir haben eine *Kontingenztafel* (mit 4 Zellen) erhalten. Für den Fall nominalen Meßniveaus, wo beide Variablen jeweils nur in zwei Merkmalsausprägungen auftreten *(dichotomisierte Merkmale)*, ist der sog. *φ-Korrelationskoeffizient* nach PEARSON berechenbar. Hat man den Chi^2-Test als Signifikanztest bereits durchgeführt, so kann die so gewonnene Prüfgröße unmittelbar in die Berechnung von φ Eingang finden, denn die Formel nimmt folgendes Aussehen an:

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

Dieser Formel ist sofort anzusehen, daß φ offensichtlich dann 0 wird, wenn die Prüfgröße $Chi^2 = 0$ ist, d.h. also, wenn schon der Signifikanztest belegt hat, daß keine Beziehung zwischen den Variablen besteht, daß also die empirische der theoretischen Verteilung unter der Gültigkeit der Nullhypothese entspricht. φ nimmt den Wert 1 dann an, wenn die Prüfgröße Chi^2 genauso groß ist, wie die in die Tabelle eingehende Fallzahl. Dieses wäre der maximale Zusammenhang. Die konkrete Berechnung von φ für unser Beispiel bringt einen Koeffizienten von $\varphi = 0{,}36$, sodaß von einem mäßigen bis brauchbaren Zusammenhang zwischen den beiden Variablen ausgegangen werden kann.

Die Logik des Korrelationskoeffizienten φ kann jedoch besser verdeutlicht werden, wenn man eine andere Formel heranzieht, die zu demselben Resultat führt, nämlich:

$$\varphi = \frac{ad - bc}{\sqrt{s_1 \cdot s_2 \cdot s_3 \cdot s_4}}$$

wobei gilt:

			Σ
	a	b	s_3
	c	d	s_4
Σ	s_1	s_2	

Eine perfekte Korrelation wäre offensichtlich gegeben, wenn auf die Vierfelder-Tafel bezogen, nur jeweils die Zellen einer Diagonale besetzt wären, denn genau dann fallen niedrige Werte der einen Variablen mit niedrigen der anderen und hohe Werte der einen mit hohen Werten der anderen Variablen zusammen. Im Falle einer negativ-perfekten Korrelation würde die andere Diagonale besetzt sein.

Tab. 7. Perfekte Korrelationen in Kreuztabellen

Tab. 7a: perfekter positiver Zusammenhang

	ja	nein	Σ
ja	50	–	50
nein	–	80	80
Σ	50	80	130

Tab. 7b: perfekter negativer Zusammenhang

	ja	nein	Σ
ja	–	40	40
nein	70	–	70
Σ	70	40	110

Der Korrelationskoeffizient prüft also offenkundig, wieviele Meßwerte vorhanden sind, die eine positive und wieviele eine negative Korrelation indizieren; man setzt diese zueinander in Beziehung, wie es die Formel ausweist. Man spricht auch von konkordanten und diskordanten Paaren:

Tab. 8. Konkordante und diskordante Paare

	+	–	Σ
+	++	–+	
–	+–	––	
Σ			

konkordante Paare: ++; ––
diskordante Paare: –+; +–

Da der Korrelationskoeffizient φ jedoch nur für jeweils dichotomisierte Merkmale, also nur für Vierfelder-Tafeln gilt, sind seine Anwendungsmöglichkeiten relativ restriktiv. Wir müssen uns daher nach einem Korrelationskoeffizienten umsehen, der für jede, nur mögliche Tabellengröße geeignet erscheint. Hierzu kann der *Kontingenzkoeffizient C* herangezogen werden. Er ist – wie jeder Korrelationskoeffizient – eine Maßzahl für das Ausmaß der Stärke zwischen zwei Variablen. Er ist anwendbar, wenn die beiden Variablen nominale Informationen enthalten, liegt also ebenfalls auf niedrigstem Meßniveau und kann für höhere Meßniveaus gültig Verwendung finden. Unabhängig davon, wie die Merkmale und die Merkmalsausprägungen in einer Tabelle angeordnet werden, der Kontingenzkoeffizient liefert immer das gleiche Maß.

Tab. 9. Berechnung eines Kontingenzkoeffizienten unabhängig von der Ordnung der Variablen in der Kreuztabelle

Tab 9a)

	ja	unent-schieden	nein	Σ
ja	6	24	18	48
nein	15	12	6	33
Σ	21	36	24	81

$\chi^2 \approx 23{,}9$; df = 2;
C = 0,48

Tab. 9b)

	nein	unent-schieden	ja	Σ
ja	18	24	6	48
nein	6	12	15	33
Σ	24	36	21	81

Der Kontingenzkoeffizient baut in seiner Berechnung ebenfalls auf dem Chi^2-Test auf und errechnet sich nach folgender Formel:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Die Berechnung des Kontingenzkoeffizienten ergibt für die obigen Tabellen einen mittelstarken Zusammenhang von 0,48 zwischen den beiden Variablen. Allerdings gilt folgendes dabei zu berücksichtigen: Der Kontingenzkoeffizient hat die unangenehme Eigenschaft, daß sein maximal möglicher Wert von der Größe der Tabelle abhängig ist und nicht in jedem Falle bei |1| liegt, wie das die anderen Korrelationskoeffizienten tun. Für den Fall, daß die zugrundeliegende Kreuztabelle quadratisch ist (also gleiche Anzahl von Zeilen und Spalten), läßt sich allerdings der Maximalwert des Kontingenzkoeffizienten nach der Formel berechnen:

$$\sqrt{\frac{k-1}{k}}$$

wobei k die Zahl der Spalten oder Zeilen der quadratischen Matrix ist. So liegt der maximale Wert des Kontingenzkoeffizienten für eine Vierfelder-Tafel bei 0,707 und der einer 3x3-Tabelle bei 0,816. Hat man nun keine quadratische Tabelle, sondern z.B. eine 3x2-Tabelle, so kann man den maximalen Wert annähernd durch die Vermittlung zwischen dem Maximalwert einer 2x2- und einer 3x3-Tabelle berechnen, nämlich als Mittel aus 0,707 und 0,816, was einen Maximalwert von annähernd 0,762 entspricht. Steht der jeweilige Maximalwert zur Verfügung, so kann man den aus der Tabelle errechneten Wert des Kontingenzkoeffizienten mit dem

reziproken Maximalwert multiplizieren („gewichteten") und erreicht damit, daß der so berechnete Wert sein Maximum bei |1| hätte. Für die obige Tabelle ergibt sich folgende Gewichtung:

$$C_{korr} = C \cdot \frac{1}{C_{max}}$$

$$C_{korr} = 0{,}48 \cdot \frac{1}{0{,}762} = 0{,}63$$

Der so korrigierte Kontingenzkoeffizient ist natürlich größer als der ursprüngliche. Wir können daher die Stärke des Zusammenhangs zwischen den in der Tabelle kombinierten Variablen noch günstiger beurteilen.

Der bekannteste Korrelationskoeffizient auf ordinalem Meßniveau ist der *SPEARMAN'SCHE Rangkorrelationskoeffizient* ρ. Eine bestimmte Anzahl von n Individuen wird nach zwei Variablen geordnet (ordinales Meßniveau), z.B. Zulassung zum Medizinstudium nach den Kriterien „Abiturnotendurchschnitt und Wartezeit". Die Korrelation zwischen diesen Variablen wäre dann perfekt, wenn Abiturnotendurchschnitt und Wartezeit für alle Bewerber in der aufgestellten Rangfolge übereinstimmen würden. Es erscheint daher logisch, die möglichen Differenzen zwischen diesen beiden Reihen als ein Maß der Gleichheit bzw. Ungleichheit für die Korrelation zu nehmen. Je größer diese Differenzen sind, desto geringer ist die Korrelation zwischen beiden Variablen. Hierbei wäre jedoch wieder zu berücksichtigen, daß negative Differenzwerte die positiven Werte eliminieren könnten, wollten wir ein Maß über die beiden Rangreihen hinweg finden. Quadriert man jedoch diese Differenzen und summiert sie auf, so können diese unerwünschten Effekte ausgeschaltet werden.

Eine zweite Schwierigkeit würde darin bestehen, daß die absolute Höhe des Koeffizienten von der Anzahl der Merkmalsausprägungen abhängig werden würde (in unserem Beispiel der Anzahl der Bewerber für einen Studienplatz). Um diese und weitere Schwierigkeiten zu vermeiden, wurde der Rangkorrelationskoeffizient von SPEARMAN entwickelt. Seine Ableitung soll hier nicht vorgeführt werden, sie kann in jedem statistischen Lehrbuch nachgelesen und nachvollzogen werden. Die Formel für die Anwendung des SPEARMAN'schen Rangkorrelationskoeffizienten lautet:

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n^3 - n}$$

Das folgende Beispiel soll die Anwendung des SPEARMAN'schen Rangkorrelationskoeffizienten ρ demonstrieren.

Tab. 10. Berechnung des Rangkorrelationskoeffizienten ρ

V_p	x_i	y_i	x_i'	y_i'	$d_i = x_i' - y_i'$	d_i^2
A	2,3	5	3	2	+1	1
B	2,1	7	2	3	−1	1
C	1,9	4	1	1	±0	0
D	2,4	9	4	4	±0	0
E	3,0	15	7	7	±0	0
F	2,5	12	5	6	−1	1
G	2,6	10	6	5	+1	1
Σ						4

$n = 7$

x_i = Abiturnote
y_i = Wartezeit in Monaten
x_i' = Rangreihe der Abiturnoten
y_i' = Rangreihe der Wartezeiten

$$\rho = 1 - \frac{6d_i^2}{n^3 - n} = 1 - \frac{24}{336} =$$

$$\rho = 0{,}93$$

Wir stellen also eine Korrelation zwischen den beiden Variablen in Höhe von 0,93 fest. Da ρ auf ordinalem Meßniveau liegt, kann es sowohl positive wie negative Werte annehmen und mithin positive und negative Korrelationen indizieren. Der Maximalwert liegt jeweils bei $|1|$. Die von uns festgestellte Stärke des Zusammenhangs kann daher als fast perfekt bezeichnet werden.

Der letzte hier zu besprechende Korrelationskoeffizient liegt auf mindestens Intervallniveau und setzt bei seiner Anwendung voraus, daß die beiden miteinander kombinierten Variablen jeweils normalverteilt sind. Der Korrelationskoeffizient nach BRAVAIS-PEARSON errechnet sich nach der folgenden Formel:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}}$$

Zu seiner Berechnung sei das folgende Beispiel herangezogen, wo die beiden Variablen, Alter und Weitsichtigkeit (operationalisiert, z.B. über die Gläserstärke der Brillen in Dioptrien ohne Vorzeichen) miteinander korreliert werden.

Tab. 11. Berechnung des Produkt-Moment-Korrelationskoeffizienten

V_p	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
A	50	1,7	−10	−0,3	+3	100	0,09
B	52	1,5	− 8	−0,5	+4	64	0,25
C	60	1,9	± 0	−0,1	±0	0	0,01
D	65	2,0	+ 5	±0	±0	25	0,00
E	55	2,0	− 5	±0	±0	25	0,00
F	70	2,5	+10	+0,5	+5	100	0,25
G	53	2,1	− 7	+0,1	−0,7	49	0,01
H	75	2,3	+15	+0,3	+4,5	225	0,09
Σ	480	16			15,8	588	0,7

x_i = Alter
y_i = Weitsichtigkeit

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{480}{8} = 60$$

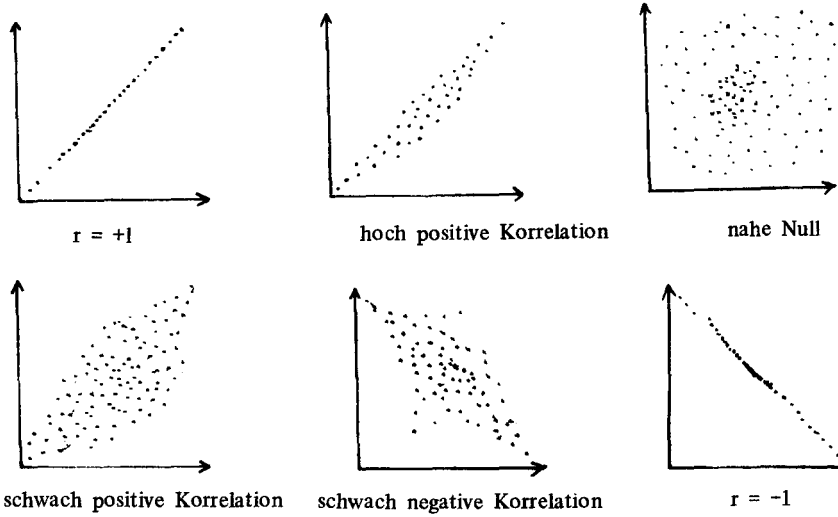
$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{16}{8} = 2$$

In den entsprechenden Spalten der obigen Tabelle wurden alle, für die Anwendung der obigen Formel notwendigen Größen bereits berechnet, sodaß wir nur einzusetzen brauchen;

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} = \frac{15{,}8}{\sqrt{588 \cdot 0{,}7}} = 0{,}78$$

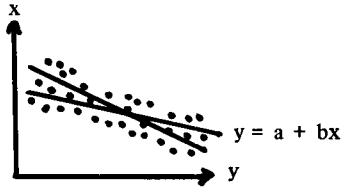
Wir stellen also einen recht hohen Zusammenhang zwischen Alter und Weitsichtigkeit fest.
Auch der BRAVAIS-PERSON'sche Produkt-Moment-Korrelationskoeffizient kann positive und negative Werte annehmen und hat seine Grenzwerte bei – 1 und + 1. Die folgende Graphik möge zur Verdeutlichung dessen, was Korrelationskoeffizienten auszusagen vermögen, dienen:

Abb. 7. Korrelation



Stellt man eine Korrelation grafisch unter der Voraussetzung der Metrisierbarkeit der Variablen dar und berechnet jene Gerade, die die Meßpunkte der Variablen optimal wiedergibt (Regressionsgerade), was mit Hilfe der Methode der kleinsten Quadrate geschieht, so können zwei Regressionsgeraden angegeben werden.

Abb. 8. Korrelationsdarstellungen als Regressionsgeraden



Die zwei Geraden ergeben sich daraus, daß einmal x und einmal y als unabhängige Variable gesehen wird. Mit Hilfe der Steigungen (b) der beiden Regressionsgeraden kann nun der Korrelationskoeffizient r berechnet werden als:

$$r = \sqrt{b_{xy} \cdot b_{yx}}$$

Der so errechnete Korrelationskoeffizient r unterscheidet sich nur im Berechnungsmodus von dem vorher dargestellten. Eine dritte Methode, denselben Korrelationskoeffizienten zu berechnen, wird bei der Varianzanalyse in dem Abschnitt 2.2.3 vorgestellt werden, um damit die Möglichkeit zu eröffnen, das Konzept der Korrelation auf unterschiedlichste Weise kennen und verstehen zu lernen.

Eine vollständig positive Korrelation ist dann gegeben (vgl. Abb. 7), wenn die beiden Meßreihen der Variablen x und y auf einer Geraden liegen. Dies hängt damit zusammen, daß die Korrelationskoeffizienten eine *lineare Beziehung der Variablen* unterstellen und nur eine solche messen. So wäre durchaus vorstellbar, daß eine Beziehung perfekt nonlinear ist, die Berechnung des Korrelationskoeffizienten jedoch einen Wert von 0, also keine Beziehung zwischen den Variablen ergibt.

Eine positive Korrelation liegt dann vor, wenn die Zunahme der einen Variablen mit einer Zunahme der anderen verknüpft ist bzw. Abnahme der einen mit einer Abnahme der anderen einhergeht.

Eine negative Korrelation ist gegeben, wenn eine Zunahme in x von einer Abnahme in y begleitet wird und vice versa.

Eine Nullkorrelation ist immer dann gegeben, wenn keine lineare Beziehung zwischen den beiden Variablen festgestellt werden kann, d.h. die Variationen der Merkmalsausprägungen der einen Variablen völlig unabhängig von der der anderen erfolgen.

Bei der Interpretation der Korrelationskoeffizienten ist zu berücksichtigen, daß diese tatsächlich nur einen Zusammenhang, soweit er linear ist, wiedergeben und daß sie keine Aussagen über die Art des zugrundeliegenden Zusammenhangs machen können. Zwar kann man zweifelsfrei davon ausgehen, daß ein maximaler Korrelationskoeffizient von + 1 oder − 1 zu erwarten wäre, wenn kausal-deterministische Beziehungen unterstellt sind. Andererseits kann jedoch nicht aus einem maximalen Korrelationskoeffizienten auf ein kausales Verhältnis geschlossen wer-

den. Die Korrelationskoeffizienten liefern also nur Anhaltspunkte für eine gültige Interpretation der Beziehung zwischen zwei Variablen.

Auch erlauben Korrelationskoeffizienten wie Signifikanztests keine Aussagen darüber, in welcher Richtung die Beziehung zwischen den Variablen verläuft, welches also die unabhängige und welches die abhängige Variable ist. Daraus läßt sich grundsätzlich folgern, daß alle statistischen Verfahren Hilfsmittel sind, denen nur beschränkter Aussagewert zukommt. Ihre Aussagefähigkeiten und Interpretationsmöglichkeiten ergeben sich ausschließlich aus den zugrundegelegten statistischen Modellen, wobei die Basis aller empirischer Forschung grundsätzlich die Theorie bleibt. Sie determiniert und entscheidet, welche Erkenntnisse überhaupt möglich sind. Die Statistik wird nur dazu eingesetzt, die Fülle der empirisch gewonnenen Informationen auf ein menschlich überschaubares Maß zu reduzieren und intersubjektiv abgesicherte Überprüfungs- und Interpretationsmöglichkeiten zu schaffen.

Ergänzende und vertiefende Literatur:

1. Methodologie und Methoden:

FRIEDRICHS, J., Methoden der empirischen Sozialforschung, Reinbek 1974

HARTMANN, H., Empirische Sozialforschung, München 1970

MAYNTZ, R. u.a., Einführung in die Methoden der empirischen Soziologie, Köln und Opladen 1972

2. Kritischer Rationalismus:

POPPER, K.R., Logik der Forschung, Tübingen 1966

PRIM, R. und TILMANN, H., Grundlagen einer kritisch-rationalen Sozialwissenschaft, Heidelberg 1977

3. Statistische Methoden:

CLAUSS, G. und EBNER, H., Grundlagen der Statistik, Thun und Frankfurt 1977

KRIZ, J., Grundlagen der Statistik, Reinbek 1973

OPPEL, U.G. und RÜGER, B., Biomathematik, medizinische Statistik und Dokumentation, München 1975

2 DAS EXPERIMENT

Wenn das Streben nach Erkenntnis, d.h. nach Erweiterung und nach „Bestätigung" des Wissens eines der wesentlichsten Ziele der Wissenschaften ist, so bedarf es zur Zielrealisierung bestimmter Vorschriften darüber, auf welche Weise und mit welchen Methoden dieses Ziel erreicht werden kann. So hat jede einzelwissenschaftliche Disziplin eigene Methoden und Techniken entwickelt, die ihrem selbstgestellten Anspruch auf Wissenschaftlichkeit genügen sollen.

Neben diesen einzelwissenschaftlichen Methoden und Techniken scheint es einige grundlegende Methoden zu geben, die für viele (wenn nicht gar für alle) Einzelwissenschaften in gleicher, ähnlicher oder modifizierter Form anwendbar sind, weil deren Logik und Vorgehensweise so basal sind, daß behauptet werden könnte, diese Methoden setzten eher selbst das Kriterium der Wissenschaftlichkeit als daß die Methoden aus unabhängigen Wissenschaftskriterien abgeleitet würden. Eine solche grundlegende Methode mit Anspruch auf Wissenschaftlichkeit in Logik und Vorgehensweise ist das Experiment, das ob seiner umfassenden Potenz sowohl in den Naturwissenschaften, wie auch in den Sozialwissenschaften eingesetzt wird.

Da der Begriff des Experiments alltagssprachlich, aber auch in der Wissenschaftssprache mit unterschiedlichen Konnotationen belegt sein kann, soll der wissenschaftlich engere Begriff des Experiments präzisiert und von den anderen Verständismöglichkeiten des Begriffsinhalts abgehoben werden. Der wissenschaftliche Begriff des Experiments unterscheidet sich von dem umgangssprachlichen des Versuchs oder einer Neuerung (vgl. das politische Experiment) ebenso, wie von dem alltagssprachlichen Experimentieren als einem versuchsweisen Verfahren des Ausprobierens (trial and error). Der hier im engeren wissenschaftlichen Sinne gebrauchte Begriff des Experiments als Methode hebt sich aber auch von dem gelegentlich in der Wissenschaft gebrauchten Begriff des Gedankenexperiments, bei dem eine gedankliche Durchdringung eines Objektbereiches im Sinne eines Verfahrens, einer Beweisführung gemeint ist, ab. Der hier im weiteren zu präzisierende Begriff des Experiments unterscheidet sich von den oben genannten Vorstellungsinhalten nicht zuletzt durch das Kriterium der Wissenschaftlichkeit. Wissenschaftliches Arbeiten setzt in allen Disziplinen voraus, daß das Suchen nach Erkenntnis nicht dem Zufall überlassen bleibt, sondern systematisch betrieben wird.

Diese normative Festsetzung impliziert aber keineswegs, daß nur auf systematischem Wege Erkenntnis gewonnen werden kann. So sind eine Fülle von naturwissenschaftlichen Erkenntnissen durch Zufall als *Serendipität,* als Abfall- oder Nebenprodukte bei Forschungen entstanden, die zwar selbst systematisch betrieben wurden, die aber anderen Zielen und Zwekken dienten. Auch die zufällige Erkenntnis wird im Regelfalle im Rahmen von systematisch angestrebter Erkenntnis gewonnen.

Das Insistieren auf systematischer Suche nach Erkenntnis und das normative Ausschließen zufälliger Erkenntnis als einem wesentlichen Kriterium wissenschaftlichen Arbeitens schlägt notwendigerweise voll auf die einzusetzenden Methoden durch. An einem einfachen Beispiel kann dieser Sachverhalt demonstriert werden (vgl. KÖNIG, Beobachtung und Experiment, S.31). Dieses Beispiel soll dazu dienen,

die Begrenztheit der Erkenntnismöglichkeiten einer „anekdotischen Beobachtung" aufzuzeigen, um daraus abzuleiten, daß eine unsystematische, zufällige, spontane Beobachtung durchaus ein erster Schritt auf dem Wege nach wissenschaftlicher Erkenntnis sein kann, daß aber die Unsystematik im weiteren Verlauf durch systematische, standardisierte Beobachtung abgelöst werden muß:

Ein erstmalig die Schweiz Bereisender wurde in einem Restaurant von einem rothaarigen Kellner bedient. Als derselbe am nächsten Tag in einem anderen Restaurant nochmals einem rothaarigen Kellner begegnete, berichtete er nach der Rückreise in sein Heimatland: „Alle Schweizer Kellner haben rote Haare".

Welche Einwände sind nun gegen diese Generalisierung vorzubringen? Zunächst einmal wird man entgegenhalten müssen, daß die Beobachtung von zwei Kellnern wohl quantitativ nicht ausreichend sein kann, um auf alle Schweizer Kellner zu schließen. Zum zweiten wird man einwenden müssen, daß, was als Schweizer Spezifikum hingestellt wurde, nicht notwendigerweise ein Schweizer Charakteristikum sein muß, weil durchaus vorstellbar wäre, daß in anderen Ländern ebenfalls Kellner mit roten Haaren anzutreffen sind. Die quantitative Ausweitung des Untersuchungsmaterials hätte also erfolgen müssen, um die Aussage wissenschaftlich absichern zu können. Hierbei wäre wahrscheinlichkeitstheoretisch zu erwarten gewesen, daß nicht nur rothaarige Kellner angetroffen worden wären, womit der generalisierende Satz: „Alle Schweizer Kellner haben rote Haare", der nur auf zwei zufälligen Beobachtungen beruhte, falsifiziert worden wäre.

Aber die quantifizierende Ausweitung der Untersuchungspopulation allein würde nicht ausreichen, die unwissenschaftliche, spontane Beobachtung zu einer wissenschaftlichen und mehr oder weniger standardisierten zu machen. So wäre durchaus vorstellbar, daß die Ausweitung der zufälligen Beobachtung ausschließlich weitere rothaarige Kellner zum Resultat gehabt hätte. Vielmehr bedarf auch die ausgeweitete, spontane Zufallsbeobachtung einer Systematisierung und Standardisierung der Art, daß die Überprüfung einer solchen Aussage an einer gezielt und systematisch ausgewählten Population vorgenommen wird, wobei die Systematik jede Form von Subjektivität in der Auswahl ausschließen sollte (z.B. nach wahrscheinlichkeitstheoretischen Gesichtspunkten).

Diese auswahltechnischen Gesichtspunkte müssen aber, um wissenschaftlichen Kriterien zu genügen, noch durch eine gewisse Form der Intersubjektivität der Beobachtung selbst ergänzt werden. So könnte man sich als Extremtypus vorstellen, daß der Beobachter farbenblind wäre und die Kellner nur mangels Farbunterscheidungsvermögens als rothaarig klassifiziert werden. Es wäre also mindestens ein zweiter Beobachter notwendig, um zu einigermaßen gesicherten Aussagen zu kommen. Eine solche gegenseitige Kontrolle ist nicht notwendigerweise an derselben Population und zum selben Zeitpunkt erforderlich. Vielmehr würde es dem Kriterium der Wissenschaftlichkeit voll genügen, wenn man davon ausgehen kann, daß ein anderer Beobachter bei gleichem Objektbereich zu demselben Resultat kommt.

Die Loslösung von der Subjektivität der Beobachtung hin zur *Intersubjektivität* ist eines der wesentlichsten Kriterien für wissenschaftliches Arbeiten (sofern dieses empirisch angelegt ist) und unterscheidet die unsystematische, spontane Zufallsbeobachtung von der systematischen, standardisierten experimentellen Beobachtung.

Neben der Intersubjektivität muß auch noch das Kriterium der Zielgerichtetheit der Beobachtung in die Diskussion eingebracht werden. Standardisierte oder experimentelle Beobachtung setzt voraus, daß der Beobachter weiß, was zu beobachten ist, weil ohne dieses Wissen eine Standardisierung und Systematisierung der Beobachtung nicht möglich erscheint. Das Wissen um die zu beobachtenden Faktoren oder Variablen ergibt sich als zielgerichtetes Überprüfen der einmal aufgestellten Hypothesen. Sind keine Hypothesen zu einem Objektbereich formuliert, kann man auch nicht wisssen, welche Sachverhalte im einzelnen zu beobachten sind. In diesem Falle wäre die Beobachtung auf rein zufällige Resultate angewiesen.

Das Aufstellen von Hypothesen ist allerdings keine hinreichende Bedingung dafür, alle relevanten Sachverhalte erhoben zu haben, denn in der theoretisch-hypothetischen Durchdringung eines Objektbereiches können durchaus wichtige Gesichtspunkte vergessen, vernachlässigt und übersehen worden sein. Mit dem Aufstellen von Hypothesen ist aber gesichert, daß sich die Beobachtung auf die in den Hypothesen genannten Variablen bezieht und diese somit systematisch überprüft werden können.

Unsystematisches Arbeiten unterscheidet sich demnach vom systematischen durch folgende Kriterien:

1. Die Untersuchungseinheiten werden nach einem wissenschaftlichen Design (wie auch immer technisch realisiert) und nicht zufällig im Sinne von willkürlich gewonnen.

2. Die Subjektivität der Beobachtung wird abgelöst durch Intersubjektivität, d.h., daß verschiedene Beobachter bei gleichen Sachverhalten (Situationen) zu dem gleichen Resultat kommen.

3. Dies setzt voraus, daß die Beobachtung zielgerichtet entlang von Hypothesen erfolgt, die vor der Beobachtung (z.B. durch die Konstruktion eines spezifischen Erhebungsinstrumentes) ermöglichen.

2.1 Methode und Logik des Experiments

Allgemeine Funktion jeder wissenschaftlichen Methode ist, den Erkenntnisstand zu erweitern. Dabei dienen die Methoden im speziellen dazu, aufgestellte theoretische Hypothesen empirisch auf deren Richtigkeit hin zu überprüfen. Von dem Experiment als wissenschaftlicher Methode nimmt man nun im besonderen an, daß es in der Lage wäre, Hypothesen, die kausale Beziehungen zwischen Variablen aufzeigen, zu überprüfen. Kausale Beziehungen werden verstanden als Ursache-Wirkungs-Relation mit deterministischem Charakter. Um solche Kausalbeziehungen zu entdecken bzw. als zutreffend auszuweisen, hat John Stewart MILL um die Jahrhundertwende ein System der Logik entwickelt, indem er 5 Möglichkeiten kausaler Schlußfolgerungen beim Experiment aufzeigt. Von diesen 5 Möglichkeiten sollen hier zwei kurz genannt werden, um paradigmatisch an ihnen aufzuzeigen, daß eine kausale Zurechnung von Faktoren im Experiment durchaus problematisch ist und daß die MILL'schen Methoden dem heutigen Stand der Wissenschaft nicht mehr adäquat sind.

Bei der *Methode der Übereinstimmung* geht MILL davon aus, daß das mehrmalige, gemeinsame Auftreten jeweils einer unabhängiger mit jeweils derselben abhängigen Variablen dafür spricht, daß die unabhängige Variable die Ursache für die abhängige Variable ist.

Abb. 9: Methode der Übereinstimmung:*

I : a, b, c, d ⟶ y

II: ~a, ~b, ~c, d ⟶ y

* Das Zeichen „ ~ " bedeutet die Negation eines Sachverhaltes, also z.B. dessen Nichtvorliegen.

Es ist aus dem obigen Schema unmittelbar einsichtig, daß diese Schlußfolgerung nicht notwendigerweise zutreffen muß. So wäre denkbar, daß die folgende Konfiguration eine gültigere Beziehung zwischen den genannten Variablen darstellt und somit die Methode der Übereinstimmung als ungenügend erscheint, kausale Beziehungen festzustellen.

Abb. 10: Problem der Methode der Übereinstimmung

I : a, b, c, d ⟶ y
II: ~a, ~b, ~c, x ⟶ y oder
III: ~a, ~b, ~c, x, d ⟶ y

Bei der *Methode der Differenz* wird davon ausgegangen, daß unter sonst ceteris-paribus-Bedingungen das Auftreten einer unabhängigen Variablen x mit dem Auftreten einer abhängigen Variablen y korrespondiert, während beim Nichtvorhandensein von x auch die abhängige Variable y nicht auftritt.

Abb. 11: Methode der Differenz:

I : a, b, c, d ⟶ y
II: a, b, c, ~d ⟶ ~y

Diese Methode der Differenz ist günstiger zu beurteilen als die der Übereinstimmung und entspricht annäherungsweise auch den klassischen Versuchsanordnungen des Experiments (vgl. hierzu 2.1.4). Gleichwohl muß gegen diese Methode eingewandt werden, daß sie zwar in der Lage ist, das Kovariieren von zwei Merkmalen anzuzeigen, daß sie aber keinen Anhaltspunkt dafür liefern kann, welche Variable die Ursache und welche die Wirkung darstellt (dies kann bestenfalls durch zusätzliche, plausibel-theoretische Überlegungen geschehen). Zudem muß berücksichtigt werden, daß es selten gelingen wird, Variablenkonstellationen so zu konstruieren, wie es die ceteris-paribus-Bedingungen in der Methode der Differenz verlangen: Die Untersuchungspersonen dürfen sich ja nur in einem Merkmal, nämlich der unabhängigen Variablen unterscheiden. Diese Schwierigkeit kann aber dadurch kompensiert werden, daß man nicht wie MILL Gleichheit aller anderen Variablen unterstellt, sondern deren Zufallsstreuung annimmt bzw. konstruiert (vgl. hierzu 2.1.3).

Mit dem Experiment wird angestrebt, kausale Beziehungen, also ein Ursache-Wirkungs-Verhältnis zwischen Faktoren festzustellen. Kausale Beziehungen als Ursache-Wirkungs-Relationen müssen dabei (wenn von wissenschaftstheoretischen Präzisierungen hier abgesehen wird) dem Alltagsdenken entsprechend folgende Bedingungen erfüllen:

1. Eine Wirkung folgt immer einer Ursache und nicht umgekehrt.
2. Eine kausale Beziehung ist deterministisch, d.h. sie gilt immer und in jedem Falle.
3. Ursache und Wirkung liegen räumlich beieinander.

Inwieweit ist nun das Experiment geeignet, diese drei Bedingungen an einem konkreten, zu überprüfenden Sachverhalt auszumachen? Statistisch und technisch gesehen, gibt es keine Methode, die in der Lage wäre, anzugeben, welche Variable eine Ursache und welche eine Wirkung darstellt. Statistisch stellen sich solche Relationen immer als *Kovariationen* heraus. Die Entscheidung darüber, wie die Richtung einer solchen Relation geartet ist, kann immer nur aufgrund theoretischer oder plausibler Überlegungen getroffen werden. (So wird nicht bestritten werden können, daß das Geschlecht unter sonst gleichen Bedingungen die Einkommenshöhe determiniert und nicht das Einkommen das Geschlecht.) Das Experiment als Methode und Versuchsanordnung ist daher auch nicht fähig, die Qualitäten von Ursache und Wirkung den jeweils untersuchten Variablen zuzuschreiben.

Das Problem einer *deterministischen Beziehung* als einem Element der Kausalität kann jedoch durchaus experimentell gelöst werden. Hierzu das folgende Beispiel:

Als Hypothese sei angenommen, daß eine bestimmte Form der sexuellen Delinquenz (z.B. Notzuchtverbrechen) durch Kastration des jeweiligen Täters dazu führt, daß diese nicht wieder auftritt. Nehmen wir weiter an, wir würden diese Hypothese experimentell überprüfen wollen, indem wir zwei Gruppen von Sexualdelinquenten bilden, wobei die eine Gruppe kastriert wird und die andere nicht. (Ethische Bedenken sind hier für die Illustration des Determinismusproblems nicht relevant.) Aufgrund unseres Experiments würden wir, sofern die Beziehungen der Hypothese deterministisch verläuft, erwarten können, daß nach einem gewissen Zeitraum alle Kastrierten nicht rückfällig geworden sind, während die Nichtkastrierten alle erneut Straftaten begangen haben. Eine tabellarische Darstellung hätte also in etwa die folgende Form:

Tab. 12. Deterministische Relationen im Experiment (perfekte Korrelation)

	Kastration	~ Kastration	Σ
~ Rückfall	80	—	80
Rückfall	—	100	100
Σ	80	100	180

Wir stellen also in unserem Experiment eine deterministische Beziehung zwischen Kastration und Rückfallhäufigkeit fest (vollständige positive Korrelation, ausschließlich konkordante Paare). Aufgrund unseres theoretischen Wissens werden wir in der Lage sein, die Kastration als Ursache und die Rückfallhäufigkeit als Wirkung zu interpretieren und nicht umgekehrt. Das Experiment selbst hat aber hierzu keine Informationen geliefert.

Deterministische Beziehungen sind allerdings in der Realität nur äußerst selten anzutreffen. (Dies gilt im übrigen auch für die Naturwissenschaften, in denen nachgewiesen wurde, daß viele, ursprünglich kausal vermutete Beziehungen, nur stochastisch ablaufen.) Ein ähnlich gelagertes Beispiel möge dies demonstrieren:

Wir gehen von der Hypothese aus, daß die Behandlung von Sexualdelinquenten mit einem bestimmten Hormon X deren Rückfallhäufigkeit entscheidend reduziere. Die experimentelle Versuchsanordnung sähe wie oben aus: es werden zwei Gruppen gebildet, wo unter sonst gleichen Bedingungen nur die eine mit dem Hormon behandelt wird. Nach einem gewissen Zeitraum (z.B. von 5 Jahren) werde überprüft, in welchen Häufigkeiten in den beiden Gruppen Rückfalldelinquenz aufgetreten ist. Das Ergebnis zeigt die folgende Tabelle:

Tab. 13. Probabilistische Relationen im Experiment

	Hormon X	~ Hormon X	Σ
~ Rückfall	200	80	280
Rückfall	50	220	270
Σ	250	300	550

Die in dem ersten Experiment als deterministisch ausgewiesene Beziehung war eine solche, die in jedem Falle galt. In der Tabelle 12 äußerte sich dies darin, daß zwei Zellen der Matrix unbesetzt geblieben waren. Würde für das zweite Experiment eine deterministische Beziehung unterstellt werden, so dürfte in jedem Falle, in dem ein Delinquenter mit dem Hormon behandelt wurde, keine weitere Delinquenz auftreten und in all den Fällen, wo keine Behandlung erfolgte, müßten weitere Rückfälle zu verzeichnen gewesen sein. Tatsächlich zeigt aber die Tabelle an fiktiven Daten, daß die unterstellte Ursache-Wirkungs-Beziehung zwischen hormoneller Behandlung und Rückfallshäufigkeit nur stochastisch-statistisch abläuft. Inhaltlich interpretiert bedeutet dies, daß zwar eine hormonelle Behandlung durchaus zu einer Rückfallreduktion führt, daß es aber auch andere Gründe dafür gibt, daß jemand nicht rückfällig wird (vgl. die Zelle der Matrix, die jene Personen wiedergibt, die nicht hormonell behandelt worden sind und gleichwohl nicht rückfällig wurden). Die hormonelle Behandlung ist also keine hinreichende Bedingung für die Reduktion der Rückfallshäufigkeit; offensichtlich gibt es noch andere Variablen, die die Rückfallhäufigkeit determinieren. Da diese Variablen aber experimentell nicht kontrolliert worden sind (d.h. nicht in das Experiment durch vorherige Formulierung der Hypothesen einbezogen wurden) kann über deren Wirkung durch das Experiment nichts ausgesagt werden. Das Experiment überprüft also nur jene Bedingungen, die hypothetisch erfaßt und in die experimentelle Versuchsanordnung einbezogen worden sind.
Das Vermögen des Experiments reduziert sich immer darauf, Beziehungen zwischen den in der Hypothese verknüpften Variablen nachzuweisen, wobei die Qualität dieser Beziehungen vom Experiment unberücksichtigt bleibt. (Auch in der scheinbar kausalen (deterministischen) Beziehung zwischen Kastration und Rückfallhäufigkeit wäre durchaus denkbar, daß das Ursache-Wirkungs-Verhältnis ein scheinbares ist und zusätzlich oder allein durch eine weitere, nicht in die Betrachtung und Analyse einbezogene Variable ausgelöst wird. (Vgl. hierzu das Konzept der intervenierenden Variablen in 2.1.1).

Wenn wir abschließend noch einmal die beiden experimentellen Beispiele miteinander vergleichen und danach fragen, worin die Unterschiede bestanden, so ergaben sich im Hinblick auf die Logik des angewandten Verfahrens keine, im Hinblick auf die Qualität der unterstellten und geprüften Beziehungen jedoch wichtige

Unterschiede, wobei selbst in dem scheinbar eindeutigen Fall einer deterministischen Relation dieser Determinismus als nicht gesichert angesehen werden mußte. Somit stellt sich die Frage, ob das Experiment als Methode der Kausalanalyse aufzugeben ist, weil es offensichtlich nicht in der Lage ist, kausale Beziehungen aufzudecken. Diese Frage muß eindeutig verneint werden; erstens deshalb, weil die Unterstellung von deterministischen Abläufen pure Dezision und wissenschaftstheoretisch wie empirisch nicht abgesichert ist und nicht abgesichert werden kann. Der Verzicht auf die Annahme von kausalen Beziehungen bedeutet jedoch keine Einschränkung der Erkenntnismöglichkeiten; im Gegenteil, man würde viele gerade auch pragmatisch nützliche Erkenntnisse als irrelevant ablehnen, weil sie nicht in deterministischen Relationen ablaufen, nur solche aber als wissenschaftlich anerkannt werden.

Stochastische Aussagen erhalten zweitens qualitativ mindestens so viel Erklärungspotential wie deterministische. Zu wissen, daß eine hormonelle Behandlung in z.B. 90 % aller Fälle weitere Straftaten verhindert, ist theoretisch wie praktisch bedeutsam und trägt dazu bei, die Ursachen für Sexualdelinquenz zu ermitteln und zu erklären. Der Verzicht auf die Unterstellung von Kausalität reduziert also den Erkenntnisspielraum in keiner Weise, sondern er ist geradezu eine notwendige Bedingung dafür, daß das Experiment als eine Methode der Überprüfung einer Beziehungsstruktur zwischen Variablen universell eingesetzt werden kann, sofern dem keine pragmatischen, ethischen oder materiellen Restriktionen entgegenstehen.

2.1.1 Absicht und Definition des Experiments

Absicht des Experiments ist es, eine theoretisch vermutete Beziehung zwischen zwei Variablen empirisch auf deren Richtigkeit hin zu testen, zu überprüfen, welche Ursachen welche Wirkungen haben (wobei Ursache und Wirkung jetzt nicht mehr deterministisch, sondern stochastisch verstanden werden). Da im Regelfalle nicht allein die Kovariation zweier Variablen interessiert, sondern eine Richtung in der Beziehung der Variablen unterstellt wird, kann man die in eine Hypothese gefaßte Relation zwischen zwei Variablen formal wie folgt darstellen:

$$H: \; x \longrightarrow y$$

x = unabhängige Variable
y = abhängige Variable

Hierbei deutet der Pfeil die Richtung der Beziehung an. Man unterstellt, daß x y determiniert. Wenn aber x y bestimmt, so ist y von x abhängig. Wenn zugleich gilt, daß y nicht x determiniert, dann ist x unabhängig von y. Daher spricht man von x als der *unabhängigen oder determinierenden Variablen und von y als der abhängigen oder determinierten Variablen.* Unterstellt man solche bivariate (oder bivariable) Beziehungen, so hat man eine Einfachstruktur, die methodologisch und statistisch durchaus problematisch ist (vgl. das Problem von Scheinkorrelation und Intervention). So wäre beispielsweise denkbar, daß im Experiment und statistisch

abgesichert festgestellt wird, daß x y in einem bestimmten Ausmaß (= Korrelationskoeffizient) und mit einer bestimmten Sicherheit (= Signifikanz oder Irrtumswahrscheinlichkeit) determiniert. Die hypothetisch zugrundegelegte Einfachstruktur einer bivariaten Hypothese kann aber theoretisch zu falschen Schlußfolgerungen und Interpretationen führen, weil eine Drittvariable nicht berücksichtigt wurde. Tritt eine solche Drittvariable zwischen die Beziehung von x und y in dem Sinne, daß diese Drittvariable z sich zwischen x und y schiebt, x also nur via z wirkt, so spricht man von z als einer *intervenierenden Variablen* und auf die Struktur solcher Variablenbeziehungen bezogen, von einer *Intervention* ($x \longrightarrow z \longrightarrow y$).

Wenn die Logik des Experiments zur Verdeutlichung nur für bivariate Variablenrelationen erarbeitet wird, so nur deshalb, weil der gedankliche Nachvollzug bei solchen Einfachststrukturen keine Schwierigkeiten bereiten sollte. In der Realität experimenteller Forschung wird man über solche Einfachststrukturen weit hinausgehen müssen, um zu gültigen Beziehungen zwischen den Variablen zu kommen. So werden insbesondere mit der Entwicklung multivariabler statistischer Verfahren in letzter Zeit immer häufiger sog. faktorielle Versuchspläne aufgestellt (vgl. 2.1.4), die experimentell multivariable Beziehungen überprüfen und testen wollen. Für diese faktoriellen Versuchspläne gilt allerdings die Logik des einfachen Experiments analog. Eine Ausweitung der bivariaten in eine multivariate Struktur dürfte gedanklich leicht nachvollziehbar sein, nachdem die Logik des Experiments verstanden wurde (vgl. Beispiel 3 in 2.2).

Die experimentelle Überprüfung einer bivariaten Hypothese setzt zwei Dinge voraus:

1. *Es muß durch die experimentelle Versuchsanordnung ermöglicht werden, daß die abhängige Variable gemessen werden kann,* d.h. es geht darum festzustellen, ob ein Merkmal auftritt oder nicht auftritt (nominal), ob ein Merkmal stark, weniger stark oder überhaupt nicht auftritt (ordinal), oder ob in einer Quantifizierung Differenzen in den Merkmalsausprägungen festgestellt werden können (intervall). Die Feststellbarkeit, die Beobachtbarkeit der abhängigen Variablen muß – auf welchem Meßniveau auch immer – gegeben sein.

2. *Die unabhängige Variable muß durch den Versuchsleiter manipulierbar sein.* Er muß die Möglichkeit haben, die Merkmalsausprägungen zu variieren. Gleich der abhängigen Variablen können bei der unabhängigen Variablen unterschiedliche Meßniveaus auftreten. (In unserem zweiten Beispiel war die Manipulationsmöglichkeit dadurch gegeben, daß in dem einen Falle das Hormon verabreicht wurde, im anderen nicht. Man könnte sich auch vorstellen, daß unterschiedliche Hormondosen verabreicht werden, womit denn mindestens ordinales Meßniveau vorliegt.)

Sind die Meßbarkeit der abhängigen und die Manipulierbarkeit der unabhängigen Variablen gegeben, so erfolgt – wie aus den bisher gebrachten Beispielen ersichtlich – die Überprüfung der aufgestellten Hypothese dadurch, daß zwei sonst gleiche Gruppen konstituiert werden (natürlich kann es sich auch um mehrere Gruppen handeln), bei denen die unabhängige Variable in der vom Untersuchungs-

design vorgeschriebenen Weise variiert und manipuliert wird und nach der Manipulation die jeweils abhängige Variable gemessen wird. Die im einfachsten Fall zu konstruierenden beiden Gruppen werden als Experimental- und Kontrollgruppe bezeichnet. *Die Experimentalgruppe unterscheidet sich von der Kontrollgruppe dadurch, daß die unabhängige Variable in unterschiedlichen Merkmalsausprägungen (Werten) auftritt.* (Im einfachsten psychologischen Experiment könnte man sagen, daß die Experimentalgruppe einem Stimulus X ausgesetzt wird, während in der Kontrollgruppe dieser Stimulus X nicht gegeben wird).

Nach diesen Vorüberlegungen sind wir nun in der Lage, eine umfassendere Definition des Experiments zu geben. Das eingangs erarbeitete Kriterium der Wissenschaftlichkeit setzte Intersubjektivität der Erkenntnisgewinnung voraus, Intersubjektivität kann operational definiert werden als die wiederholte oder wiederholbare Beobachtung unter systematischen, kontrollierten Bedingungen. Die Kontrolle der Beobachtungsbedingungen erstreckt sich auch darauf, daß gleiche Sachverhalte in unterschiedlichen Situationen gleich beobachtet werden. Es wurde gezeigt, daß das Experiment dazu dient, eine zuvor aufgestellte Hypothese, die einen Zusammenhang zwischen Variablen behauptet, zu überprüfen. Letzendlich war ausgeführt worden, daß für eine solche Hypothesenprüfung die Notwendigkeit besteht, daß die abhängige Variable gemessen und die unabhängige Variable gemäß dem in der Hypothese behaupteten Zusammenhang variiert und manipuliert werden kann. Demnach kann das Experiment nun definiert werden als „wiederholbare Beobachtung unter kontrollierten Bedingungen, wobei (eine oder mehrere) unabhängige Variable(n) derartig manipuliert wird (werden), daß eine Überprüfungsmöglichkeit der zugrundeliegenden Hypothese ... in unterschiedlichen Situationen gegeben ist" (ZIMMERMANN, E., Das Experiment in den Sozialwissenschaften, Stuttgart 1972, S. 37).

2.1.2 Kriterien des Experiments

In der oben entwickelten Definition sind die wesentlichsten Kriterien des Experiments enthalten:

a) willkürliche Herstellbarkeit der experimentellen Situation durch die Manipulierbarkeit der unabhängigen Variablen,
b) die Wiederholbarkeit der Durchführung von Experimenten,
c) die Variierbarkeit der experimentellen und situativen Bedingungen sowie
d) die Kontrolle der experimentellen Bedingungen.

Diese vier Kriterien sollen im einzelnen noch erläutert und begründet werden.

Willkürliche Herstellbarkeit ist hier selbstverständlich nicht im Sinne arbiträrer Konstruktion eines Experiments, sondern als intentionaler, gezielter, von der vorab aufgestellten Hypothese geleiteter Entwurf eines Experiments zu verstehen. Willkürliche Herstellbarkeit meint also, solche Bedingungen schaffen zu können, die dem Zweck des Experiments dienen, nämlich den Einfluß der variierten Bedingungen auf einen hypothetisch vermuteten Sachverhalt festzustellen. Die willkür-

liche Herstellbarkeit experimenteller Bedingungen ist in deren Realisation natürlich von dem jeweiligen Untersuchungsgegenstand abhängig. So kann der Fall eintreten, daß in Hypothesen Sachverhalte erfaßt werden, die sich einer experimentellen Überprüfung aus pragmatischen, ethischen und/oder anderen Gründen entziehen. In diesen Fällen muß das Experiment durch andere Methoden der Hypothesenprüfung ersetzt werden.

Die Manipulierbarkeit bzw. Variierbarkeit der unabhängigen Variablen war gefordert worden, um deren Einfluß auf die abhängige Variable feststellen zu können. Manipulierbarkeit unterscheidet sich von Variierbarkeit insoweit, als Variierbarkeit die Manipulierbarkeit miteinschließt, Manipulierbarkeit ein Spezialfall der Variierbarkeit ist. An einem Beispiel soll dies demonstriert werden.

Um den Einfluß des Geschlechtes auf eine bestimmte abhängige Variable ermitteln zu können, muß das Geschlecht variiert werden, d.h., es werden Untersuchungs- und Kontrollgruppen geschlechtspezifisch konstituiert. Diese Variation ist aber keine Manipulation, da die unabhängige Variable des Geschlechts bei den in die Untersuchungsgruppen gekommenen Personen nicht bewußt manipuliert und damit verändert wird. Soll jedoch der Einfluß von Lärm auf Schlafstörungen untersucht werden, so kann in der experimentellen Versuchsanordnung die unabhängige Variable des Lärms beliebig manipuliert werden.

Manipulation heißt also: bewußte Veränderung der Merkmalsausprägungen der unabhängigen Variablen, während Variation von unterschiedlichen Merkmalsausprägungen ausgeht und diese den Kontroll- bzw. Experimentalgruppen nur zuordnet. Bei den unabhängigen Variablen, die nur variiert und nicht manipuliert werden können, spricht man auch von *präexperimentellen Variablen* (vgl. Ingeborg STELZL, Experimentelle Versuchsanordnungen, S. 139 in: KOOLWIJK van, J. und WECKEN-MAYSER, M. (Hrsg.), Techniken der empirischen Sozialforschung, Bd. 6, Statistische Forschungsstrategien, München 1974).

Als drittes Kriterium für die Durchführung von Experimenten war die *Wiederholbarkeit* genannt worden, wobei dieses nicht eine ausschließlich experimentelle Bedingung ist, sondern vielmehr ein allgemeines Kriterium wissenschaftlichen Vorgehens für alle empirischen Methoden darstellt. Die Wiederholbarkeit empirischer Untersuchungen – insbesondere experimenteller Versuchsanordnungen – ist eine Forderung, die der Intersubjektivität und der Nachprüfbarkeit gewonnener Erkenntnisse dienen soll. Daß die Wiederholbarkeit von experimentellen Versuchen eine Notwendigkeit ist, hat sich mehrfach, insbesondere bei sozialpsychologischen Experimenten gezeigt. Denn dort ergab sich bei wiederholter Durchführung von Experimenten zu gleichen Sachverhalten keine Übereinstimmung in den experimentellen Resultaten. Mit als Ursache für solche divergierenden Ergebnisse kann angeführt werden, daß noch nicht erkannte Störvariablen einen jeweils unterschiedlichen Einfluß auf die abhängige Variable ausgeübt haben. Schon aus diesem Grunde kann die Empfehlung, Versuchsanordnungen zu replizieren, von erheblicher erkenntnistheoretischer Relevanz sein.

Die Wiederholbarkeit als Kriterium des Experiments kann unter die Forderung nach *Kontrolle* des Experiments subsumiert werden. Kontrolle meint hier die am theoretischen Konzept gemessene, systematische Zuführung der Versuchsbedin-

gungen zu einer intersubjektiven und replizierbaren Überprüfung. Kontrolle als Element eines Experiments ist ebenfalls ein Kriterium allgemein wissenschaftlicher Methoden und nicht ein spezifisch experimentelles, wenngleich beim Experiment spezielle Kontrolltechniken entwickelt wurden (vgl. 2.1.3).

2.1.3 Kontrolltechniken

Kontrolltechniken dienen dazu, eventuell auftretende Störgrößen, die die experimentelle Hypothesenüberprüfung beeinflussen können, auszuschalten. Geht es bei der Hypothesenprüfung durch das Experiment um die Feststellung, welche Ursachen welche Wirkungen erzeugen, so müssen in einer schlüssigen Beantwortung dieser Frage *alle* hypothetisch relevanten Faktoren, sowohl auf der Seite der unabhängigen, wie auf der Seite der abhängigen Variablen kontrolliert werden. Ohne deren gegenseitige Kontrolle kann eine Zurechnung von Ursache und Wirkung und insbesondere eine solche des Ausmaßes der Wirkung nicht vorgenommen werden.

Während in einem naturwissenschaftlichen Experiment es in der Regel völlig hinreichend ist, einen Objektbereich zu messen, im Anschluß daran, einen bestimmten Stimulus zu setzen, und in der Folge diesen Objektbereich wieder zu messen, genügt diese einfachste aller Versuchsanordnungen für sozialwissenschaftliche Experimente normalerweise nicht. In den Naturwissenschaften kann nämlich vorausgesetzt werden, daß alle nichtmanipulierten Faktoren bei der Vorher- und Nachhermessung konstant geblieben sind, weil der Meßvorgang das Meßobjekt selbst nicht verändert. In den Sozialwissenschaften hingegen, wo der *Objektbereich aus Subjekten besteht,* die auf den Meßvorgang selbst reagieren und mithin jede Variable verändern können, müssen kompliziertere Kontrolltechniken angewandt werden, um solche und ähnliche Probleme zu verhindern. Will man daher in den Sozialwissenschaften den Einfluß einer unabhängigen Variablen x auf eine abhängige Variable y gültig messen, muß die einfache naturwissenschaftliche Versuchsanordnung, bei der die Vorher- und Nachhermessung an einem Objekt (oder mehreren Objekten) vorgenommen wurde, erweitert werden. Diese Erweiterung führt zu der sog. *klassischen Versuchsanordnung des Experiments,* bei der zwei Gruppen gebildet werden: Die *Experimentalgruppe und die Kontrollgruppe.* In beiden Gruppen wird eine Vorher- und Nachhermessung der abhängigen Variablen vorgenommen, wobei in der Experimentalgruppe zwischen erster und zweiter Messung ein Stimulus gesetzt wird, der bei der Kontrollgruppe entfällt. Damit gelingt es besser, bestimmte, gemessene Wirkungen in der abhängigen Variablen der unabhängigen Variablen (dem Stimulus) zuzuordnen. (Mit dieser Versuchsanordnung sind jedoch längst nicht alle Probleme gemeistert, wie der Abschnitt 2.1.4 noch zeigen wird.)

Welches sind nun die Einflüsse, die im Experiment kontrolliert werden müssen, um zu gültigen Aussagen kommen zu können. KISH (Some Statistical Problems in Research Design, in: American Sociological Review Nr. 24 1959, S. 328 – 338) hat eine Typologie solcher Einflußfaktoren entwickelt. Er trennt analytisch 4 Variablengruppen, die im Experiment kontrolliert werden müssen:

1. Die unabhängige Variable, die variiert bzw. manipuliert und deren Einfluß auf die abhängige Variable gemessen werden soll, muß kontrolliert werden können (durch die Tatsache, daß diese Variable manipuliert wird, ist im Regelfall deren Kontrollierbarkeit auch gegeben).

2. Neben diesen manipulierten Variablen gibt es möglicherweise weitere unabhängige Variablen, die einen Einfluß auf die abhängige Variable ausüben, die aber durch das Experiment kontrolliert – z.B. durch Konstanthalten, Abschirmung etc. – werden können. Je nach theoretischer Durchdringung des Objektbereiches wäre zudem vorstellbar, daß weitere, determinierende Variablen existieren, die dem Forscher aber bislang verborgen geblieben sind. Auch deren Einfluß sollte im Experiment kontrolliert werden können.

3. Hier handelt es sich um solche Faktoren, die die abhängige Variable determinieren, die aber in keinerlei Beziehung zu der im Experiment selbst manipulierten, unabhängigen Variablen stehen.

4. Weiter ist denkbar, daß es Variablen gibt, die einerseits die abhängige Variable determinieren und andererseits zugleich mit der unabhängigen Variablen, die im Experiment variiert wird, korrelieren. So ist der Fall in den Sozialwissenschaften gar nicht so selten, daß eine unabhängige Variable x eine abhängige y determiniert (scheinbar), tatsächlich aber eine weitere unabhängige Variable z für die scheinbaren Beziehungen verantwortlich ist, z.B. in dem Sinne, daß z sowohl x als auch y bestimmt (= Scheinkorrelation: x ⟵ z ⟶ y).

Zweck der Kontrolltechniken ist es nun, nur Variablen der ersten und zweiten Gruppe zuzulassen. Intention des Versuchsleiters muß sein, die Variablen aus der vierten Gruppe auszuschalten, abzuschirmen oder zu eliminieren und die Faktoren der dritten Gruppe in solche der zweiten Gruppe überzuführen (vgl. ZIMMERMANN, E., Das Experiment in den Sozialwissenschaften, Stuttgart 1972, S. 62).

Der Realisierung dieses Bestrebens des Versuchsleiters dienen im wesentlichen vier Kontrolltechniken, die im folgenden abgehandelt werden sollen. Die ersten beiden kurz zu nennenden Kontrolltechniken, nämlich *Ausschaltung und Abschirmung,* sind sozusagen extra-inhaltliche Kontrolltechniken, die im Grunde genommen für jede experimentelle Versuchsanordnung gelten, gleichgültig, welche Variablen in die Untersuchung einbezogen werden. Die beiden weiteren Kontrolltechniken *Matching und Randomisierung* hingegen stehen in einem direkten inhaltlichen Bezug zu den zu testenden Variablen und versuchen, inhaltliche Störgrößen auszuschalten.

Unter Ausschaltung versteht man den Versuch, erkannte, experimentelle Störgrößen zu neutralisieren. Ist die Wirkung einer solchen Störgröße bekannt, so kann man Gegenstrategien entwickeln, die die potentielle Wirkung zu kompensieren in der Lage sind. Ist die Wirkung im Konkreten nicht bekannt, weiß man jedoch, daß Einflüsse von der Störvariablen ausgehen, so versucht man die Störvariable unmittelbar auszuschalten. Da es in den Sozialwissenschaften nur in den seltensten Fällen möglich sein wird, solche Störgrößen zu eliminieren, wird es in der Regel darauf ankommen, solche Faktoren durch Gegenstrategien zu neutralisieren.

Sind sowohl Eliminierung wie Ausschaltung von Störgrößen nicht praktikable und realisierbare Kontrolltechniken, so kann versucht werden, die Störvariable abzuschirmen. *Bei der Abschirmung wird angestrebt, die potentielle Störquelle durch ein anderes Merkmal so zu überlagern, daß ihre Wirksamkeit reduziert oder völlig ausgeschlossen wird.* Dabei muß natürlich davon ausgegangen werden, daß die abschirmenden Variablen selbst keinen Einfluß auf die unabhängige Variable ausüben.

Die Randomisierung ist eine Kontrolltechnik, die auf der Wahrscheinlichkeitstheorie basiert und aus allgemeinen, stichprobentheoretischen Überlegungen für die experimentelle Versuchsanordnung abgeleitet werden kann. Geht man im einfachsten Falle davon aus, daß eine Kontrollgruppe und eine Experimentalgruppe konstituiert werden müssen, wobei sich das inhaltliche Kriterium für die Auswahl der Gruppen aus den zu testenden Hypothesen ergibt, so stehen für Experimental- und Kontrollgruppe die gleichen Populationen zur Verfügung (z.B. Studenten der Universität München). Zieht man nun aus der Grundgesamtheit eine Stichprobe zufällig, d.h. daß jeder Student die gleiche (mindestens aber eine berechenbare und von Null verschiedene) Chance hat, in die Stichprobe zu gelangen, so kann man bei genügend großer Stichprobe davon ausgehen, daß diese Teilpopulation in allen Merkmalen der Gesamtpopulation in etwa entspricht. Zieht man aus dieser Teilpopulation nochmals eine Zufallsstichprobe derart, daß zwei Gruppen daraus gebildet werden, so entsprechen bei ausreichender Gruppengröße beide in etwa der ursprünglichen Gesamtpopulation. *Somit wären Experimental- und Kontrollgruppe derart konstituiert, daß sie repräsentativ für die Gesamtpopulation sein können.*

Dieses, relativ komplizierte, zweistufige Stichprobenverfahren kann mit demselben Effekt durch ein einfacheres abgelöst werden, indem aus der ursprünglichen Grundgesamtheit gleich zwei zufällige Stichproben gezogen werden. Handelt es sich wieder um Zufallsstichproben gemäß der Wahrscheinlichkeitstheorie, so gelten diese beiden Stichprobengruppen sowohl als repräsentativ für die ursprüngliche Grundgesamtheit, wie auch als untereinander gleich zusammengesetzt.

Mit diesem Randomisierungsverfahren gelingt es, alle nur denkbaren Einflußfaktoren im Experiment zu kontrollieren, denn nach dem Gesetz der Wahrscheinlichkeitstheorie sind (unter bestimmten einschränkenden Bedingungen, z.B. ausreichend große Zahl der Stichprobe etc.) alle Variablen der Population in der Stichprobe repräsentativ vertreten, somit auch in Experimental- und Kontrollgruppe, sodaß potentielle Einflußfaktoren sowohl in der Experimental- wie auch in der Kontrollgruppe (und zwar in gleicher Weise) wirken, sodaß Unterschiede in der abhängigen Variablen zwischen Untersuchungs- und Kontrollgruppe auf die manipulierte, unabhängige Variable zurückgeführt werden können.

Tab. 14. Wirkung der Randomisierung

a) Verteilung der Merkmale in der Grundgesamtheit

	+	–	Σ
+	83% (1000)	63% (500)	1500
–	17% (200)	37% (300)	500
Σ	1200	800	2000

b) Verteilung der Merkmale in der Experimentalgruppe

	+	–	Σ
+	81% (47)	62% (26)	73
–	19% (11)	38% (16)	27
Σ	58	42	100

c) Verteilung der Merkmale in der Kontrollgruppe

	+	–	Σ
+	81% (48)	68% (28)	76
–	19% (11)	32% (13)	24
Σ	59	41	100

Tab. 14 verdeutlicht in Form von Kreuztabellen die Wirkung der Randomisierung an zwei exemplarisch herausgegriffenen Variablen.

Dieses fiktive Beispiel für die Randomisierung macht deutlich, daß im Hinblick auf die in den Tabellen erfaßten Variablen, sich Kontroll- und Experimentalgruppe nur innerhalb statistischer Schwankungsbreiten in der Verteilung der Variablen unterscheiden. Sowohl die Zellenbesetzungen wie die Randverteilungen sind nicht identisch, weil durch das Prinzip der Zufallsstreuung statistische Schwankungen möglich sind. Gleichwohl bleibt gewährleistet, daß alle Faktoren innerhalb der Zufallsstreuung im Experiment durch die Randomisierung kontrolliert sind. Damit gelingt es mit dem relativ einfachen Verfahren der Stichprobenziehung auch, die Einwände gegen das sozialwissenschaftliche Experiment ad absurdum zu führen, die davon ausgehen, daß nicht alle Variablen, die als potentielle Stör- und Einflußgrößen gelten können, kontrolliert werden können, weil deren Abschirmung, Ausschaltung, Konstanthaltung etc. aus den verschiedensten Gründen heraus nicht möglich wäre.

Bei der Kontrolltechnik des *Matching bzw. Parallelisierens wird angestrebt, den Einfluß der unabhängigen Variable im Experiment dahingehend zu kontrollieren, daß sie in der Kontroll- und in der Experimentalgruppe in gleicher Verteilung auftreten.* Die Gleichheit der Versuchseinheiten kann sich dabei einmal darauf beziehen, daß Kontroll- und Experimentalgruppe so konstituiert werden, daß sich über die Summe aller Personen hinweg, eine gleiche Merkmalsverteilung der unabhängigen Variablen ergibt, was man als *matched groups oder parallelisierte Gruppen* bezeichnet. Zum zweiten können innerhalb der Gruppen die einzelnen Individuen sich jeweils als Paar gegenüberstehen, wobei die Paare im Hinblick auf die zu kontrollierenden unabhängigen Variablen identisch sind, was als *matched pairs oder parallelisierte Paare* gilt. Daraus ergibt sich, daß die parallelisierten Paare ein Spezialfall der parallelisierten Gruppen sind, wie die nachfolgende tabellarische Darstellung ausweist.

Tab. 15. Wirkung des Parallelisierens

a) Parallelisierte Paare

Experimentalgruppe

	+	–	Σ
+	23	32	55
–	17	28	45
Σ	40	60	100

Kontrollgruppe

	+	–	Σ
+	23	32	55
–	17	28	45
Σ	40	60	100

b) Parallelisierte Gruppen

Experimentalgruppe

	+	−	Σ
+	24	31	55
−	16	29	45
Σ	40	60	100

Kontrollgruppe

	+	−	Σ
+	25	30	55
−	15	30	45
Σ	40	60	100

Aus der obigen Darstellung ergibt sich im Vergleich mit der zur Randomisierung, daß in letzterer sich sowohl die Individuen (Zellenbesetzung), wie auch die Gruppen (Randverteilung) im Hinblick auf die unabhängigen Variablen unterscheiden können. Bei den parallelisierten Gruppen ist die Randverteilung in Experimental- und Kontrollgruppe gleich; es können sich aber durchaus die Individuen in den Häufigkeiten unterscheiden, während bei den parallelisierten Paaren gleiche Randverteilung und Zellenbesetzung auftritt. In anderer Formulierung: Bei den parallelisierten Gruppen handelt es sich um eine Gleichverteilung der interessierenden Kontrollvariablen für die Gruppen insgesamt, während bei den parallelisierten Paaren die Individuen der Experimental- und Kontrollgruppe jeweils in gleicher Häufigkeit gleiche Werte bei den unabhängigen Variablen annehmen. Bei der Randomisierung handelt es sich *tendenziell* um paarweises Matching, wobei zwischen Kontrollgruppe und Experimentalgruppe im Hinblick auf die Zellenbesetzungen und die Randverteilungen innerhalb der statistischen Schwankungsbreiten eine Gleichverteilung zu erwarten ist.

Es ist einleuchtend, daß es für den Versuchsleiter einfacher ist, parallelisierte Gruppen herzustellen als parallelisierte Paare. Die größere Praktikabilität muß allerdings erkauft werden mit einem Verlust an Erkenntnis- und Interpretationsmöglichkeiten, denn im Falle der parallelisierten Gruppen können nur Aussagen für die Gruppe insgesamt gemacht werden. Rückschlüsse auf die Individuen könnten fehlerbehaftet sein. Zudem muß gelten, daß bei den parallelisierten Gruppen Häufigkeiten in den einzelnen Zellen nicht notwendigerweise zwischen Experimental- und Kontrollgruppe gleich bzw. ähnlich sein müssen, weil durch die bewußte Auswahl der Randhäufigkeiten nicht automatisch auch die Kombinationen der Merkmalsausprägung der beiden Variablen sich entsprechen müssen. Im Extremfall (insbes. in kleinen Gruppen) wäre z.B. folgende Situation denkbar:

Tab. 16. Parallelisierte Gruppen mit signifikant verschiedenen Zellenbesetzungen

Experimentalgruppe

	+	−	Σ
+	20	30	50
−	20	30	50
Σ	40	60	100

Kontrollgruppe

	+	−	Σ
+	5	45	50
−	35	15	50
Σ	40	60	100

Probleme, wie sie sich aus der obigen Darstellung ergeben, können durch das paarweise Parallelisieren bzw. durch Randomisieren vermieden werden.

Für das Parallelisieren allgemein gilt, daß zwei gravierende Probleme mit dessen Anwendung verbunden sind, die durch Randomisieren nicht auftreten können: Wenn man nämlich durch Parallelisieren den Einfluß der unabhängigen Variablen kontrollieren will, so muß man immer schon wissen, welches die unabhängigen Variablen sind, die potentialiter die abhängige Variable determinieren. Wenn man aber theoretisch-hypothetisch nicht weiß, welches die potentiellen, unabhängigen Variablen sind, so fällt die Methode des Matching als Kontrolltechnik aus. Glaubt man ein oder zwei unabhängige Variablen zu kennen, so kann zwar nach diesen Variablen parallelisiert werden, doch gilt wiederum – wegen der nicht zufälligen Auswahl –, daß damit nicht notwendigerweise alle anderen denkbaren, unabhängigen Variablen kontrolliert werden. Es sind daher Störeinflüsse zu erwarten, die experimentell nicht erfaßt sind.

Unter der Voraussetzung, man würde alle Einflußgrößen kennen und parallelisieren wollen, gilt eine zweite Restriktion des Parallelisierens, nämlich, daß mit steigender Zahl der Variablen bzw. mit zunehmender Zahl der Merkmalsausprägungen einer konstanten Zahl von Variablen die Gruppengröße und damit die Zahl der Versuchspersonen so sehr steigen würde (und um statistisch gesicherte Aussagen machen zu können, steigen müßte), daß sich aus Praktikabilitätsgründen die Durchführung eines solchen Experimentes verbietet.

Tab. 17. Konsequenzen einer Erweiterung der Merkmalsausprägungen einer Variablen beim Parallelisieren: Notwendigkeit höherer Fallzahlen.

Variable y \ Variable x	1	2	3	4	5	Σ
+	25	30	25	30	25	135
−	25	30	15	20	25	115
Σ	50	60	40	50	50	250

n = 250

Gilt dieses Argument tendenziell auch für das Randomisierungsverfahren, so hebt sich letzteres vom ersteren doch dadurch ab, daß bei der Randomisierung die unabhängigen Variablen für die Konstituierung der experimentalen Kontrollgruppe nicht bekannt sein müssen. Dieser Vorteil wird jedoch kompensiert durch den erheblichen Nachteil, daß bei sehr ungleich auftretender Verteilung von Merkmalswerten (z.B. Geschlechterproportion 1:20) sich in den Experimental- und Kontrollgruppen dieselbe Proportion ergibt, womit die Zellenbesetzungen so klein werden, daß statistische Absicherungen kaum mehr vorgenommen werden können. In solchen Fällen würde sich in der Tat anbieten zu matchen (also eine Geschlechterproportion von 1:1 in Experimental- und Kontrollgruppe zu haben). Auch bietet sich eine Parallelisierung solcher Faktoren an, die in Realität bereits in entsprechender, z.B. dichotomisierter, Form vorliegen, wie dies bei der Geschlechtszugehörigkeit der Fall ist.

Da Parallelisierung und Randomisierung qualitativ unterschiedliche Vor- bzw. Nachteile aufweisen, die sich z.T. auch kompensieren können, bietet sich als optimale Strategie die kombinierte Kontrolltechnik des Randomisierens und Parallelisierens an. Grundsätzlich wird immer eine Entscheidung für oder gegen eine bestimmte Kontrolltechnik nur in Abhängigkeit von dem jeweiligen Untersuchungsgegenstand und den ihm zugrundeliegenden Variablen getroffen werden. Die hier entwickelten, generellen Argumente können immer nur Leitlinie und Richtschnur für solche Entscheidungen sein.

2.1.4 Experimentelle Konfigurationen

In diesem Abschnitt geht es nicht darum, alle nur möglichen, denkbaren experimentellen Designs vorzustellen; für den Praktiker, der selbst ein Experiment durchführen möchte, sei diesbezüglich auf die ausgezeichnete Darstellung bei CAMPBELL und STANLEY hingewiesen (CAMPBELL, D.T. und STANLEY, J.C.: Experimental and Quasi-Experimental Designs for Research, Chicago 1963). Vielmehr soll hier versucht werden, aus der Darstellung *pseudo-experimenteller Versuchsanordnungen* die mit ihnen verbundenen Probleme aufzuzeigen und daraus die wichtigste, sog. *klassische Versuchsanordnung* des Experiments abzuleiten, die diese Probleme vermeiden hilft. Die unvollkommenste, pseudo-experimentelle Versuchsanordnung zeichnet sich dadurch aus, daß ein Stimulus gesetzt wird und im Anschluß daran eine Messung vorgenommen wird (Stimulus = S, Messung = M).

1) S M

Da jede Vergleichs- und Kontrollmöglichkeit der einmaligen Messung mit vorhergehenden oder parallelen anderen Messungen fehlt, ist die Aussagekräftigkeit einer solchen Untersuchung eines Einzelfalles praktisch gleich Null.

Ein erweitertes, pseudo-experimentelles Design stellt jene Versuchsanordnung dar, bei der an einer Gruppe eine Messung vorgenommen, dann ein Stimulus gegeben und im Anschluß daran eine zweite Messung durchgeführt wird.

2) M_1 S M_2

Diese Methode, die für die Naturwissenschaften durchaus brauchbare Erkenntisse liefern kann, ist für sozialwissenschaftliche Zwecke aus den folgenden Gründen mehr oder weniger ungeeignet:

1. Durch die zweimalige Messung derselben Versuchspersonen in einem gewissen zeitlichen Abstand ist nicht auszuschließen, daß zeitliche Einflüsse, die außerhalb der experimentellen Versuchsanordnung liegen, die abhängige Variable, die gemessen wird, verändert haben. Dabei spielt es eine relativ untergeordnete Rolle, wie groß der Zeitabstand zwischen den Messungen war. Prinzipiell gilt, daß der zeitliche Einfluß durch diese Versuchsanordnung nicht festgestellt werden kann.

2. Da wir es in den Sozialwissenschaften bei den Meßobjekten mit Subjekten zu tun haben, kann bei dieser Versuchsanordnung jedenfalls nicht ausgeschlossen werden, daß unbeabsichtigte Meßeffekte wirksam werden. So können Lern-, Sensibilisierungs-, Situationseffekte usw. die zweite Messung in erheblichem Maße beeinflussen, ohne daß eine solche, die Meßergebnisse verzerrende, Abweichung festgestellt werden könnte.

3. Da wir in den Sozialwissenschaften normalerweise nicht mit exakten mechanischen Meßinstrumenten arbeiten können, sondern die Meßinstrumente häufig die zu beobachtenden Subjekte sind, ist auch nicht auszuschließen, daß bei wiederholter Anwendung des Meßinstruments im Instrument selbst Lern- oder Ermüdungseffekte auftreten, die das Meßinstrument weniger valide machen. Auch solche Veränderungen werden durch die obige Versuchsanorddung nicht kontrolliert.

4. Insbesondere durch sozialpsychologische Experimente konnte bei wiederholter Messung an derselben Population der sog. *Regressionseffekt* festgestellt werden. Dies bedeutet, daß bei mehrmaliger Messung sich der Variationsspielraum der Meßwerte in Richtung auf den gemessenen Mittelwert einpendelt, daß er auf diesen hin regrediert. (Technisch kann man auch davon sprechen, daß sich die gemessene Spannweite der Merkmalsvariablen und die Varianz bzw. Standardabweichung der Meßwerte reduziert.) Als Erklärung für solche Regressionseffekte kann man Auswahlverzerrungen, zufällige Extremwerte bei der ersten Messung etc. anführen.

Da die vier genannten Fehlerquellen als Störgrößen und quantitativ nicht erfaßbare Verzerrungen in pseudo-experimentellen Versuchsanordnungen wirken, müssen bei echten experimentellen Versuchsanordnungen solche Fehlerquellen ausgeschaltet werden. Kriterium für die Zuordnung von „experimentellen" Konfigurationen zu pseudoexperimentellen oder echten experimentellen Versuchsanordnungen ist nun die Kontrolle der oben genannten vier Fehlerquellen. Die klassische Versuchsanordnung der Vorher-Nachhermessung mit einer Kontroll- und einer Experimentalgruppe (für die Diskussion dieser Konfiguration kann unentschieden belassen werden, ob die Kontrollgruppe durch Parallelisierung oder Randomisierung entstanden ist), wobei die Experimentalgruppe einem Stimulus ausgesetzt wird, der bei der Kontrollgruppe fehlt, genügt offensichtlich den Bedingungen, die für eine optimale Variablenkontrolle genannt werden müssen.

3)	M_1	S	M_3	Experimentalgruppe	Randomisierung oder
	M_2		M_4	Kontrollgruppe	Parallelisierung

Experimental- und Kontrollgruppe werden vor dem eigentlichen Experiment (Einführung der experimentellen Variablen als Stimulus oder Variation der experimentellen Variablen) im Hinblick auf die abhängigen Variablen gemessen. Nach Durchführung des Experiments erfolgt in beiden Gruppen die erneute Messung der ab-

hängigen Variablen. Mit dieser experimentellen Konfiguration gelingt es, jeweils unter ceteris-paribus-Bedingungen, die oben angesprochenen Fehler zu vermeiden, bzw. zu kontrollieren:

1. Eventuell auftretende Zeiteinflüsse, die für Experimental- und Kontrollgruppe in gleicher Weise gewirkt haben müssen, werden durch den Vergleich der beiden Messungen in der Kontrollgruppe festgestellt und bei der aufgetretenen Meßwertdifferenz in der Experimentalgruppe mitberücksichtigt. Ohne diesen potentiellen Veränderungseffekt in der Kontrollgruppe hätte man diesen der unabhängigen Variablen zugerechnet.

2. Meßeffekte durch wiederholte Messung treten ebenfalls in der Experimental- und Kontrollgruppe in gleicher Weise auf. Die Meßeffektgrößen-Berücksichtigung der Kontrollgruppe innerhalb der Experimentalgruppe führt zu einer richtigen Interpretation des Effektes, der durch die unabhängige Variable verursacht wurde.

3. Auch Meßinstrumenteffekte und -Veränderungen sollten sich innerhalb statistischer, zufälliger Schwankungen in den beiden Gruppen in gleicher Weise auswirken, sodaß eine Differenzierung in der Zuordnung der Meßeffekte bzw. der Wirkungseffekte durch die unabhängige Variable möglich ist.

4. Die Regressionseffekte werden ebenso durch diese Versuchsanordnung kontrolliert, weil sie in ähnlicher Weise in der Experimental- und Kontrollgruppe auftreten sollten.

Die jeweils analytisch herausgegriffenen Fehlermöglichkeiten, die durch diese Versuchsanordnung kontrolliert werden, können natürlich nicht in ihrer isolierten, quantititativen Ausprägung erfaßt werden. Vielmehr ist es so, daß die gemeinsame Summe dieser potentiellen Störfaktoren durch die zweimalige Messung in der Kontrollgruppe ermittelt und kontrolliert wird. Eine quantitative Zuordnung von Meßwertdifferenzen auf die einzelnen Fehlerquellen ist nicht möglich. Aus der Differenz zwischen den Effekten der Experimentalgruppe und den Effekten der Kontrollgruppe ergibt sich jedoch der quantitative Einfluß der unabhängigen, experimentellen Variablen auf die abhängige Variable.

Diese klassische Versuchsanordnung kann unter der Voraussetzung, daß Experimental- und Kontrollgruppe durch Randomisierung gewonnen wurden, vereinfacht werden, wobei zusätzlich erreicht wird, daß insbesondere bei reaktiven Meßverfahren, Meßeffekte durch wiederholte Messung ausgeschlossen werden können. Sie hat zudem den Vorteil, daß sie ökonomischer ist, weil zwei Messungen entfallen.

4)	S	M_1		Experimentalgruppe	Randomisierung
		M_2		Kontrollgruppe	

Aus der Kombination dieser beiden echten experimentellen Versuchsanordnungen ergibt sich die sog. *Solomon-Vier-Gruppen-Anordnung*, die die Vorteile der beiden Konfigurationen in sich vereint:

5)	M_1	S	M_3
	M_2		M_4
		S	M_5
			M_6

Aus den Meßwerten, die man jeweils anderen Meßwerten als Differenzen gegenüberstellen kann, lassen sich einzelne Effekte isolieren. (Der Leser überlege, welche Effekte bei welchen Meßwertdifferenzen ermittelt werden können.) Vernachlässigt man, daß diese Versuchsanordnung wegen ihrer Komplexität nur schwer zu realisieren ist (Kosten- und Praktikabilitätsgründe), so könnten mit ihr doch zusätzlich *Interaktionseffekte* festgestellt werden. Allerdings muß gleichzeitig einschränkend konzediert werden, daß es kein statistisches Verfahren gibt, das in der Lage wäre, diese experimentelle Konfiguration als ganzheitliche zu testen, sodaß deren erkenntniserweiternder Wert wiederum reduziert wird.

Nach der bisher vorgenommenen Definition des Begriffes der Kontrollgruppe müssen wir bei der Solomon-Vier-Gruppen-Anordnung feststellen, daß es sich dabei um zwei Experimental- und zwei Kontrollgruppen handelt (Kontrollgruppe war ja definiert worden als eine mit der Experimentalgruppe vergleichbare Gruppe, bei der die unabhängige Variable nicht vorhanden war oder nicht manipuliert wurde.) Streng genommen muß man jedoch davon ausgehen, daß die Vorgabe eines Stimulus bzw. dessen Nicht-Vorgabe logisch gleichwertig sind, eine Differenzierung in Experimental- und Kontrollgruppe also nicht gerechtfertigt erscheint. So wäre denkbar, daß man ein zu testendes Medikament einer Gruppe 1 in der Dosis x_1, der Gruppe 2 in der Dosis x_2, der Gruppe 3 in der Dosis x_3 verabreicht. Jede Gruppe ist somit Experimentalgruppe, im Vergleich der Gruppen untereinander jedoch zugleich auch immer Kontrollgruppe. Es ist also nicht in jedem Falle streng zwischen Experimental- und Kontrollgruppe zu trennen.

Auf diesen Sachverhalt weisen die *faktoriellen Versuchsanordnungen* hin, wo die experimentellen Gruppen sich wechselseitig und gegenseitig kontrollieren. Für den einfachen Fall, wo nur zwei Variablen in der experimentellen Versuchsanordnung variiert werden, wobei die eine Variable dichotomisiert und die andere trichotomisiert ist, ergibt sich dann die folgende Versuchsanordnung (der einfachste Fall wäre der, wo beide Variablen nur dichotomisiert sind):

6)

G_1: x_1y_1	G_4: x_2y_2
G_2: x_1y_2	G_5: x_2y_2
G_3: x_1y_3	G_6: x_2y_3

G = Gruppe
x = Variable 1
y = Variable 2

Die Indizes geben die jeweiligen Merkmalsausprägungen an bzw. die Gruppennummer.

Dieser faktoriellen Darstellung kann entnommen werden, daß mindestens 6 Untersuchungsgruppen benötigt werden, um die zwei unabhängigen Variablen in ihrem Einfluß auf die abhängige testen zu können. Allgemein gilt, daß sich die Zahl der benötigten Versuchsgruppen aus der Formel n x k errechnet, wobei n die Zahl der

Variablen und k die Zahl der Werte (= Merkmalsausprägungen) angibt. Aus der multiplikativen Zunahme der erforderlichen Gruppenzahl bei faktoriellen Versuchsanordnungen – in Abhängigkeit von der Zahl der Variablen und deren Werten – ergibt sich eine relativ restriktive Anwendungsmöglichkeit faktorieller Versuchspläne, weil man im Regelfall keine ausreichend große Zahl von Versuchspersonen zur Verfügung haben wird, bzw. weil sich aus ökonomischen Gründen die Durchführung des Experimentes verbietet.

(Die relativ komplizierten faktoriellen Versuchspläne können je nach theoretisch-statistischem Raffinement erweitert und in Spezialfälle transformiert werden; so z.B. in das lateinische Quadrat, griechisch-lateinische Quadrat etc., die aber alle mehr theoretischen als praktischen Wert haben. Für den interessierten Leser sei auf ZIMMERMANN, D , Das Experiment in den Sozialwissenschaften, Stuttgart 1972, verwiesen.)

2.1.5 Fehlerquellen im Experiment

Wie bei jeder empirischen Methode treten auch beim Experiment Fehlerquellen auf, die zwar meist nicht exakt quantitativ abschätzbar, oft jedoch qualitativ erkennbar und hypothetisch vermutbar sind. Solche Fehlerquellen können klassifiziert werden in systematische und zufällige Fehler. *Der systematische Fehler verzerrt die Ergebnisse immer in einer bestimmten Richtung.* Da diese Verfälschung des Ergebnisses keinen Variationen unterliegt und für die gesamte Untersuchung in dieselbe Richtung weist, spricht man vom systematischen Fehler. Wegen ihrer Gleichrichtung können sie durch den Meßvorgang selbst nicht entdeckt werden, obgleich die Meßergebnisse ungültig sind. (So kann man sich vorstellen, daß ein Wert y_1 der abhängigen Variablen der gültige und richtige wäre, bei jeder Messung jedoch ein Wert von y_2 auftritt, der sich durch den Betrag z von y_1 unterscheidet.)

Der zufällige Fehler ist ein Meßfehler, der in der Untersuchung oder der Beobachtung entsteht, *der aber das Meßresultat nicht einseitig verzerrend beeinflußt.* Solche Meßfehler streuen zufällig und weichen daher von dem tatsächlichen Meßresultat nach oben und unten ab, sodaß über eine ausreichend große Population der zufällige Fehler um den wahren Meßwert schwankt und sich die zufälligen Fehler gegenseitig aufheben.

Systematische Fehler und zufällige Fehler können sich jeweils auf die Beobachtung selbst oder auf die Untersuchung insgesamt beziehen (vgl. TRAXEL, W., Grundlagen und Methoden der Psychologie, Stuttgart 1974). *Beobachtungsfehler kommen dabei durch die Beobachtung oder das Meßinstrument zustande, während Untersuchungsfehler sich darauf beziehen, daß die Fehlerquelle in der Untersuchung selbst, also z.B. in der Anlage und Durchführung der Untersuchung liegen.*

Ein *systematischer Untersuchungsfehler* läge z.B. dann vor, wenn man in der Absicht, die Frequenz der Bibliotheksbenutzung feststellen zu wollen, jeweils morgens zwischen 8 und 9 Uhr eine Stichprobe der Bibliotheksbenützer ziehen würde, weil man davon ausgehen kann, daß frühaufstehende Studenten atypisch für alle Studenten sind. Man wird zu diesem Zeitpunkt also eine systematische Verzerrung derart erzielen, daß man die Bibliotheksfrequenz vermutlich unterschätzt. Ein *zufälliger Untersuchungsfehler* läge vor, wenn bei eben dieser Bibliotheksfrequenzfeststellung täglich eine unterschiedliche Geschlechtsproportion ermittelt würde, wenn davon ausgegangen werden kann, daß sich die Geschlechter im Hinblick auf frühmorgendliche Aktivitäten nicht unterscheiden.

Zufällige Fehler gleichen sich mit zunehmender Größe der Population tendenziell aus und variieren um den tatsächlich gültigen Mittelwert der Meßreihe. Systematische Fehler hingegen verzerren die Untersuchungsergebnisse, sodaß es – wie bereits ausgeführt – im Experiment darauf ankommt, die systematischen Fehlerquellen zu beseitigen, mindestens aber zu reduzieren. Hierzu dienen die verschiedenen Kontrollmethoden bzw. die unterschiedlichen experimentellen Konfigurationen. Während der Zufallsfehler durch statistische Verfahren abschätz- und berechenbar ist, kann man bei den systematischen Fehlern nie sicher sein, sie ausgeschaltet oder reduziert zu haben. Es erscheint wichtig darauf hinzuweisen, daß nicht notwendigerweise Unabhängigkeit zwischen Zufallsfehlern und systematischen Fehlern besteht. Vielmehr ist es so, daß dann, wenn systematische Fehlerquellen nicht ausgeschaltet werden können, die statistische Absicherung der erzielten Ergebnisse unbrauchbar wird oder werden kann. Andererseits hat der zufällige Fehler den Vorteil, keine systematischen Verzerrungen zu bewirken und statistisch kontrolliert werden zu können, sodaß jeder Versuchsleiter im Experiment sein Hauptaugenmerk darauf wird richten müssen, systematische Störfaktoren und Fehlerquellen zu ermitteln und auszuschalten.

Zwei wichtige systematische Untersuchungsfehler sind in der Sozialpsychologie durch eine Reihe von Experimenten belegt und wegen ihrer Bedeutsamkeit mit eigenen Namen versehen worden. Einmal *der Versuchsleitereffekt, benannt nach ROSENTHAL, und der Versuchspersonenfehler, benannt nach einem Experiment der Industriesoziologie: der HAWTHORNE-Effekt.*

Als in den Hawthorne-Werken ein Experiment durchgeführt wurde, das klären sollte, in welchem Zusammenhang die Lichthelligkeit und die Steigerung der Arbeitsleistung stünden, machte man folgende interessante Entdeckung: Gleichgültig, wie die unabhängige Variable (Lichthelligkeit) variiert wurde, es wurde jedesmal eine Leistungssteigerung verzeichnet. Da im Experiment gleichwohl alle anderen Faktoren konstant gehalten wurden (wie man glaubte), war das erzielte Ergebnis praktisch nicht zu interpretieren. Erst die weitere Überlegung (und das Einführen des weiteren Faktors), daß die erhöhte Arbeitsleistung möglicherweise darauf zurückzuführen ist, daß die Personen, die am Experiment teilnahmen, ein Bewußtsein entwickelt hatten, das ihre Bedeutung und Wichtigkeit als Versuchspersonen und das Interesse, das an sie herangetragen wurde, in den Vordergrund rückte, was zu einer Steigerung der Leistungsmotivation führte, welche letztendlich auch eine erhöhte Leistung hervorbrachte. *Die Störvariablen „Bewußtsein" und „Erwartungen an das Experiment"* im Zusammenhang mit der Tatsache, als Untersuchungspersonen für das Experiment wichtig zu sein, führte dazu, daß Störeffekte in der abhängigen Variablen zu verzeichnen waren. *Wenn Erwartungen von Versuchspersonen bezüglich des Untersuchungsgegenstandes eine verfälschende Rolle spielen, spricht man daher vom Hawthorneeffekt.* Die experimentelle Situation als solche, die Rekrutierung der Experimental- und Kontrollgruppen, die Versuchsanordnungen und die Anweisungen, die im Experiment gegeben werden, können also via Erwartungen der Versuchspersonen zu verzerrenden Ergebnissen führen.

Als Rosenthaleffekt bezeichnet man den Sachverhalt, daß durch den Versuchsleiter unbewußt Einflüsse auf die Untersuchungspersonen derart ausgeübt werden, daß die Untersuchungsergebnisse im Sinne der Untersuchungshypothesen verzerrt werden. Da jeder Forscher danach trachtet, seine Untersuchungshypothese zu verifizieren – (forschungspsychologisch ist die wissenschaftstheoretische Forderung nach Falsifikation der aufgestellten Hypothesen relativ unsinnig) – wird der Forscher, wenn er zugleich Versuchsleiter ist, sich unbewußt so verhalten, daß seine positiven Erwartungen bestätigt werden. Sozialpsychologische Erkenntnisse zeigen nun, daß durch diese Erwartungshaltung verbale und nonverbale Kommunikationen unbewußt stattfinden und von den Versuchspersonen aufgenommen werden, was letztere dazu veranlaßt, sich im Sinne der von ihnen erwarteten Erwartungen des Versuchsleiters zu verhalten. Der Rosenthaleffekt ist also ein Versuchsleitererwartungseffekt, dessen Wirksamkeit im übrigen nicht nur in menschlichen, sondern gerade auch in Tierexperimenten nachgewiesen werden konnte. Seine verzerrende Wirkung im Sinne einer self-fulfilling prophecy der Hypothese wurde erst sehr spät erkannt und kann, da die zugrundeliegenden Prozesse unbewußt ablaufen und nicht in jedem Falle völlig auszuschalten sind, nur sehr schwer kontrolliert werden.

Der sog. Pygmalioneffekt als Spezialfall des Rosenthaleffektes kann die tatsächliche Beeinflussung durch Erwartungshaltungen bei der abhängigen Variablen (durchaus auch über weitere intervenierende Variable) belegen: die vermeintliche Intelligenzprüfung in einer Schulklasse und die Mitteilung der fiktiven Intelligenzquotienten der einzelnen Schüler an den Lehrer führten dazu, daß die als scheinbar intelligenter eingestuften Schüler nach einer gewissen zeitlichen Periode tatsächlich einen größeren Punktezuwachs im IQ aufwiesen als jene Schüler, denen vermeintlich ein schlechterer IQ zugedacht worden war. Die Erwartungshaltung des Lehrers, daß der Schüler X ein überdurchschnittliches Intelligenzvermögen habe, führte unbewußt dazu, daß der Lehrer häufiger mit X interagierte, ihn damit unbewußt stärker förderte, was seine Intelligenzleistung im Test tatsächlich in die Höhe steigen ließ.

Neben dem Versuchsleitererwartungseffekt können durchaus andere Versuchsleiterwirkungen auftreten, wie z.B. äußere Erscheinung, Geschlecht etc. Will man diese eliminieren, so müßte man in seiner experimentellen Versuchsanordnung solche und ähnliche Faktoren konstant halten, was allerdings in letzter Konsequenz bedeuten würde, daß kaum eine experimentelle Untersuchung durchgeführt werden könnte, weil die allzu häufig sehr manifest vorhandenen ökonomischen, zeitlichen und materiellen Restriktionen der Kontrolle aller Faktoren im Wege stehen.

Rosenthaleffekt und Hawthorneeffekt waren Erwartungswirkungen; bei dem Versuchsleiter im ersten, bei den Versuchspersonen im zweiten Falle. Solche Erwartungen können bewußt oder unbewußt provoziert und erzeugt werden. Sind die Erwartungen durch die Versuchsperson *selbst produziert, so spricht man von Autosuggestion, werden sie durch andere als durch die Versuchspersonen hervorgerufen, so handelt es sich um Heterosuggestion.* Auto- und Heterosuggestion spielen

in der medizinisch-pharmakologischen Forschung eine erhebliche Rolle, weil als Spezialfall davon der *Placeboeffekt* diskutiert wird. *Mit Placebos werden solche Medikamente bezeichnet, die sich äußerlich nicht von wirksamen Medikamenten unterscheiden, tatsächlich aber keine Wirkstoffe enthalten.*

Der Placeboeffekt äußert sich in zweifacher Weise: gibt man einem Patienten ein Placebo, beläßt ihn aber in dem Glauben, es handele sich um ein wirksames Präparat, so kann gelegentlich festgestellt werden, daß eine vermeintliche oder tatsächliche Wirkung des Placebos auftritt. Die vermeintliche Wirkung äußert sich in dem subjektiven Befinden des Patienten, während eine tatsächliche Wirkung als objektives Krankheitsbild festgestellt wird. Es ist klar, daß das Placebo als wirkungslose Droge nicht unmittelbar irgendwelche Effekte hervorrufen kann. Wenn aber trotzdem solche Wirkungen eintreten, so geschieht dies über die Autosuggestion des Patienten, die sich in einer Erwartungshaltung bezüglich der Heilwirkung äußert. Der Glaube an die Wirksamkeit der Droge und die Erwartung einer Besserung des Krankheitszustandes führen zu einer Verbesserung des Krankheitsverlaufs, zu einer Reduzierung der Schmerzen etc. Obgleich der Placeboeffekt methodisch als Autosuggestion zu behandeln ist und zu verzerrenden experimentellen Resultaten führen kann (wenn es z.B. darum geht, die Wirksamkeit eines Medikamentes zu testen), ist der Placeboeffekt auch insoweit positiv zu würdigen, als er in ärztlicher Therapie medikamentös und psychologisch genutzt werden kann.

So wie der Arzt bewußt einen Placeboeffekt herbeiführen kann, indem er dem Patienten suggeriert, das zu verabreichende Medikament wäre ein besonders wirksames, so können auch unbewußte, heterosuggestive Einflüsse vom behandelnden Arzt ausgehen, wenn er selbst sich von einem zu erprobenden Medikament eine entsprechend gute Wirkung erhofft (vgl. hierzu den Versuchsleitereffekt). Der Placeboeffekt kann also sowohl durch Autosuggestion wie auch durch Heterosuggestion in der Therapie psychologisch zum Wohle des Kranken genutzt werden.

Beim Testen von Medikamenten jedoch wirkt der Placeboeffekt als Störvariable, die es auszuschließen gilt. Die Kontrolle des Placeboeffektes bzw. allgemein der Autosuggestion und Heterosuggestion im psychologischen Experiment erfolgt durch den Blindversuch oder den doppelten Blindversuch. *Beim einfachen Blindversuch erfährt der Proband nichts über die Art, den Inhalt und den Zweck des psychologischen Experimentes,* bzw. auf die ärztliche Praxis angewandt: dem Patienten wird verschwiegen, ob er ein Placebo oder eine wirksame Droge verabreicht bekommt. Zwar ist die Unwissenheit keine Garantie dafür, daß bei den Versuchspersonen doch nicht irgendwelche Vermutungen über Sinn und Zweck des Experiments aufkommen, doch wird durch die Unwissenheit immerhin eine erhöhte Wahrscheinlichkeit dafür gegeben sein, daß autosuggestive Effekte nur reduziert, wenn überhaupt auftreten.

Während der einfache Blindversuch nur autosuggestive Effekte ausschließt, will der doppelte Blindversuch zusätzlich potentielle, heterosuggestive Wirkungen kontrollieren. *Beim doppelten Blindversuch wissen sowohl die Versuchspersonen wie*

auch die Versuchsleiter (hier verstanden als jene Personen, die mit den Versuchspersonen im Experiment in unmittelbarem Kontakt stehen) *nichts über die konkrete Zusammensetzung der Experimental- und Kontrollgruppe.* Auf das pharmakologische Experiment bezogen heißt dies, daß der Placeboeffekt dadurch auszuschalten versucht wird, daß weder der die Medikamente verabreichende Arzt noch der die Drogen konsumierende Patient wissen, ob mit einem Placebo oder mit einem wirksamen Medikament behandelt wird. Nur der eigentliche Untersuchungsleiter, der Forscher selbst, ist natürlich darüber informiert, welche Personen zur Kontroll- und welche zur Experimentalgruppe gehören.

Tab. 18. Der Placeobeffekt und seine Kontrolle

Placeboeffekt / Kontrolle	Autosuggestion	Heterosuggestion
Blindversuch (Versuchspersonen nicht informiert)	–	+
doppelter Blindversuch (Versuchspersonen und Versuchsleiter nicht informiert)	–	–
keine Kontrolle	+	+

\+ = Effekt tritt auf
– = Effekt tritt tendenziell nicht auf

In der Auswertung der Untersuchungsergebnisse durch den Forscher sollte sich zeigen, ob durch den doppelten Blindversuch – als optimaler Fehlervermeidungsstrategie von suggestiven Einflüssen – der Placeboeffekt vermieden und die Wirksamkeit der zu testenden Droge festgestellt werden können. Immerhin kann als gesichert festgehalten werden, daß der doppelte Blindversuch zu gültigeren Resultaten führt als der einfache Blindversuch, und dieser wiederum gültigere Resultate liefert als das einfache Experiment, bei dem suggestive Wirkungen überhaupt nicht kontrolliert werden. (Vgl. hierzu auch die ausführliche Darstellung des Placeboeffekts in RATHGEBER, W., (Hrsg.), Medizinische Psychologie, München 1977, 2. Aufl., S. 300 – 308.)

2.2 *Anwendung und Beispiel*

Für die drei theoretisch und abstrakt angesprochenen Gruppen von experimentellen Versuchsanordnungen, nämlich vorexperimentelle, experimentelle und faktorielle Versuchspläne, soll im weiteren je ein Beispiel bearbeitet werden, an dem – von der Hypothesenkonstruktion bis hin zur Auswertung, Analyse und Interpretation der Ergebnisse – die einzelnen Untersuchungsschritte exemplarisch vorgestellt werden, um den ganzheitlichen Charakter empirischer Forschung deutlich zu machen und das Verständis für die integrale Zusammengehörigkeit der einzelnen Untersuchungsschritte einerseits, wie die der Methodologie, der Methoden und der Statistik andererseits zu fördern.

Das erste Exempel beschäftigt sich mit arbeitsmedizinischen bzw. arbeitssoziologischen Fragestellungen: In einem Betrieb werde unter Einsatz von Maschinen, die z.T. beträchtlichen Lärm erzeugen, ein Gut X produziert. Da der Lärmpegel bei der Produktion einen medizinisch nicht mehr vertretbaren Stand erreicht, sind Arbeitsmediziner und Betriebssoziologe der Auffassung, daß Maßnahmen getroffen werden müßten, den Lärm zu reduzieren. Beide sind sich auch darin einig, daß das permanente Tragen eines Lärmschutzes zwar subjektiv und objektiv die Lärmaufnahme durch die beschäftigten Mitarbeiter reduziert, daß eine solche Maßnahme jedoch andererseits von den Beschäftigten als lästig und unangenehm abgelehnt werden könnte. Beider Anliegen wird es nun sein, die Betriebsleitung davon zu überzeugen, daß Lärmschutzmaßnahmen an den Maschinen erforderlich sind; da diese mit erheblichen Kosten verbunden sein können, wird die Betriebsleitung nur dann Investitionsmaßnahmen der geforderten Art durchführen, wenn die Kosten kompensiert werden können (sei es durch erhöhte Produktivität, durch höhere Preise oder ähnliches). Um der Betriebsleitung die Investition schmackhaft zu machen, wird man die Hypothese aufstellen, daß mit Lärmschutzeinrichtungen die Arbeitsleistung der Mitarbeiter gesteigert würde. Die zu überprüfende Nullhypothese würde also lauten: „Es bestehe keine Beziehung zwischen der Höhe des Lärmpegels und der Arbeitsleistung der Beschäftigten." Diese Hypothese soll nun experimentell überprüft werden, wozu die Betriebsleitung ihre Zustimmung gegeben hat. Der Arbeitsmediziner und der Betriebssoziologe werden aufgefordert, die entsprechenden Maßnahmen zu ergreifen.

Der *zweite Fall* habe folgende realistische Voraussetzungen: Ein großes Pharmazieunternehmen stelle eine Droge X her, deren Wirksamkeit bisher experimentell und durch jahrelange Erprobung belegt ist. Plötzlich treten jedoch Lieferschwierigkeiten derart auf, daß ein bestimmter Bestandteil Y, der für die Drogenherstellung bisher benötigt wurde, nicht mehr beschafft werden kann. Da die Wirkungszusammenhänge der einzelnen Bestandteile innerhalb der Droge theoretisch nicht voll bekannt sind und somit nicht abgeleitet werden kann, ob der Bestandteil Y der Droge einen erheblichen Einfluß auf die Veränderung des Krankheitsbildes hat, er also ein notwendiger Bestandteil der Droge ist, wird das Forschungslabor dieses Unternehmens damit beauftragt, die potentielle Wirkgröße des Faktors Y innerhalb der Droge experimentell zu testen.

Der dritte Fall kommt aus dem klinischen Bereich: Zwei Krebsforscher können sich nicht darüber einigen, wie bestimmte Formen des Magenkrebses zu behandeln sind. Während der eine Priorität bei der operativen Entfernung des Geschwulstes sieht, glaubt der andere, daß eine radiologische Behandlung größere Erfolgsaussichten zeitigt. Da dieser Streit nicht ausschließlich theoretisch entscheidbar ist, beide aber daran interessiert sind, zu gültigen Aussagen zu kommen, beschließen sie, ein Experiment durchzuführen, das ihnen erste Aufschlüsse über die Wirksamkeit der unterschiedlichen Behandlungsmethoden geben soll.

Die zu überprüfenden und statistisch abzusichernden Nullhypothesen in den beiden letzten Fällen würden demnach lauten: „Der Bestandteil Y der Droge X ist für die Wirksamkeit des Medikamentes nicht von Bedeutung" und „Es bestehe kein Unterschied zwischen den beiden Behandlungsformen operative Entfernung bzw. Bestrahlung".

2.2.1 Instrumentarium

Nachdem die Nullhypothesen formuliert worden sind, geht es darum, das Erhebungsinstrument so zu konstruieren, daß eine optimale Überprüfung der formulierten Hypothesen möglich erscheint. Bevor das Erhebungsinstrument konstruiert werden kann, muß zunächst eine *Methodenentscheidung* getroffen werden. Diese erfolgt unter den Gesichtspunkten einer Optimierung der Hypothesenüberprüfung, d.h. es werden die vorhandenen empirischen Methoden daraufhin überprüft, ob und in welcher Weise sie (theoretisch wie praktisch) in der Lage sind und geeignet erscheinen, die vermuteten Sachverhalte zu testen. Da die vorgenannten drei Beispiele jedoch innerhalb des Kapitels zum Experiment stehen, ergibt sich die Methodenentscheidung automatisch. Wenn aber das Experiment jene Methode ist mit dem die oben formulierten Hypothesen getestet werden sollen, dann müssen im weiteren die experimentellen Versuchsanordnungen bestimmt werden, die die Möglichkeit der Hypothesenprüfung optimal eröffnen.

Beispiel 1: Wie muß also das Erhebungsinstrument im ersten Fall aussehen? Da die Betriebsleitung die Auflage gemacht hat, unter Kostenminimierungsgesichtspunkten das Experiment durchzuführen, wird man aus extra-inhaltlichen Gründen eine möglichst einfache Versuchsanordnung wählen müssen. Arzt und Soziologe verständigen sich darauf, die Arbeitsleistung zum Zwecke des Vergleichs vor und nach Installierung des Lärmschutzes zu messen. Beide sind sich darin einig, daß eine einmalige Messung nicht genügen kann, sondern daß über einen längeren Zeitraum hinweg und an mehreren Maschinen die Arbeitsleistung der Mitarbeiter festgehalten wird, wobei die Maschinen, die später mit Lärmschutzanlagen bestückt werden sollen, zufällig ausgewählt wurden. Es handelt sich also um eine vorexperimentelle Versuchsanordnung des folgenden Typus:

$$M_1 \quad S \quad M_2$$

Mit dieser Versuchsanordnung können Meßeinflüsse nicht ausgeschlossen werden,

ebensowenig solche, die aus situativen Veränderungen heraus nicht konstantgehalten werden können. Da aber nur die abhängige Variable gemessen wird (wir gehen in dem Beispiel davon aus, daß die Reduzierung des Lärmpegels durch entsprechende Schutzeinrichtungen gelingt und daß die Höhe der Reduzierung zunächst keinen eigenen Einfluß auf die abhängige Variable ausübt) und diese als Produktionszahl eine objektive Größe darstellt, können Verzerrungen durch Meßeinflüsse ausgeschlossen werden. Aus Gründen der Vereinfachung setzen wir alle anderen Variablen unter ceteris-paribus-Bedingungen als unveränderlich und konstant an.

Beispiel 2: Die Forschungsabteilung in dem Pharmazieunternehmen hat sich ebenfalls zu einer experimentellen Vorgehensweise entschlossen. Man glaubt aber dort nicht darauf verzichten zu können, echte experimentelle Versuchsanordnungen zu schaffen, um Experimental- und Kontrollgruppen miteinander vergleichen zu können. (In dem Beispiel der Lärmreduzierung wurden ja die Messungen jeweils an derselben Gruppe vorgenommen.) Sie konstruieren eine Experimental- und eine Kontrollgruppe, die beide durch Randomisierung gewonnen sind (selbstverständlich wäre auch Parallelisierung möglich gewesen). Während die Experimentalgruppe mit dem Medikament X, das den Bestandteil Y enthält, behandelt wird, erhält die Kontrollgruppe ein sonst gleiches Medikament, dem jedoch der Bestandteil Y fehlt. Die so geschaffene Versuchsanordnung entspricht der folgenden echt experimentellen Konfiguration:

$$\begin{array}{ccc} M_1 & S & M_3 \\ M_2 & & M_4 \end{array}$$

Man könnte sich durchaus vorstellen, daß die Messungen 1 und 2 entfallen und gleichwohl die Wirksamkeit des Drogenbestandteils Y überprüft werden kann. Verzichtet man jedoch auf sie, so werden die Resultate ungenauer, weil der zeitliche Einfluß nicht kontrolliert ist. Daher erscheint diese klassische Anordnung des Experimentes als bessere Vorgehensweise der Hypothesenüberprüfung. Um Placeboeffekte auszuschalten, wird das Experiment als doppelter Blindversuch durchgeführt und wir setzen voraus, daß die abhängige Variable objektiv gemessen werden kann. D.h. nicht die subjektive Befindlichkeit des Patienten nach Verabreichung der Droge mit bzw. ohne Bestandteil Y wird ermittelt, sondern das objektive Krankheitsbild ist feststellbar (z.B. Blutdruck).

Beispiel 3: Die beiden konkurrierenden Therapieformen des Magenkrebses werden ebenfalls experimentell getestet. Allerdings sind die beiden rivalisierenden Kollegen übereingekommen, nicht nur überprüfen zu wollen, ob die eine oder die andere Methode die bessere ist, sondern ob eine Kombination beider Methoden eventuell zusätzliche Erfolgschancen zeitigt. Daher hat man sich zu einem faktoriellen Design entschlossen, das die Beantwortung dieser Frage ermöglicht. Auch konnte man sich darauf verständigen, daß die Bestrahlungsdosis variiert wird. Während also die eine Variable nur dichotomisiert ist (operativ entfernt = y oder nicht entfernt = $\sim$ y), soll die Bestrahlungstherapie trichotomisiert werden (niedrige = x_3, mittlere = x_2 und hohe Bestrahlungsdosis = x_1). Da diese Variable nicht jeweils

unabhängig gemessen werden soll, was der „normalen" echten, experimentellen Versuchsanordnung entsprechen würde, sondern die Variablen jeweils in Kombination ihrer Merkmalsausprägungen auftreten sollen, um interaktive Wirkungen feststellen zu können, erfordert der faktorielle Versuchsplan nicht nur eine Kontroll- und eine Experimentalgruppe, sondern insgesamt 6 Untersuchungsgruppen (2 x 3 Merkmalsausprägungen der beiden Variablen). Die faktorielle Versuchsanordnung hat demnach das Aussehen der im Abschnitt 2.1.4 vorgestellten faktoriellen Versuchsplanung. Dieser faktorielle Versuchsplan bietet die Möglichkeit, die Wirksamkeit der einzelnen Therapieformen, wie auch der kombinierten Therapie abzuschätzen und statistisch abzusichern.

Unklar ist bisher geblieben, wie die abhängige Variable (die Wirksamkeit der Therapieformen) operationalisiert und gemessen werden soll. Da die subjektive Befindlichkeit der Patienten nach Anwendung der Therapie durchaus fehlerbehaftet sein kann und keineswegs mit dem objektiven Zustand kompatibel sein muß, wird eine Operationalisierung gesucht, die eine gültige Messung der abhängigen Variablen ermöglicht. Auch hierbei verständigen sich die beiden Mediziner: sie setzen als Operationalisierung der abhängigen Variablen die Überlebensdauer in Monaten nach der Therapierung fest. (Es versteht sich von selbst, daß dieses Beispiel – wie alle anderen – fiktiv ist und daß davon ausgegangen wird, daß alle potentiellen Störgrößen kontrolliert sind, im letzten Fall also insbesondere auch das Lebensalter der Patienten.)

2.2.2 Die empirischen Daten

Beispiel 1: An zufällig ausgewählten Wochentagen und Maschinen wird die durchschnittliche Arbeitsleistung der Versuchspersonen gemessen, bevor die Lärmschutzeinrichtungen installiert werden. Man habe dabei an den 14 ausgewählten Tagen folgende Arbeitsleistung erhalten:

150, 146, 149,152, 145, 155, 160, 148, 150, 153, 151, 149, 150, 145
$\bar{x} = 150{,}2 \quad s = 4{,}02$

Nach Erstellung der Lärmschutzeinrichtungen wurde eine zweite Meßreihe analog den genannten Bedingungen festgestellt. Diese Meßreihe enthalte die folgenden Werte:

155, 150, 160, 163, 148, 152, 158, 163, 167, 159, 161, 155, 148, 149
$\bar{x} = 156{,}4; \quad s = 6{,}12$

Der Vergleich beider Meßreihen scheint zunächst anzudeuten, daß die aufgestellte Hypothese zutrifft: nach Installierung der Lärmschutzeinrichtungen ist die Arbeitsleistung angestiegen. Auch das arithmetische Mittel der zweiten Meßreihe liegt deutlich über dem der ersten. Aus dieser deskriptiven Information auf eine Regelmäßigkeit oder gar Gesetzmäßigkeit zu schließen, wäre jedoch verfrüht. Wir müssen statistische Überlegungen darüber anstellen, ob nicht rein zufällig eine Stei-

gerung der Arbeitsleistung hätte entstehen können, ohne daß die Lärmschutzeinrichtungen dafür ursächlich verantwortlich zu machen wären. Diese Frage kann durch eine statistische Analyse der Daten beantwortet werden, die im Abschnitt 2.2.3 vorgenommen wird.

Beispiel 2: Die Experimental- und die Kontrollgruppen bestehen jeweils aus 100 Personen ($n_1 = n_2 = 100$), wobei alle 100 Personen an Hypertonie leiden. Die Messungen M_1 und M_2 (also jeweils vor Verabreichung des Medikamentes bzw. des Placebos) erbrachte eine Meßwertverteilung mit dem arithmetischen Mittel $\bar{x}_1$ für die Messung M_1 von 200 und $\bar{x}_2$ für die Messung M_2 von 190. Nach Verabreichung der Droge mit dem Bestandteil Y ergibt sich ein arithmetisches Mittel des Blutdruckes in der Experimentalgruppe von $\bar{x}_3 = 150$, während in der Kontrollgruppe das Medikament X ohne Bestandteil Y nur zu einer Reduzierung des Bluthochdruckes auf 180 führt. Auch in diesem Beispiel scheint die Hypothese bestätigt zu werden, daß der Bestandteil Y der Droge X einen erheblichen Beitrag zur Blutdruckssenkung leistet. Aber auch hier muß erst ein statistischer Test ergeben, ob diese Blutdruckreduktion zufällig entstanden ist oder tatsächlich dem Bestandteil Y zugeschrieben werden kann.

Beispiel 3: Die Durchführung der faktoriellen Versuchsanordnung erbrachte jene Resultate, wie sie in der Tab. 19 zusammengestellt sind.

Tab. 19. Ergebnisse des Experiments (Überlebensdauer in Monaten)

	x_1	x_2	x_3
y	20	18	20
	18	20	17
	25	22	17
	23	21	17
	24	19	19
~ y	19	18	17
	17	16	14
	19	18	13
	18	14	15
	17	14	16

x = Bestrahlungsdosis
y = operative Entfernung

Zunächst bietet diese Tabelle ein verwirrendes Bild, das eine einheitliche Interpretation allein aufgrund der Inspektion der Tabelle offensichtlich nicht zuläßt. Es müssen daher Verfahren gefunden werden, die die unübersichtliche Tabelle auf informationshaltigere Aussagen reduzieren können, die im Sinne der formulierten Hypothesen interpretierbar erscheinen. Eine solche Methode ist die Varianzanalyse,

die wir im Abschnitt 2.2.3 kennenlernen werden und die es ermöglicht, die in der Tabelle aufscheinende Variation der abhängigen Variablen auf die eine, die andere oder beide unabhängige Variable zurückzuführen.

2.2.3 Analyse der Daten

Beispiel 1: Um uns eine systematische Übersicht über die Datenlage zu geben, stellen wir die beiden Meßwertreihen einander gegenüber und ermitteln die jeweiligen quadrierten Differenzen der Meßwerte sowie das arithmetische Mittel der einzelnen Meßwertreihen und deren Differenz. Aus der Summe der quadrierten Differenzen könnte man nun eine Maßzahl für den Einfluß der unabhängigen Variablen konstruieren, denn offensichtlich gilt: je größer die Summe der quadrierten Differenzen wird, desto stärker der Einfluß der unabhängigen Variablen auf die Arbeitsleistung.

Tab. 20. Berechnung des t-Tests für abhängige Stichproben

Stichproben	x_{i_1}	x_{i_2}	$d_i = x_1 - x_2$	d_i^2
1	150	155	− 5	25
2	146	150	− 4	16
3	149	160	−11	121
4	152	163	−12	144
5	145	148	− 3	9
6	155	152	+ 3	9
7	160	158	+ 2	4
8	148	163	−15	225
9	150	167	−17	289
10	153	159	− 6	36
11	151	161	−10	100
12	149	155	− 6	36
13	150	148	+ 2	4
14	145	149	− 4	16
n = 14	$\bar{x}_1 = 150{,}2$	$\bar{x}_2 = 156{,}4$	$\bar{d} = -6{,}14$	1034

Da die beiden Meßreihen an ein und derselben identischen Stichprobe gewonnen wurden, sind also die beiden Stichproben voneinander abhängig: man spricht von korrelierenden Stichproben. (Durch die Auswahl eines Meßobjektes z bei der ersten Meßwertreihe war die zweite Meßwertreihe im Hinblick auf das Meßobjekt z bereits determiniert, d.h. z mußte notwendigerweise auch in die zweite Meßwertreihe gelangen; daher keine unabhängigen Stichproben.) Für den Fall korrelierender Stichproben und einem Meßniveau der abhängigen Variablen, das mindestens intervallskaliert ist (wobei eine Normalverteilung dieses Merkmales in der Grundgesamtheit vorausgesetzt wird), ist in der Statistik der *t-Test für abhängige Stich-*

proben entwickelt worden. Aus der obigen Tabelle kann ein bestimmter t-Wert nach der noch anzugebenden Formel berechnet werden. Dieser t-Wert wird mit einem „theoretischen" t-Wert verglichen, der die Grenze dafür angibt, ob der empirisch festgestellte Unterschied zufällig oder systematisch entstanden ist. Für solche kritischen t-Werte oder Grenzwerte von t findet man in jedem Statistik lehrbuch Tabellen, die zur Ermittlung des kritischen t-Wertes bei einer bestimmten Zahl von Freiheitsgraden dienen. Die Zahl der Freiheitsgrade errechnet sich dabei aus der Summe aller Meßwerte minus 2 (generell, also auch für unabhängige Stichproben, errechnet sich die Zahl der Freiheitsgrade aus Stichprobengröße der 1. Gruppe minus 1 und Stichprobengröße der 2. Gruppe minus 1). Mit diesen Informationen kann nun aufgrund der folgenden Formel der Signifikanztest durchgeführt werden:

$$t = \frac{\bar{d}}{\frac{\Sigma d_i^2 - n \cdot \bar{d}^2}{n(n-1)}} = \frac{-6{,}14}{\frac{1034 - 14 \cdot (-6{,}14)}{14\,(14-1)}} = 2{,}48$$

Werden die errechneten Werte der Tab. 20 in die Formel eingesetzt, so ergibt sich daraus ein t-Wert von 2,48. Der kritische Tabellenwert für 26 Freiheitsgrade und ein Signifikanzniveau von $\alpha = 0{,}05$ (bei zweiseitiger Fragestellung) liegt bei 1,703. Da unser empirisch gewonnener Wert größer ist als der kritische Tabellenwert, können wir die aufgestellte Nullhypothese, daß keine Beziehung zwischen der Lärmreduktion und der Arbeitsleistung der Arbeiter besteht, zurückweisen. Die Irrtumswahrscheinlichkeit, die wir dabei in Kauf nehmen, beträgt aufgrund des gewählten Signifikanzniveaus 5%. Von daher sind wir nun in der Lage, die Alternativhypothese, daß eine Reduktion des Lärms zu einer erhöhten Arbeitsleistung führt, zu akzeptieren. (Man beachte allerdings,daß diese Alternativhypothese nicht mit einer Wahrscheinlichkeit von 95 % zutrifft!)

Beispiel 2: Analog dem t-Test-Verfahren im Beispiel 1 könnte man auf die Untersuchungsgruppen bezogen jeweils einen abhängigen t-Test berechnen, der Erkenntnisse darüber liefern könnte, inwieweit sich Veränderungen zwischen den ersten und den zweiten Meßwerten ergeben haben ($M_1 - M_3$ bzw. $M_2 - M_4$). Eine solche Vorgehensweise wäre aber wenig fruchtbar, weil sich daraus keine Aussage darüber ableiten läßt, inwieweit sich die Meßwerte der zweiten Messung innerhalb der Experimentalgruppe von der der Kontrollgruppe unterscheiden. Dies ist aber das eigentliche Ziel der Untersuchung, das in der Hypothese formuliert wurde.

Der unmittelbare Vergleich der Messung M_3 und M_4 wäre ebenfalls nicht aussagekräftig genug, weil die Vorhermessungen M_1 und M_2 nicht in die Analyse miteinbezogen sind. (So kann man sich vorstellen, daß Differenzen zwischen M_4 und M_3 durchaus auch in den Messungen M_2 und M_1 auftreten können. Die Nichtberücksichtigung von M_1 und M_2 würde dann bei alleiniger Analyse von M_3 und M_4 zu Ergebnissen führen, die man als signifikante Unterschiede interpretieren würde, die aber tatsächlich nicht auf den zwischengeschalteten Stimulus zurückzuführen sind,

sondern bereits in den Ausgangsmessungen begründet liegen). Daher können gültige und richtige Aussagen nur dann vorgenommen werden, wenn die Messung M_2 mit M_4 auf der Basis der Messungen M_1 und M_3 verglichen werden. Daher verwendet man zur Signifikanzprüfung in dem Falle des klassischen Experimentes die Meßwertdifferenzen zwischen M_3 und M_1 und M_4 und M_2. Auch hierbei wird wieder der t-Test angewandt, diesmal allerdings in modifizierter Form, indem nämlich die Differenz der Messungen, die in der Experimentalgruppe aufgetreten sind, in den Zähler geschrieben und die Meßwertdifferenz der Kontrollgruppen in den Nenner genommen werden. Überschreitet dieser Bruch wieder einen bestimmten kritischen Wert, so wird der aufgetretene Unterschied in den Differenzen als signifikant interpretiert, d.h. die im Experiment variierte, unabhängige Variable ist mit einer bestimmten Irrtumswahrscheinlichkeit für die Meßwertdifferenz verantwortlich. Die Durchführung dieser Rechenoperation erbringt einen t-Wert, der größer ist als der kritische Tabellenwert, sodaß auch in diesem Beispiel davon ausgegangen werden kann, daß mit einer Irrtumswahrscheinlichkeit von 95 % die aufgestellte Nullhypothese zurückgewiesen werden kann. Der fehlende Bestandteil Y der Droge X hat eine den Hochdruck reduzierende Wirkung.

Beispiel 3: Die Tabelle, die die Resultate der faktoriellen Versuchsanordnung wiedergegeben hat, kann zur Grundlage der *Varianzanalyse* gemacht werden, die die folgenden Fragestellungen zu beantworten hat:

1. Unterscheiden sich die einzelnen Meßwerte rein zufällig oder sind die Unterschiede auf das Wirken der unabhängigen Variablen zurückzuführen (dies ist die Frage der Signifikanzprüfung).

2. Es wird gefragt, welche Therapieformen (unabhängige Variablen) für die in der Tabelle feststellbaren Variationen der abhängigen Variablen verantwortlich zu machen ist (Operation allein, Behandlungstherapie allein oder beide gemeinsam).

3. In welchem Ausmaß sind die unabhängigen Variablen in der Lage, die abhängige Variable zu determinieren?

Das gewählte Beispiel weist in der Tabelle eine Einfachststruktur auf, um das Verfahren der Varianzanalyse leicht nachvollziehbar und verstehbar zu machen. So wäre es nicht notwendig gewesen, jeweils eine gleiche Gruppengröße von 5 zu haben, oder die Variablen zu dichotomisieren oder zu trichotomisieren, oder sich auf zwei Variablen zu beschränken. Tatsächlich können alle die genannten Faktoren modifiziert und variiert werden, ohne daß das Verfahren der Varianzanalyse sich in der grundlegenden Methode verändern würde. Die oben gewählte Einfachststruktur kann also jederzeit erweitert werden, wenn das Verfahren selbst verstanden ist.

Die Varianzanalyse ist ein Verfahren, das versucht, die in der abhängigen Variablen aufgetretenen Variationen der oder den unabhängigen Variablen zuzuordnen. Dabei wird von der Überlegung ausgegangen (die für die Varianzanalyse grundlegend ist), daß sich die gesamte Variation der abhängigen Variablen zusammensetzt aus der Variation, die durch die unabhängigen Variablen jeweils einzeln verursacht, die durch die Interaktion(en) der unabhängigen Variablen determiniert wird und aus einer Fehler- oder Restvariation, die im Rahmen des varianzanalytischen

Modells (Beschränkung auf wenige Variablen etc.) nicht erklärt werden kann.

$$V = V_x + V_y + V_{xy} + V_e$$

Für unseren Beispielfall ergäbe sich somit die oben formulierte Grundgleichung der Varianzanalyse. Wie kommt man nun mit der Varianzanalyse zur Lösung der Grundgleichung?

Zunächst sei die Fehlervarianz V_e besprochen. Wir greifen uns willkürlich eine der 6 Untersuchungsgruppen des Experiments heraus und betrachten die Meßwerte der abhängigen Variablen. Da innerhalb einer Untersuchungsgruppe die unabhängigen Variablen gleich gewirkt haben (z.B. in der ersten Untersuchungsgruppe x_1 und y), sollte eigentlich zu erwarten sein (da alle anderen Faktoren kontrolliert wurden), daß die abhängige Variable bei allen Versuchspersonen die gleichen Meßwerte aufweist. Tatsächlich jedoch stellen wir Differenzen in den Meßwerten fest, d.h. die abhängige Variable variiert *innerhalb* der Gruppe. Da aber in dieser Gruppe die gleichen Bedingungen geherrscht haben, insbesondere auch in den unabhängigen Variablen, kann diese Variation nicht den unabhängigen Variablen zugeschrieben werden. Diese interne Gruppenvariation kann daher nicht erklärt werden und ist die Rest- bzw. Fehlervarianz.

Diese Fehlervariation kann für jede Gruppe berechnet und aufaddiert werden, sodaß man letztendlich zur gesamten Fehlervariation des Experimentes gelangt. Die Variation für die einzelnen Gruppen berechnet sich aus der quadrierten Abweichung der Meßwerte von dem arithmetischen Mittel der Meßwerte der jeweiligen Gruppe. Wir berechnen zunächst die arithmetischen Mittel der Meßwerte für die einzelnen Untersuchungsgruppen:

$$\begin{aligned}
\bar{z}_1 &= (20 + 18 + 25 + 23 + 24) : 5 = 22 \\
\bar{z}_2 &= (18 + 20 + 22 + 21 + 19) : 5 = 20 \\
\bar{z}_3 &= (20 + 17 + 17 + 17 + 19) : 5 = 18 \\
\bar{z}_4 &= (19 + 17 + 19 + 18 + 17) : 5 = 18 \\
\bar{z}_5 &= (18 + 16 + 18 + 14 + 14) : 5 = 16 \\
\bar{z}_6 &= (17 + 14 + 13 + 15 + 16) : 5 = 15
\end{aligned}$$

Nun kann die Fehlervarianz berechnet werden:

$$V_e = \Sigma(z_i - \bar{z})^2$$

$$\begin{aligned}
V_{e_1} &= (20-22)^2 + (18-22)^2 + (25-22)^2 + (23-22)^2 + (24-22)^2 = 34 \\
V_{e_2} &= (18-20)^2 + (20-20)^2 + (22-20)^2 + (21-20)^2 + (19-20)^2 = 10 \\
V_{e_3} &= (20-18)^2 + (17-18)^2 + (17-18)^2 + (17-18)^2 + (19-18)^2 = 8 \\
V_{e_4} &= (19-18)^2 + (17-18)^2 + (19-18)^2 + (18-18)^2 + (17-18)^2 = 4 \\
V_{e_5} &= (18-16)^2 + (16-16)^2 + (18-16)^2 + (14-16)^2 + (14-16)^2 = 16 \\
V_{e_6} &= (17-15)^2 + (14-15)^2 + (13-15)^2 + (15-15)^2 + (16-15)^2 = 10
\end{aligned}$$

$$V_e = 34 + 10 + 8 + 4 + 16 + 10 = 82$$

Die gesamte Fehlervarianz für das Experiment beträgt demnach 82. Dieser Wert kann offenkundig durch unser Experiment nicht erklärt werden.

Die Begriffe Variation und Varianz werden im Zusammenhang mit der Varianzanalyse synonym gebraucht, obwohl nur der der Variation korrekt ist. Variation war nämlich die Summe der quadrierten Abweichungen der Meßwerte vom arithmetischen Mittel aller Meßwerte. Die Varianz hingegen ist die Summe der quadrierten Abweichung der Meßwerte von dem arithmetischen Mittel aller Meßwerte durch die Zahl der Meßwerte (n) dividiert.

Die Summe aller Variationen für das gesamte Experiment oder Gesamtvarianz (V) errechnet sich aus der Summe der quadrierten Abweichungen der einzelnen Meßwerte vom arithmetischen Mittel aller Meßwerte der gesamten Tabelle. Die folgende Tabelle gibt die Berechnung der Gesamtvariation für die obige Tabelle wieder.

$$\bar{z}_{gesamt} = (\bar{z}_1 + \bar{z}_2 + \bar{z}_3 + \bar{z}_4 + \bar{z}_5 + \bar{z}_6) : 6 = 18{,}17$$

$$V = \Sigma (z_i - \bar{z}_{ges.})^2$$

$$\begin{aligned} V = {} & (20 - 18{,}17)^2 + (18 - 18{,}17)^2 + (25 - 18{,}17)^2 + (23 - 18{,}17)^2 + (24 - 18{,}17)^2 \\ & + (18 - 18{,}17)^2 + (20 - 18{,}17)^2 + (22 - 18{,}17)^2 + (21 - 18{,}17)^2 + (19 - 18{,}17)^2 \\ & + (20 - 18{,}17)^2 + (17 - 18{,}17)^2 + (17 - 18{,}17)^2 + (17 - 18{,}17)^2 + (19 - 18{,}17)^2 \\ & + (19 - 18{,}17)^2 + (17 - 18{,}17)^2 + (19 - 18{,}17)^2 + (18 - 18{,}17)^2 + (17 - 18{,}17)^2 \\ & + (18 - 18{,}17)^2 + (16 - 18{,}17)^2 + (18 - 18{,}17)^2 + (14 - 18{,}17)^2 + (14 - 18{,}17)^2 \\ & + (17 - 18{,}17)^2 + (14 - 18{,}17)^2 + (13 - 18{,}17)^2 + (15 - 18{,}17)^2 + (16 - 18{,}17)^2 \\ = {} & 246{,}17 \end{aligned}$$

Die gesamte Variation der abhängigen Variablen, die durch V_x, V_y, V_{xy} und die Fehlervariation V_e zusammengesetzt ist, beträgt 246,17.

Welches ist nun die durch x verursachte Variation? Die quadrierten Abweichungen des arithmetischen Mittels für die jeweiligen x-Gruppen vom gesamten arithmetischen Mittel multipliziert mit der jeweiligen Anzahl der Gruppengrößen ergibt die durch x verursachte Variation der abhängigen Variablen.

Wir berechnen zunächst die arithmetischen Mittel für die x-Gruppen:

$$\bar{z}_{x_1} = (20 + 18 + 25 + 23 + 24 + 19 + 17 + 19 + 18 + 17) : 10 = 20$$

$$\bar{z}_{x_2} = (18 + 20 + 22 + 21 + 19 + 18 + 16 + 18 + 14 + 14) : 10 = 18$$

$$\bar{z}_{x_3} = (20 + 17 + 17 + 17 + 19 + 17 + 14 + 13 + 15 + 16) : 10 = 16{,}5$$

Das arithmetische Mittel der Meßwerte der beiden Gruppen, die dem Stimulus x_1 ausgesetzt waren, beträgt 20, das bei $x_2 = 18$ und das bei $x_3 = 16{,}5$. In allen drei Fällen handelt es sich um jeweils 10 Untersuchungspersonen, sodaß die Variation, die durch x verursacht wurde, wie folgt berechnet werden kann:

$$V_x = \Sigma (\bar{z}_{x_i} - \bar{z}_{ges.})^2$$

$$V_x = (20 - 18{,}17)^2 \cdot 10 + (18 - 18{,}17)^2 \cdot 10 + (16{,}5 - 18{,}17)^2 \cdot 10 =$$

$$V_x = \mathbf{61{,}6}$$

Mit dieser Berechnung haben wir die durch x verursachte Variation ermittelt, gleichgültig ob y als unabhängige Variable gewirkt oder nicht gewirkt hat (technisch gesprochen über die Zeilen der Tabelle hinweg). In analoger Weise errechnen wir nun die durch y verursachte Varianz – unabhängig davon, in welcher Merkmalsausprägung die unabhängige Variable x vorlag (also technisch gesprochen über alle Spalten hinweg). Diese Variation läßt sich berechnen aus der Summe der quadrierten Differenzen zwischen den Mittelwerten für y und ~ y und dem gesamten arithmetischen Mittel multipliziert mit der jeweiligen Gruppengröße. In beiden Fällen wird also die quadrierte Abweichung mit 15 multipliziert, weil jeweils 15 Personen der unabhängigen Variablen ausgesetzt wurden. Die arithmetischen Mittel für die y-Gruppen betragen:

$$\bar{z}_y = (20 + 18 + 25 + 23 + 24 + 18 + 20 + 22 + 21 + 19 + 20 + 17 + 17 + 17 + 19) : 15 = 20$$

$$\bar{z}_{\sim y} = (19 + 17 + 19 + 18 + 17 + 18 + 16 + 18 + 14 + 14 + 17 + 14 + 13 + 15 + 16) : 15 = 16{,}3$$

Die durch y verursachte Variation errechnet sich nach:

$$V_y = \Sigma(\bar{z}_y - \bar{z}_{ges.})^2$$

$$V_y = (20 - 18{,}17)^2 \cdot 15 + (16{,}3 - 18{,}17)^2 \cdot 15 = 102{,}6$$

Nachdem wir Fehlervariation, Variation durch x und Variation durch y sowie die gesamte Variation der abhängigen Variablen berechnet haben, können wir aus der Varianzgrundgleichung die Interaktionsvarianz, also jenen Effekt, der durch das gemeinsame Wirken von x und y auftreten kann, residual bestimmen.

$$V = V_x + V_y + V_{xy} + V_e$$

$$246{,}2 = 61{,}6 + 102{,}6 + 82 + V_{xy}$$

$$V_{xy} = 0$$

Das obige Beispiel wurde so gewählt, daß eine Interaktionsvarianz nicht gegeben ist. Dieses ist allerdings ein Spezialfall der allgemeinen Varianzanalyse. Um die Interaktionsvarianz inhaltlich zu verdeutlichen, sei noch folgende Illustration gegeben:

Man kann annehmen, daß eine psychosomatische Erkrankung allein durch medizinische oder allein durch psychotherapeutische Behandlung jeweils im Krankheitsbild so verändert wird, daß von einer Besserung gesprochen werden kann. Man kann sich aber auch vorstellen, daß der gemeinsame Einsatz medizinischer und psychotherapeutischer Behandlung keine Besserung des Krankheitsbildes im additiven Sinne der jeweiligen Einzelwirkung der Therapie bringt, sondern daß darüber hinaus ein zusätzlicher Effekt durch das gemeinsame, interaktive Wirken der beiden unabhängigen Variablen erzielt wird. In unserem Beispielfall der Tabelle ist es jedoch so, daß eine Interaktionsvarianz nicht auftritt.

Aus den absoluten Variationsgrößen, die bislang berechnet wurden, schließen zu wollen, daß die durch y verursachte Variation von 102,6 weit über der durch x verursachten von 61,6 liegt, und somit y der wesentlichere Faktor ist, ist verfrüht. Man kann sich nämlich leicht vorstellen, daß die absolute Höhe der errechneten

Variation von der Gruppengröße und ähnlichen Faktoren abhängt. Unterschiedliche Größen sind aber nicht miteinander vergleichbar. Es muß also versucht werden, einen Maßstab zu finden, der solche Vergleiche zuläßt. Hier hilft wieder das Konzept der Freiheitsgrade. Beziehen wir nämlich die Varianzen auf die jeweils zugehörigen Freiheitsgrade, so erhalten wir *Varianzen je Freiheitsgrad,* haben also eine gleiche Basis vorausgesetzt und können die Varianzen sinnvoll und fruchtbar miteinander vergleichen.

Welches sind nun die *Freiheitsgrade für die einzelnen Variationen?* Die Gesamtvarianz hat n – 1 Freiheitsgrade, also Anzahl der Versuchspersonen minus 1. Die durch x verursachte Variation hat so viele Freiheitsgrade, wie die Variable x Merkmalsausprägungen minus 1 aufweist. Die durch y verursachte Variation hat so viele Freiheitsgrade, wie die Variable y Merkmalsausprägungen minus 1 aufweist. Die Zahl der Freiheitsgrade für die Interaktionsvarianz zwischen x und y ergibt sich aus der Multiplikation der Freiheitsgrade von x und y. Die Zahl der Freiheitsgrade für die Fehlervarianz kann residual bestimmt werden.

Berechnen wir, wie in der unten stehenden Tabelle 21 geschehen, die Variation je Freiheitsgrad für x und y, so stellen wir fest, daß in der Tat y offensichtlich einen stärkeren Einfluß auf die abhängige Variable ausübt als x (102,6 zu 30,8). Nun wäre aber durchaus denkbar, daß die nun festgestellten determinativen Wirkungen der unabhängigen Variablen x und y, wie auch die festgestellte unterschiedliche Stärke der Wirkungen rein zufällig entstanden ist. An die bisher durchgeführten Rechenoperationen muß sich demnach ein Signifikanztest anschließen, der die Frage überprüft, ob die Wirkungen der Variablen zufällig oder signifikant sind. Diese Überprüfung, ob Varianzen auf einen systematischen Einfluß von unabhängigen Variablen zurückzuführen sind, wird in dem *F-Test* vorgenommen. Hierbei werden die durch die unabhängigen Variablen erklärten Varianzen jeweils in Beziehung

Tab. 21. Analyse der Varianzen (ANOVA)

	Variation	Freiheitsgrade df	Variation je Freiheitsgrad $\bar{V} = \frac{V}{df}$	F-Wert für den F-Test $F = \frac{\bar{V}}{V_e}$	Signifikanzniveau	Korrelation
Variation durch x	61,6	3 – 1 = 2	$\frac{61,6}{2} = 30,8$	$F_x = \frac{30,8}{3,4} = 9,1$	< 0,001	$r_{xz} = 0,44$
Variation durch y	102,6	2 – 1 = 1	$\frac{102,6}{1} = 102,6$	$F_y = \frac{102,6}{3,4} = 30,2$	< 0,001	$r_{yz} = 0,74$
Interaktionsvariation	0	2 · 1 = 1	$\frac{0}{1} = 0$			
Fehlervariation	82	24	$\frac{82}{24} = 3,4$			
Gesamtvariation	246,2	30 – 1 = 29				

zur unerklärten Fehlervarianz gesetzt. Der Bruch liefert eine Maßzahl, die wiederum mit einem bestimmten Tabellenwert für die entsprechenden Freiheitsgrade verglichen wird. Ist die errechnete Maßzahl aus den empirischen Daten größer als der angegebene Tabellenwert für das vorgesehene Signifikanzniveau, so kann die Nullhypothese mit der im Signifikanzniveau angegebenen Irrtumswahrscheinlichkeit zurückgewiesen werden.

Der für die Variable x errechnete F-Wert beträgt 9,1, der für y 30,2. Wir suchen in einem statistischen Lehrbuch die Tabelle der F-Verteilung für ein Signifikanzniveau von 0,05 und 2, bzw. 24 Freiheitsgraden für den F-wert von x und einen, bzw. 24 Freiheitsgraden für den F-Wert von y. Wir finden in den entsprechenden Tabellen für das Signifikanzniveau von 5 % einen F-Wert von 3,40 für x und einen solchen von 4,26 für y. Da unsere empirisch errechneten Werte weitaus größer sind als der jeweils angegebene Tabellenwert, können wir die Nullhypothese, daß keine Beziehung zwischen den Variablen besteht, verwerfen. Offensichtlich sind sowohl die radiologische Therapie wie die operative Entfernung eines Krebsgeschwulstes brauchbare therapeutische Methoden.

Die eingangs formulierten Fragestellungen können nun beantwortet werden: Die Variation in den Meßwerten der abhängigen Variablen ist nicht zufällig, sondern systematisch entstanden. Als Ursache für diese Variation können die Variablen x und y gelten, wobei festgestellt wurde, daß eine Interaktionsvarianz nicht gegeben war. Die Varianzanalyse konnte weiter zeigen, daß offensichtlich y jener Faktor ist, der stärker auf die abhänigige Variable der Überlebensdauer wirkt.

Will man die Stärke der Beziehung zwischen der jeweiligen unabhängigen Variablen und der abhängiger Variablen noch exakter bestimmen, so stehen hierfür Korrelationskoeffizienten zur Verfügung. Die folgenden Formeln geben den sog. Interklassenkorrelationskoeffizienten an, der für den Beispielfall eine adäquate Maßzahl darstellt. Auch hier zeigt sich der stärkere Zusammenhang zwischen y und z.

$$r_{xz} = \frac{\bar{V}_x - \bar{V}_e}{\bar{V}_x + (n_x - 1)\bar{V}_e} = \frac{30{,}8 - 3{,}4}{30{,}8 + (10 - 1) \cdot 3{,}4} = 0{,}44$$

$$r_{yz} = \frac{\bar{V}_y - \bar{V}_e}{\bar{V}_y + (n_y - 1)\bar{V}_e} = \frac{102{,}6 - 3{,}4}{102{,}6 + (15 - 1) \cdot 3{,}4} = 0{,}74$$

Die Anwendung der Varianzanalyse auf unsere experimentell gefundenen Daten war bislang als richtig unterstellt worden. Tatsächlich jedoch müssen vor deren Anwendung die folgenden Bedingungen überprüft werden:

1. Die unabhängigen Variablen können ordinal mehrwertig bzw. nominal zweitig sein.
2. Die abhängige Variable muß mindestens auf dem Meßniveau einer Intervallskala liegen.

3. Die Grundgesamtheit, aus der die Untersuchungsgruppen als Stichproben gezogen werden, muß normal verteilt sein.

4. Die einzelnen Untersuchungsgruppen müssen als unabhängige Stichproben voneinander gelten können.

5. Die internen Gruppenvarianzen müssen innerhalb zufälliger Schwankungen (Zufälligkeit hier im statistischen Sinne) gleich sein.

Sind diese Voraussetzungen gegeben, so kann die Varianzanalyse legitimerweise angewandt werden; sie liefert unter diesen Vorausetzungen gültige und leicht interpretierbare Informationen.

2.2.4 Interpretation der Ergebnisse

Beispiel 1: Die Durchführung des t-Tests für korrelierende Stichproben als Signifikanztest hatte ergeben, daß die Nullhypothese, die keinerlei Beziehung zwischen Reduzierung des Lärmpegels und Arbeitsleistung unterstellt hatte, mit einer Irrtumswahrscheinlichkeit von 5% zurückgewiesen werden konnte. Wir sind deswegen in der Lage, begründet anzunehmen, daß die Einführung von Lärmschutzeinrichtungen die Arbeitsleistung steigern wird. Vorsicht ist jedoch geboten, wenn es darum geht, Aussagen darüber zu machen, ob diese Arbeitsleistungssteigerung auch längerfristig erhalten bleibt. So wäre durchaus vorstellbar, daß durch Gewöhnungseffekte die erzielte Leistungszunahme sich langfristig wieder reduziert und auf einem niedrigeren Stand stagniert. Daraus läßt sich nun methodologisch ableiten, daß die Zurückweisung der Null- und die Annahme der Alternativhypothese nur vorläufigen Charakter haben kann, daß es jederzeit möglich ist, scheinbar gültige Sachverhalte erneut zu falsifizieren. Eventuell könnte die oben formulierte Hypothese derart spezifiziert werden, daß nicht der geringere Lärmpegel an sich eine erhöhte Arbeitsleistung hervorbringt, sondern daß jedwede Lärmpegelveränderung für die modifizierte Arbeitsleistung verantwortlich zu machen ist, sodaß jede Lärmpegelstagnation entweder die abhängige Variable in ihrer Merkmalsausprägung nicht modifiziert oder aber sie sogar reduziert.

Neben der Frage der Signifikanz der Beziehungen von unabhängiger und abhängiger Variable kann auch die nach der Stärke der Beziehung gestellt werden. Solche Fragen beantworten die Korrelationskoeffizienten, je nach Meßniveau der kovarierenden Variablen müssen verschiedene Koeffizienten zu ihrer Berechnung herangezogen werden. (Vgl. hierzu die Beispiele zur Befragung!)

Beispiel 2: Die Anwendung des t-Tests auf die klassische Versuchsanordnung hatte ergeben, daß der Bestandteil Y der Droge X ein wirksamer Bestandteil ist, weil die Nullhypothese, die das Gegenteil behauptet hatte, zurückgewiesen werden konnte. Diese statistische Aussage kann deshalb als erklärungskräftig übernommen werden, weil die Versuchsanordnung so angelegt war, daß die potentiellen Fehlerquellen, wie zeitliche Einflüsse, Meßeinflüsse etc. durch die Kontrollgruppe kontrolliert wurden. Jede andere Versuchsanordnung, die in ihrer Konfiguration als „minderwertiger" anzusehen ist (z.B. vorexperimentelle Versuchsanordnung), hät-

ten statistisch gesehen zu einem ähnlichen Ergebnis führen können. Man hätte aber nie sicher darüber sein können, daß dieses statistische Ergebnis auch theoretisch richtig ist, weil bei den vorexperimentellen Konfigurationen die potentiellen Fehlerquellen nicht ausgeschlossen waren.

Auch im Beispiel 2 prüfte (wie grundsätzlich) der Signifikanztest nur die Frage, ob eine Beziehung zwischen den Variablen zuverlässig besteht oder nicht besteht, d.h. ob der Drogenbestandteil Y einen den Bluthochdruck reduzierende Wirkung hat. Wie stark der hochdruckreduzierende Effekt von Y ist, müßte wiederum mit einem Korrelationskoeffizienten berechnet werden.

Beispiel 3: In der faktoriellen Versuchsanordnung wurden zwei unabhängige Variablen daraufhin überprüft, ob und welchen Einfluß sie auf die abhängige Variable ausüben. Der durchgeführte Signifikanztest (F-Test) erbrachte für beide unabhängigen Variablen eine überzufällige Beziehung mit der abhängigen Variable. Der F-Test prüft dabei, inwieweit die jeweils unabhängige Variable dafür verantwortlich zu machen ist, daß eine Variation der Merkmalsausprägungen in der abhängigen Variablen auftrat. Anders als beim t-Test, der Unterschiede zwischen arithmetischen Mittelwerten daraufhin überprüfte, ob sie zufällig oder signifikant sind, bezieht sich der F-Test auf die Variation bzw. Varianzen der Variablen; er prüft die Verteilungen der Merkmalsausprägungen im Hinblick auf Unterschiede in den Varianzen.

Aus dem fiktiven Zahlenbeispiel ergab sich, daß die operative Behandlung eine größere Erfolgswahrscheinlichkeit – gemessen an der Überlebensdauer des Patienten – aufwies als die Bestrahlungstherapie. Aus der durchgeführten Varianzanalyse mit dem F-Test war unmittelbar ersichtlich, daß eine Interaktionswirkung beider Variablen nicht vorlag. Aus ihr war auch der Korrelationskoeffizient ableitbar, der die Stärke der Beziehung zwischen den jeweils unabhängigen und der abhängigen Variablen angab. Auch hier zeigte sich, daß die Bestrahlungstherapie einen geringeren Erfolg zu verbuchen hat.

Ergänzende und vertiefende Literatur:

BREDENKAMP, J., Experiment und Feldexperiment, in: Graumann, C.F. (Hrsg.), Handbuch der Psychologie, Bd. 7, 1. Halbband, Sozialpsychologie, Göttingen 1969

KÖNIG, R., Beobachtung und Experiment in der Sozialforschung, Köln 1972(8)

ZIMMERMANN, E., Das Experiment in den Sozialwissenschaften, Stuttgart 1972

3 TESTVERFAHREN

Testverfahren werden häufig nicht als eigenständige empirische Methode angesehen, weil sie einerseits in der Untersuchungsanlage mit dem Experiment vergleichbar bzw. aus dem Experiment ableitbar sind und andererseits auch andere empirische Methoden (z.B. Befragung) verwenden. Mit dem Fortschreiten der wissenschaftlichen Erkenntnis, das auch darin bestand, daß immer mehr und bessere Testverfahren entwickelt wurden, hat sich eine Verselbständigung der Tests als Methode ergeben, obgleich die verwandtschaftliche Nähe zu anderen empirischen Methoden nach wie vor gegeben ist. Die psychologischen Testverfahren haben weiten Einzug in die Spezialdisziplinen der klinischen Psychologie, Arbeitspsychologie, Schulpsychologie u.s.w. gefunden und nehmen dort in der Psychodiagnostik einen so hohen Stellenwert ein, daß es gerechtfertigt erscheint, die Testverfahren gleichwertig als eigenständige Methode dem Experiment gegenüberzustellen.

3.1 Methode und Logik des Tests

Basis eines jeden Experiments waren relationale Hypothesen, in denen eine theoretische Vermutung über die Beziehungsstruktur zwischen unabhängigen und abhängigen Variablen angestellt wurde. Im Experiment selbst wurde versucht, die Bedingungen der unabhängigen Variablen zu variieren oder zu manipulieren, um deren Einfluß auf die abhängige Variable feststellen zu können. In der Überprüfung der theoretischen Hypothese auf deren empirische Richtigkeit hin wollte man zu „kausalen" Beziehungen zwischen den Variablen derart kommen, daß man eine Regelmäßigkeit oder Gesetzmäßigkeit der Beziehung unterstellen konnte *(nomothetischer Aspekt).* Genau in diesem Merkmal unterscheiden sich Tests methodologisch vom Experiment. *Sie versuchen unter ähnlich definierten Versuchsbedingungen individuelle Merkmale zu erfassen, indem von einer hypostasierten Beziehungsstruktur ausgegangen und diese benutzt wird, um die individuellen Merkmale zu messen (ideographischer Aspekt).* Ein erstes Unterscheidungskriterium zwischen Experiment und Test liegt also darin, daß Experimente zum Zwecke der Hypothesenprüfung durchgeführt werden, um zu generalisierten Aussagen zu kommen, während Tests von ungeprüften oder geprüften Hypothesen ausgehen, um die individuellen Eigenschaften zu messen. Als illustratives Beispiel hierfür sollen die Intelligenztests dienen:

Intelligenz ist ein hypothetisches Konstrukt, das nicht unmittelbar erfahrbar und meßbar ist. Wenn man aber theoretisch begründet davon ausgehen kann, daß logisches Schließen mit ein wesentliches Definitionskriterium von Intelligenz ist, so kann das logische Schließen z.B. durch einen Analogietest geprüft und gemessen werden. Aus den Meßresultaten für den Analogietest schließt man auf die Intelligenzleistung. Die Richtigkeit eines solchen Schlusses setzt natürlich voraus, daß logisches Schließen ein Element der Intelligenz ist. Die Hypostasierung einer solchen Beziehung bzw. die Plausibilisierung oder theoretisch argumentative Absicherung ist also notwendige Voraussetzung für die Entwicklung und Durchführung von Tests.

Im Experiment wurden unabhängige Variablen variiert bzw. manipuliert, um im

Anschluß daran die abhängige Variable zu messen. Beim Test hingegen werden die Variablen weder manipuliert noch variiert, sondern es wird schlicht versucht, eine oder mehrere Variable zu messen, wobei zwar bestimmte Stimuli vorgegeben werden (z.B. Fragen oder Items), die aber keine Veränderung in der zu messenden Variablen hervorrufen sollen, sondern nur dazu dienen, die nicht unmittelbar beobachtbare Variable meßbar zu machen. *Das Fehlen einer planmäßigen Variation der unabhängigen Variablen wird somit zu einem wesentlichen Unterscheidungskriterium des Tests vom Experiment.*

Da zur Durchführung von Tests keine relationalen Hypothesen aufgestellt werden und erforderlich sind, unterscheiden sich Tests von Experimenten noch dadurch, daß sie rein deskriptive Aussagen über die gemessenen Eigenschaften an Individuen machen. Zwar können auch deskriptive Hypothesen mit Testverfahren überprüft werden, etwa in der Form: „Person X habe eine überdurchschnittliche Intelligenz"; doch ist der erkenntnistheoretische Wert einer solchen deskriptiven Hypothese als gering zu veranschlagen.

Während im Experiment durch die Variation der unabhängigen Variablen deren Einfluß auf die abhängige Variable gemessen wird (also nur eine einstufige Relation: unabhängige ⟶ abhängige Variable), wird beim Test die unabhängige Variable – nämlich das Erhebungsinstrument – für alle Probanden konstantgesetzt; es werden dann die abhängigen Variablen als Verhaltensrelationen auf dem Testinstrument gemessen, woran sich dann der Rückschluß von dem gemessenen overten Verhalten auf die zugrundeliegenden individuellen Persönlichkeitsmerkmale anschließt (also eine zweistufige Relation: unabhängige Variable ⟶ overtes Verhalten ⟶ Persönlichkeitsmerkmale).

Nach diesen ausgrenzenden Überlegungen sind wir nun in der Lage, die Logik der Tests zu entwickeln: zunächst können wir feststellen, daß die unabhängige Variable (der vorzugebende Stimulus oder die Stimuli) nicht variiert wird. Da wir über die Reaktionen auf den Stimulus oder die Stimuli auf bestimmte Persönlichkeitsdispositionen oder Eigenschaften schließen wollen, müssen wir wissen, welche Stimuli diese Qualität besitzen. Dies bedeutet: wir müssen mindestens eine theoretische Vermutung (wenn nicht gar Erkenntnis) darüber haben, daß zwischen Stimulus, Verhalten und Persönlichkeitseigenschaft eine Beziehung besteht. Die erste Aufgabe in der Testentwicklung wird also darin bestehen, solche Stimuli zu suchen, die als Operationalisierungen der Persönlichkeitseigenschaften dienen können. Da man aber nie sicher sein kann, daß ein einzelner Stimulus das theoretische Konzept in ausreichender Weise repräsentiert, wird das Bestreben des Forschers danach gehen, eine Folge von Stimuli zu entwickeln, die den theoretischen Begriff in seiner Vieldimensionalität adäquat widerspiegeln.

(So wie der Intelligenztest Rechenleistung , logisches Schließen etc. mißt und sich diese Dimensionen dann zu dem hypothetischen Konstrukt der Intelligenz zusammenfügen lassen, so müssen auch die einzelnen Elemente jeder Testaufgabe zum Zwecke einer gültigeren Operationalisierung in mehreren Items erfaßt werden; denn eine Rechenaufgabe allein wird sicher weniger geeignet sein, die Rechenleistung zu ermitteln als mehrere.)

Nachdem die Dimensionen eines theoretischen Konzepts aufgeschlüsselt und je-

weils Items (= Stimuli) zu deren Operationalisierung formuliert worden sind, kann der Test (hier verstanden als Erhebungsinstrument) konstruiert werden. Da keine systematische Variation vorgenommen wird und jeder Proband den gleichen Test zu bearbeiten hat, ist der Test damit in seiner Rohform entwickelt, soweit angegeben wird, in welcher Beziehung die einzelnen Items untereinander stehen und wie eine Gesamtbeurteilung der zu messenden Variablen zustandekommt (z.B. Addition der Punktwerte). Die Aussagekraft eines so kontruierten Tests ist jedoch noch relativ gering, da jeglicher Vergleichsmaßstab fehlt. Zwar kann davon ausgegangen werden, daß unterschiedliche Testergebnisse auf unterschiedliche Variablenkonstellationen zurückzuführen sind, insbesondere weil das Erhebungsinstrument und damit die unabhängigen Stimuli wie auch die Erhebungssituation allgemein konstantgehalten wurden, doch können festgestellte Differenzen nicht vernünftig interpretiert werden. Hierzu bedarf es einer Vergleichsgruppe oder vergleichender Maßzahlen, um die relative Bedeutung eines individuellen Testwertes abschätzen zu können (vgl. hierzu 3.1.3).

Selbst wenn eine solche Vergleichbarkeit hergestellt ist, kann noch nichts darüber ausgesagt werden, ob der Test auch ein gültiger Test für die zu messende Variable ist (vgl. die oben entwickelte zweistufige Relation). Nur unter der Voraussetzung, daß man konkrete und abgesicherte Informationen über diese zweistufige Relation hat, kann davon ausgegangen werden, daß das Testmaterial eine gültige Operationalisierung der theoretischen Variablen darstellt (vgl. hierzu 3.1.2).

Ist das Erhebungsinstrument erstellt, ist eine relative Vergleichbarkeit der über es zu erzielenden Meßwerte gegeben und kann davon ausgegangen werden, daß das Meßinstrument gültig ist, so können mit Hilfe des Tests konkrete Untersuchungen an Einzelpersonen (selbstverständlich auch in Gruppen) durchgeführt werden, sodaß für diese, im Hinblick auf die zu messende Variable (z.B. Intelligenz, Arbeitsleistung, Schulleistung etc.), interpretierbare Meßwerte vorliegen und praktisch verwertbare Aussagen gemacht werden können.

3.1.1 Definition und Arten von Tests

„Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad einer individuellen Merkmalsausprägung" (LIENERT, G., Testaufbau und Testanalyse, Weinheim 1969). Diese Definition ist mit den oben entwickelten Unterscheidungskriterien zwischen Test und Experiment kompatibel. Mit der Formulierung „wissenschaftliches Routineverfahren" ist gemeint: einmal entwickelte Tests können beliebig häufig angewandt werden, ohne daß das Erhebungsinstrument als solches verändert wird, während experimentelle Versuchsanordnungen im Regelfalle sich immer voneinander unterscheiden (auch wenn gleiche Hypothesen getestet werden sollen), weil in der scheinbaren Gleichheit nie alle Variablen gleichzeitig und gleich kontrolliert werden können; das wissenschaftliche Verfahren selbst ist im Test routinisiert, unterliegt also keinen Veränderungen.

Während im Experiment eine systematische Variation oder Manipulation der unabhängigen Variablen vorgenommen wurde, zeichnet sich der Test dadurch aus, daß die unabhängigen Variablen konstant sind, was sich auf den Test als *Erhebungsinstrument* selbst und auf die *Testsituation* bezieht. Man spricht davon, daß beide Elemente standardisiert sind. *Wesentliches Konstituens eines Tests ist somit seine Standardisierung im Hinblick auf die Gleichheit der situativen Faktoren in der Durchführung des Tests, wie auch im Hinblick auf die Gleichheit des Testmaterials.* Tests versuchen darüberhinaus Persönlichkeitsmerkmale in quantitativer Form zu erfassen, um bessere Differenzierungen zwischen den einzelnen Merkmalsausprägungen in Individuen lokalisiert vornehmen zu können, wobei diese Differenzierungen nur als relative, gemessen an einer, wie auch immer gearteten Kontrollgruppe, verstanden werden können.

Die psychologischen Tests können klassifiziert und Oberbegriffen zugeordnet werden. Je nach dem, welche Dimension der Klassifikation zugrundeliegt, ergeben sich unterschiedliche Zuordnungsmöglichkeiten gleicher Tests. Bezieht man sich in der Dimensionierung der Klassifikation auf die *Form der Erhebung,* so werden *Gruppen- und Individualtests* unterschieden. Klassifiziert man nach der *zeitlichen Begrenzung* in der Durchführung des Tests (wobei diese auf unterschiedliche und begründbare Absichten zurückzuführen ist), dann dichotomisiert man in *Geschwindigkeitstests (Speed-Tests)* und *Fähigkeitstests (Power-Tests).* Bezieht man seine Klassifikation darauf, was die einzelnen *Tests inhaltlich zu leisten vermögen,* so könnte man z.B. *Intelligenz-, Leistungs-, Schulaufnahme-, Einstellungs-, Konzentrationstests* usw. differenzieren (vgl. hierzu HILTMAN, H., Kompendium der psychodiagnostischen Tests, Bern, Stuttgart, Wien 1977). Solche Klassifikationen verschiedenster Art können ihre Relevanz nur dadurch legitimieren, daß sie für bestimmte Intentionen als zweckmäßig erscheinen. Da die Zweckmäßigkeit immer relativ auf das Ziel bezogen begründbar ist, finden sich in der Literatur eine Fülle von Klassifikationen, die in sich äußerst heterogen und z.T. sogar widersprüchlich sind.

Die dichotomisierende Klassifikation in *objektive und subjektive Testverfahren* kann theoretisch wie auch empirisch motiviert werden. Objektive Tests beabsichtigen, klar abgegrenzte, spezifisch formulierte und vom subjektiven Manipulationseinfluß freie Sachverhalte durch die Konstruktion des Erhebungsinstrumentes zu messen. Die Objektivität solcher Testverfahren besteht darin, daß durch die Konstruktionsweise des Tests, die mithin auch die Durchführung des Tests determiniert und die Interpretation der Resultate der persönlichen Willkür des Forschers entzieht, subjektive Einflüsse von außen ferngehalten werden. Es kommt hinzu, daß die Operationalisierung so vorgenommen wird, daß der Proband nicht in die Lage versetzt ist, durch Erwartungshaltungen, Kenntnis des Untersuchungsziels, Zusammensetzung des Tests etc. bewußt oder unbewußt manipulativ in die Datenerhebung als Verhaltensmodifikation so einzugreifen, daß eine Verfälschung der Testergebnisse auftritt. *Objektive Testverfahren zeichnen sich insgesamt dadurch aus,*

- *daß die Items des Tests standardisiert sind, also ein hoher Strukturierungsgrad des Erhebungsinstrumentes vorliegt,*
- *daß die Durchführung des Tests standardisiert ist,*
- *daß Auswertung und Interpretation nach vorgegebenen Regeln eindeutig vorgenommen*
- *und daß eine quantitativ-relationale Maßzahl gewonnen werden kann, auf der die Interpretation aufbaut und die nicht durch subjektive Elemente des Probanden verfälscht werden können.*

Subjektive Testverfahren unterscheiden sich von den objektiven nicht durch eine geringere Standardisierung der Reize. In beiden Fällen sind die Erhebungsinstrumente hochstandardisiert (können es mindestens sein). Subjektive Testverfahren unterscheiden sich jedoch von den objektiven dadurch, daß sie *unterschiedliche Inhalte* zu erfassen versuchen, daß in diesem Bemühen subjektive und damit auch manipulative Elemente einfließen können und daß die Auswertung und Interpretation der Untersuchungsergebnisse durchaus nicht forscherunabhängig sein müssen. Subjektive Testverfahren versuchen Einstellungen, Vorstellungen, Empfindlichkeiten, Gefühle, Motivationen etc. als *subjektive Äußerung des Probanden* zu erheben, so wie dieser sie wahrnimmt, empfindet oder glaubt, wahrzunehmen und zu empfinden. *Die Subjektivität besteht darin, daß sowohl bewußt wie auch unbewußt in die Datenerhebung Verzerrungen eingehen können, die bei Analyse und Interpretation der Daten nicht in jedem Falle ermittelt werden können.*

Die Zuordnung *projektiver Tests* zu objektiven oder subjektiven ist schwierig, weil sie sowohl Elemente der objektiven wie auch der subjektiven Testverfahren enthalten. Wenn hier trotzdem projektive Tests als Untergruppe der subjektiven Tests (mehr oder weniger eklektizistisch) verstanden werden, so insbesondere deshalb, weil die subjektiven Elemente in besonderer Weise bei projektiven Testverfahren Berücksichtigung finden. Sie stellen geradezu darauf ab, *das subjektive Erleben und Empfinden indirekt über standardisierte, aber relativ diffus und unstrukturiert erscheinende Reize zu erheben.* Die Logik der projektiven Tests besteht darin, daß relativ unstrukturiertes, mehrdeutiges Material als Stimulus vorgegeben wird, auf die der Proband – so die Vermutung – auf seine individuelle Weise reagieren wird, wobei seine Individualität Ausdruck einer ganzheitlichen Persönlichkeit ist. Mit anderen Worten: *projektive Tests gehen davon aus, daß in Wahrnehmung und Interpretation von unstrukturierten Stimuli sich die nicht unmittelbar beobachtbare Persönlichkeitsstruktur des Probanden artikuliert.* Definition und Interpretation des Stimulus werden als ursächlich durch die Persönlichkeit des Probanden determiniert begriffen.

Zwar ist in den Sozialwissenschaften längst bekannt, daß gleiche Situationen von verschiedenen Personen unterschiedlich definiert werden; nicht bekannt ist bislang allerdings, inwieweit solche Verschiedenartigkeit ausschließlich, überwiegend oder überhaupt auf Elemente der Persönlichkeitsstruktur oder auf soziale Bedingungen zurückzuführen ist. Unterschiedliche Definition und Interpretation von unstrukturierten Reizen können daher nur bedingt als Projektion der Persönlichkeits-

struktur auf den Reiz angesehen werden. Neben dieser allzu häufig vergessenen Problematik spielt die Subjektivität der Sinndeutungen als verzerrendes Element nur eine untergeordnete Rolle: die Probanden sind normalerweise nicht in der Lage, ein konsistent falsches Antwortverhalten als Reaktion auf die vorgegebenen Reize zu entwickeln, weil die projektiven Testverfahren Spontaneität erfordern, die die bewußte Manipulationsmöglichkeit reduziert. Allerdings sind subjektive Einflüsse bei der Interpretation des Reaktionsverhaltens der Probanden nicht nur nicht auszuschließen, sondern häufig anzutreffen: Obwohl gleiches Testmaterial und gleiche Reaktionen des Probanden zugrundeliegen, werden diese häufig von verschiedenen Forschern unterschiedlich interpretiert, d.h. die Standardisierung der Auswertung und Interpretation projektiver Testverfahren läßt zu wünschen übrig. Diese mangelnde Standardisierung ist vor allem darauf zurückzuführen, daß die oben angesprochene theoretische Absicherung der Projektion als Ausdruck der Persönlichkeitsstruktur noch nicht voll gelungen ist (was insbesondere mit daran liegt, daß unterschiedliche Persönlichkeitstheorien in die Interpretation einfließen können).

Faßt man die Wesenselemente der projektiven Tests zusammen, so erhält man eine Argumentation dafür, weshalb sie eher den subjektiven Techniken zuzuordnen sind:

1. *Objektive Testverfahren gingen analytisch vor,* indem einzelne Dimensionen oder Elemente erhoben wurden, die bei der Interpretation zu einem Gesamtbild zusammengesetzt wurden. *Die Gesamtaussage war daher eine synthetische. Projektive Verfahren* legen jedoch eine *ganzheitliche Persönlichkeit* zugrunde, die das Verhalten auf den unstrukturierten Reiz in entsprechender Weise determiniert. Bei der Interpretation wird daher davon ausgegangen, ein ganzheitliches Bild der Persönlichkeit zu entwerfen, nicht einzelne Elemente herauszugreifen (z.B. Intelligenz).

2. Zwar sind bei objektiven und projektiven Techniken die verwendeten Reize standardisiert, doch sind bei objektiven Tests die Reize selbst *sehr spezifisch, präzise und strukturiert,* während sie bei der *projektiven Technik unstrukturiert und diffus* belassen sind.

3. Während *objektive Tests eher kognitive Sachverhalte* erfassen, sind *projektive Tests* eher dazu geeignet, *emotionale Persönlichkeitsaspekte* zu ermitteln.

4. *Auswertung und Interpretation der Testergebnisse sind bei objektiven Verfahren hochstandardisiert,* während sie bei den projektiven Techniken sowohl von den angewandten Persönlichkeitstheorien (z.B. Psychoanalyse etc.) wie auch von dem die Ergebnisse interpretierenden Forscher abhängig sind.

Die projektiven Tests können im Hinblick auf das Verfahren der Stimulidarbietung und der sich daraus ergebenden Reaktionsform klassifiziert werden, wobei fünf Möglichkeiten – gemessen an ihrem quantitativen Einsatz – von Bedeutung sind:

1. Interpretation:

Die von dem Testleiter vorgegebenen Stimuli müssen interpretiert werden. Der wohl berühmteste Test dieser Gruppe ist der TAT (Thematic Apperception Test), bei dem der Proband aufgefordert wird, zu vorgegebenen Bildern Geschichten zu erfinden. Gelegentlich arbeitet man auch mit Zeichnungen und Sprechblasen, wobei die offengelassene Sprechblase vom Probanden ausgefüllt werden soll.

2. Konstitution:

Der Proband wird aufgefordert, relativ unstrukturiertem, diffusem Material eine Struktur, einen Sinn zu geben. Der bekannteste Test aus dieser Reihe ist der Rorschachtest, bei dem ein-

und mehrfarbige Kleckse als Stimuli vorgegeben werden, die durch die verbalisierten Deutungen und Interpretationen des Probanden eine bestimmte Konstitution bekommen.

3. Ergänzung:

Bei den Ergänzungstechniken werden verbale Reize vorgegeben, die verbal ergänzt werden müssen. Beim Wort-Assoziationstest muß auf ein vorgegebenes Wort spontan ein weiteres genannt werden. Aus den zugrundeliegenden Assoziationen schließt man auf Persönlichkeitsmerkmale. Ähnlich beim Satzergänzungstest oder Bildergänzungstest, bei denen unvollständige Reize vorgegeben werden, die durch die Versuchsperson ergänzt werden sollen.

4. Konstruktion:

Aus bestimmten Detailelementen, die in unterschiedlicher Weise zusammengesetzt werden können, soll die Versuchsperson eine ganzheitliche Struktur entwickeln. Der Mosaiktest ist der wohl bekannteste Test dieser Art, bei dem eine Anzahl von in Form und Farbe unterscheidbaren Pappstücken zusammengesetzt werden soll.

5. Refraktion:

Zu den refraktiven Tests würde die Graphologie gehören, die unter den genannten Methoden wohl die umstrittenste ist.

Leistungstests sind spezielle Testverfahren, die den objektiven Tests zuzuordnen sind und sich je nach Konstruktion auf unterschiedliche Inhalte beziehen. Bekannte Leistungstests sind die unterschiedlichen Intelligenztests; aber auch Schultests, Berufseignungstests etc. würden dazugerechnet werden können. Als unter die objektiven Testverfahren subsumierbar gelten für die Leistungstests die oben entwickelten Kriterien für die objektiven Tests.

Einen breiten Raum in der Psychodiagnostik nehmen die *Persönlichkeitstests* ein. Sie sind subjektive Testverfahren, die auf Fragebogen als Erhebungsinstrument aufbauen. Die Subjektivität des Tests besteht dabei darin, daß durch die Konstruktion des Erhebungsinstrumentes es im Regelfalle (Ausnahme indirekte oder projektive Fragen, vgl. hierzu Kap. 4) dem Probanden möglich ist, die Intentionen des Fragebogens als Versuch der Ermittlung bestimmter Sachverhalte zu durchschauen und er damit in die Lage versetzt wird, seine Antworten manipulativ zu verändern. Hier bezeichnet also Subjektivität Verzerrungseinflüsse. *Subjektivität bei Persönlichkeitsfragebogen kann aber auch dahingehend verstanden werden, daß im Gegensatz zu kognitiven Sachverhalten emotionale erhoben werden, die ja eindeutig subjektspezifisch erlebt werden.* Einstellungen, Meinungen, Vorlieben, Vorurteile, Interessen etc. sind eben subjektbezogene und persönlichkeitsabhängige Faktoren. In dieser Beziehung haben Persönlichkeitsfragebogen eine Gemeinsamkeit mit den projektiven Tests. Sie unterscheiden sich jedoch von letzteren, insbesondere dadurch, daß Auswertung und Interpretation der Persönlichkeitsfragebogen durchaus intersubjektiv, zuverlässig und eindeutig vorgenommen werden können, sofern das Erhebungsinstrument ausreichend standardisiert war. Insbesondere durch maschinelle Auswertung wird (wenn man von zufälligen Codierungsfehlern etc. einmal absieht) eine systematische und persönlichkeitsspezifische Verzerrung in Auswertung und Interpretation vermieden.

Leistungstests als objektive Testverfahren können für sich in Anspruch nehmen, eine durchgängige Standardisierung von Erhebungsinstrument über Erhebungssituation, bis hin zur Auswertung und Interpretation mit der Konsequenz einer optimalen Fehlerausschaltung zu schaffen. Persönlichkeitsfragebogen als subjektive Test-

verfahren können verfälschende und verzerrende subjektive Einflüsse der Probanden in der Beantwortung der vorgegebenen Fragen nicht ausschließen. Testsituation, Testverfahren, Auswertung und Interpretation können jedoch standardisiert werden. Projektive Testverfahren als subjektive Tests weisen eine hohe Standardisierung der Stimuli auf (nicht zu verwechseln mit der Tatsache, daß die Stimuli an sich unstrukturiert und diffus sind), *die durchaus in standardisierten Situationen dargeboten werden können.* Die subjektive Manipulierbarkeit der Reaktionen auf diese Stimuli durch die Probanden ist äußerst begrenzt, *hingegen ergeben sich bei Auswertung und Interpretation persönlichkeitsspezifische, subjektive Modifikationen durch den Diagnostiker.*

3.1.2 Gütekriterien der Tests

Gemäß Definition war vereinbart worden, daß von Tests dann gesprochen werden kann, wenn eine hohe Standardisierung der Stimuli und der Erhebungssituation erreicht wird. Eine solch hohe Standardisierung schafft zwar einige Vorbedingungen dafür, daß Anwendung, Durchführung und Interpretation von Tests keinen allzu großen Verzerrungen unterliegen; sie allein genügt jedoch nicht, um Testverfahren als wissenschaftlich brauchbar auszuweisen. Die in den empirischen Wissenschaften für die wissenschaftliche Brauchbarkeit von Datenerhebungsverfahren entwickelten Kriterien müssen für jedes Erhebungsinstrument gesondert auf deren Vorliegen überprüft werden. Die drei in der Literatur zur Testtheorie genannten Kriterien, *Objektivität, Zuverlässigkeit (Reliabilität) und Gültigkeit (Validität)* müssen für jedes Testverfahren *gleichzeitig* gegeben sein, damit von wissenschaftlich brauchbaren und verwertbaren Ergebnissen ausgegangen werden kann.

Objektivität ist die Basiskategorie jeglicher wissenschaftlicher Forschung. *Von ihr wird dann gesprochen, wenn eine inter-individuelle Zuverlässigkeit bzw. Nachprüfbarkeit derart gegeben ist, daß unter sonst ceteris-paribus-Bedingungen verschiedene Forscher zu demselben empirisch gewonnenen Resultat gelangen.* Der Begriff der Objektivität wird in der Literatur mehr und mehr durch den der interindividuellen Zuverlässigkeit ersetzt, weil Objektivität etwas vorzugeben scheint, was insbesondere in den Sozialwissenschaften nicht zu leisten ist. Objektivität wird allzu leicht assoziiert mit Wahrheit, reiner Erkenntnis etc., was angestrebt, jedoch realiter in konkreter empirischer Forschung nicht erreicht werden kann. Aus den wissenschaftstheoretischen und methodologischen Überlegungen wissen wir, daß auch unsere systematische, wissenschaftliche Erfahrung durch das hypothetisch-theoretische Raster, das sich in der Versuchsanordnung manifestiert, gesehen wird, sodaß von daher die Erkenntnisse immer nur im Lichte der Hypothesen oder Theorien gewonnen und interpretiert werden können. Zudem sind solche Ergebnisse stets ausschnitt- und modellhaft, weil nicht alle eventuell relevant erscheinende Faktoren in das Untersuchungsdesign miteinbezogen werden können. Unter ideologiekritischem Aspekt muß Objektivität eher als normativer Anspruch denn als Realität konkreter Forschung perzipiert werden.

Diese relativierte Stellungnahme kann nicht so interpretiert werden, als ob Objek-

tivität zugunsten von Subjektivität aufgegeben werden müßte. Vielmehr muß bei jeder empirischen Erhebung und Untersuchung darauf geachtet werden, daß interindividuelle Zuverlässigkeit und Nachprüfbarkeit im Sinne der oben entwickelten Definition von Objektivität gewährleistet ist. Nach LIENERT kann die Objektivität als interpersonelle Übereinstimmung auf drei Ebenen gefährdet sein:

1. *Die Durchführungsobjektivität* meint die Unabhängigkeit der Untersuchungsergebnisse von bewußten oder unbewußten Verhaltensweisen des Durchführenden im Verlaufe der Untersuchung. Als Regel kann man aufstellen, daß eine möglichst hohe Standardisierung des Erhebungsinstrumentes und der Erhebungssituation dazu führt, daß Durchführungsobjektivität gewährleistet ist.

2. Die *Auswertungsobjektivität* betrifft jenen Untersuchungsschritt, der sich im Anschluß an die Durchführung der Untersuchung (Datenerhebung) anschließt. Für standardisierte Erhebungsinstrumente ist normalerweise eine hohe Auswertungsobjektivität gegeben, weil die Erhebungsinstrumente standardisiert sind und z.B. für Tests Auswertungsschablonen, Musterlösungen etc. mitgeliefert werden und weil für Fragebogen mit hoher Standardisierung falsche Auswertungen durch hohe statistische Kontrolle praktisch angeschlossen sind. Auswertungsobjektivität ist also immer dann gegeben, wenn verschiedene Auswerter bei gleichen Tests und gleichen Probanden zu den gleichen Auswertungsergebnissen gelangen.

3. Die *Interpretationsobjektivität* geht davon aus, daß bei gleichen Untersuchungsresultaten verschiedene Diagnostiker, Forscher oder Untersuchungsleiter zu denselben interpretatorischen Schlußfolgerungen gelangen. Bei Leistungstests z.B. (vgl. 3.1.3) sind die Interpretationen der Untersuchungsergebnisse genormt, sodaß ein Variationsspielraum der Interpretation nicht auftreten kann. Anders jedoch bei den projektiven Testverfahren, bei denen durchaus divergierende Deutungen anzutreffen sind, die die Interpretationsobjektivität als gefährdet erscheinen lassen.

Der Zusammenhang von Durchführungs-, Auswertungs- und Interpretationsobjektivität stellt sich so dar, daß ohne Durchführungsobjektivität Auswertung und Interpretation verzerrt und problematisch sind, daß ohne Durchführungsobjektivität und Auswertungsobjektivität die Richtigkeit der Interpretation gefährdet ist. Daraus ergibt sich, daß auf allen drei Ebenen Objektivität gewährleistet sein muß, um zu unverzerrten Resultaten zu gelangen.

Das zweite Kriterium von Tests ist die *Reliabilität. Sie meint das Ausmaß der Streuung des Meßinstrumentes bei dessen wiederholter Anwendung auf den gleichen Sachverhalt.* Geht man davon aus, daß sich das zu messende Untersuchungsobjekt nicht verändert und daß Einflüsse durch den Versuchsleiter auszuschalten sind, so sollte bei wiederholter Anwendung des Meßinstrumentes zu erwarten sein, daß keine Differenzen in den Meßergebnsisen auftreten. Weil von der Konstanz des Objektbereiches und der Objektivität des Forschers ausgegangen wird, man sich also ausschließlich auf das Meßinstrument bezieht, spricht man auch von *intra-individueller Zuverlässigkeit.*

Selbst unter den gemachten Voraussetzungen wird sich bei wiederholter Anwendung eines Meßinstrumentes nicht in jedem Falle ein gleiches Meßergebnis einstellen. Die Meßergebnisse werden innerhalb statistischer Schwankungsbreiten variieren. Der Versuch, eine solche Variation auszuschließen oder im Ausmaß zu reduzieren, wird insbesondere bei psychologischen und manchen medizinischen Meßinstrumenten nicht immer möglich sein. (Man vergleiche nur die Blutdruckmessung, bei der die Meßwerte relativ ungenau abgelesen werden und zudem von der Placierung der Manschette, Lautstärkeeinflüssen der Umgebung usw. abhängig

sein können.) Zwar ist anzustreben, solche Meßwertschwankungen auszuschließen, doch ein bestimmtes Maß wird grundsätzlich toleriert werden müssen. Soweit sich solche Meßwertvariationen zufällig und innerhalb statistischer Schwankungsbreiten bewegen, also systematisch verzerrende Einflüsse ausgeschlossen sind, wird man durchaus davon ausgehen können, daß zuverlässige Meßergebnisse gewonnen wurden. Allerdings sind sie nicht als absolut geltende Maßzahlen anzusehen, sondern als empirisch gewonnene Daten, die mit einem relativ kleinen zufälligen Fehler behaftet sind, im wesentlichen jedoch das Untersuchungsmerkmal adäquat in seiner Ausprägung wiedergeben. *Der zufällig streuende Meßfehler muß in Kauf genommen werden, während der systematisch verzerrende Fehler, beeinflußt durch das Meßinstrument, ausgeschlossen werden muß.*

Um die Zuverlässigkeit eines Instrumentes nicht nur qualitativ durch den einzelnen Forscher beurteilen zu lassen, sondern auch ein quantitatives und objektiveres Maß für die Reliabilität zu haben, sind verschiedene Verfahren entwickelt worden. Vier seien kurz vorgestellt:

1. *Das Test-Retest-Verfahren:*

Kann man davon ausgehen, daß ein Test innerhalb eines bestimmten Zeitraumes wiederholt werden kann, ohne daß sich das zu messende Merkmal verändert hat und ohne daß Einflüsse des Meßinstrumentes auf das zu messende Merkmal möglich sind, so bietet das Vergleich beider Testergebnisse ein Maß für die Zuverlässigkeit des Tests. Die Berechnung eines Korrelationskoeffizienten zwischen erstem und zweitem Test wäre eine Maßzahl zur Bestimmung der Reliabilität. In diesem Falle wird der Korrelationskoeffizient als *Stabilitätskoeffizient* bezeichnet, weil er die Stabilität des Meßinstrumentes über einen bestimmten zeitlichen Abstand hinweg angibt.

Die für die Durchführung der Testwiederholungsmethode zu machenden Annahmen und Voraussetzungen sind im Regelfalle psychologischer Testverfahren nur sehr bedingt gegeben. Wählt man die Zeitspanne zwischen den beiden Tests zu groß, so kann sich das zu untersuchende Merkmal verändert haben. Der Stabilitätskoeffizient erreicht dann einen zu niedrigen Wert, der eventuell als ungenügende Reliabilität des Testverfahrens interpretiert wird, was eine ungültige Aussage wäre. Wählt man die Zeitspanne zu kurz, so besteht die Gefahr, daß Erinnerungseinflüsse Platz greifen und einen zu hohen Stabilitätskoeffizienten provozieren, also eine Zuverlässigkeit des Meßinstruments suggerieren, die realiter nicht gegeben ist.

2. *Die Split-half-Methode (Testhalbierung)*

Hierbei wird ein Test in statistisch zufällig gewonnene Hälften geteilt und einer Stichprobe von Probanden vorgelegt. Durch die statistische Zufälligkeit soll erreicht werden, daß in beiden Testhälften Aufgaben zu den gleichen Dimensionen und mit gleichem Schwierigkeitsgrad enthalten sind. (Besteht z.B. ein Intelligenztest aus einem Wortschatztest mit 10 und einem Rechentest mit 8 Items, so müßte bei dessen Aufsplittung in zwei Testhälften jede Testhälfte 5 Aufgaben zum Wortschatz und 4 Rechenaufgaben jeweils ähnlicher Schwierigkeit enthalten.) Diese beiden Testhälften werden den Probanden vorgelegt, sodaß für jeden Probanden zwei Testergebnisse vorliegen. Die Korrelation beider Testergebnisse liefert den *Koeffizienten der internen Konsistenz* und gibt den Grad der Zuverlässigkeit an. Ein hoher Korrelationskoeffizient entspricht also einem hohen Maß an Reliabilität der beiden Tests.

3. *Die Methode der äquivalenten Formen (Paralleltest)*

Entwickelt man zu einem Objektbereich nicht – wie normalerweise üblich – nur ein Testverfahren, sondern versucht man die zu messenden Variablen durch zwei unabhängig voneinander konstruierte Erhebungsinstrumente zu operationalisieren, so bieten beide Testverfahren in gegenseitiger Kontrolle die Möglichkeit, die Testergebnisse miteinander zu vergleichen und von daher auf Zuverlässigkeit der Meßresultate zu schließen. Der Korrelationskoeffizient zwischen den Meßwerten der beiden Testformen gibt das Ausmaß der Meßwertübereinstimmung an, mißt also die Zuverlässigkeit der beiden Tests und wird als *Äquivalenzkoeffizient* bezeichnet.

4. *Die Konsistenzmethode*

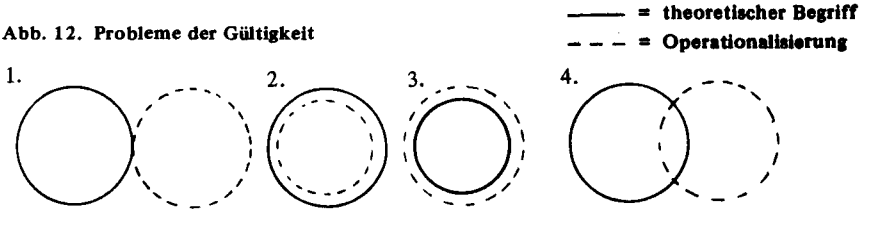
Sie ist die extremtypische Weiterführung der Split-half-Methode. Dabei wird der Test in eben-

so viele Elemente unterteilt wie Items vorhanden sind. Durch Korrelation dieser untereinander wird ein Reliabilitätskoeffizient berechnet, wobei offenkundig ist, daß die in die Berechnung eingehenden Elemente des Tests jeweils auf einer Dimension liegen, also homogen sein müssen, weil in allen anderen Fällen ein möglicherweise verzerrt niedrigerer Koeffizient berechnet wird, dessen Interpretation wegen der falschen Voraussetzungen für seine Berechnung ungültig wird.

Die Frage danach, welche Größe ein solcher Reliabilitätskoeffizient annehmen soll, damit von zuverlässigen Meßinstrumenten ausgegangen werden kann, wird in der Literatur nicht einheitlich behandelt. Da es keine absoluten Maßstäbe für die Beurteilung dieser Frage gibt, handelt es sich bei der Angabe von Grenzwerten eigentlich immer nur um Konventionen. Man kann davon ausgehen, daß ein Reliabilitätskoeffizient von 0,9 etwa ein durchaus ausreichendes Indiz für Zuverlässigkeit darstellt, auch wenn „Puristen" gelegentlich einen solchen von 0,95 verlangen. Einigermaßen gesichert kann man auch sagen, daß Reabilitätskoeffizienten, die kleiner sind als 0,8 oder 0,75 den Anforderungen nicht mehr genügen können.

Das wichtigste Gütekriterium von Tests ist die Validität (Gültigkeit). *Unter Validität versteht man die Übereinstimmung des durch das Meßverfahren erfaßten Objektbereiches mit dem theoretisch gemeinten Objektbereich, also die Frage, inwieweit ein Meßverfahren das mißt, was es vom Forscher beabsichtigt messen soll.* Wenn z.B. Intelligenz als hypothetisches Konstrukt, das nicht unmittelbar beobachtbar, meßbar und erfahrbar ist, gleichwohl empirisch erhoben werden soll, so bedarf es der Operationalisierung dieses theoretischen Konstruktes. Eine solche Operationalisierung hat die Transformation theoretischer Begriffe in empirische (beobachtbare) Sachverhalte zur Voraussetzung. Daß bei einer solchen Transformation Unschärfen unterlaufen können, dürfte unstrittig sein. Ebenso unstrittig ist aber, daß nur dann, wenn die operationale Definition genau das erfaßt, was mit dem entsprechenden theoretischen Begriff gemeint ist, wenn also völlige Deckungsgleichheit zwischen beiden besteht, ein entsprechendes Meßverfahren gültig ist (vgl. hierzu 1.5, in dem dieser Sachverhalt als eine strukturtreue Abbildung eines empirischen Relativs in ein numerisches Relativ angesprochen wurde). Bei der Testkonstruktion ist daher darauf zu achten, daß der Test tatsächlich jene Merkmale mißt, die man zu messen beabsichtigt. Daß in der empirischen Realität eine vollkommene Deckungsgleichheit zwischen Meßabsicht und tatsächlicher Messung eintritt, dürfte ein seltener Idealfall sein. Tatsächlich werden von Fall zu Fall Inkongruenzen auftreten, die die Gültigkeit mehr oder weniger tangieren. Das Problem der mangelnden Gültigkeit als nicht erzielte Deckungsgleichheit von intendierter und tatsächlicher Messung kann in Form von Venn-Diagrammen illustriert werden:

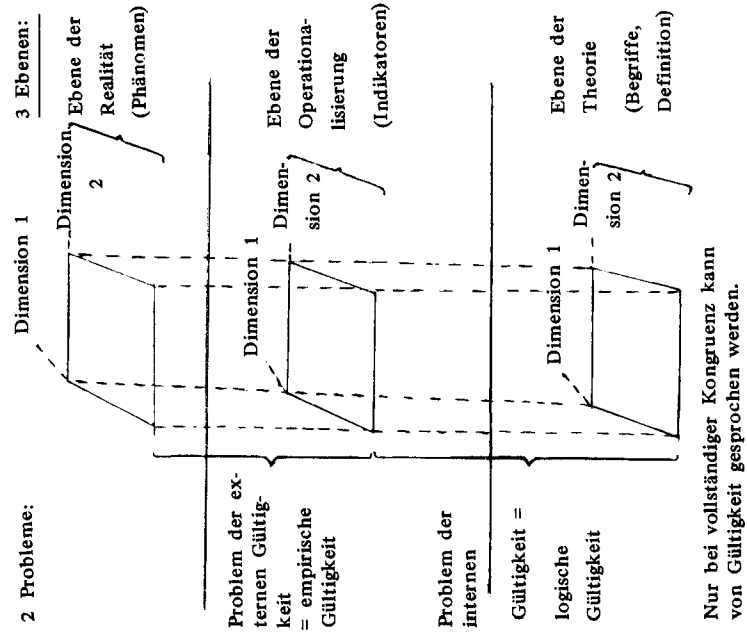
Abb. 12. Probleme der Gültigkeit



Inwieweit solche Abweichungen von dem Idealzustand beim Testverfahren auftreten, kann immer nur im konkreten Einzelfall und dort nur annäherungsweise angegeben werden. Daraus läßt sich folgern, daß Validität nicht als eine Kategorie des „Alles oder Nichts" zu verstehen ist, sondern daß es (wie es z.B. auch die Reliabilitätskoeffizienten bei den Zuverlässigkeitsprüfungen ausgedrückt haben) graduelle Abstufungen in der Beurteilung der Gültigkeit gibt. Anders als bei der Zuverlässigkeitsprüfung, wo exakte Maßzahlen zur Verfügung standen, kann der Grad der Genauigkeit, mit dem ein Meßinstrument das mißt, was es messen soll, im Regelfall nicht exakt angegeben werden, obgleich man auch hierfür Koeffizienten berechnen könnte. Diese restriktive Aussage ist damit zu begründen, daß in den Fällen, wo solche Koeffizienten unschwer (vom Technischen her gesehen) berechnet werden können, vergleichende Meßergebnisse vorliegen müssen. Für diese gelten aber die gleichen Gültigkeitsprobleme, sodaß letztlich keine stichhaltigen und in der Weise abgesicherten Maßzahlen gewonnen werden können, wie diese suggerieren wollen.

Das Problem der Gültigkeit kann analytisch in zweifacher Weise aufgefächert werden: Vergleicht man, inwieweit *theoretische Begriffe gültig operationalisiert* (d.h. in empirische Begriffe, beobachtbare Indikatoren übergeführt) worden sind, so spricht man von der *internen oder logischen Gültigkeit.* Die Gültigkeit kann aber auch gefährdet sein durch nicht gelungene Transformation in der eigentlichen Meßoperation, d.h. durch eine *mangelnde Kongruenz zwischen Operationalisierungsebene und Phänomenen.* In diesem Fall handelt es sich um die Probleme der *externen oder empirischen Gültigkeit* (vgl. hierzu Abb. 13).

Abb. 13. Logische und empirische Gültigkeit



Da in all den Fällen, wo die Gültigkeit eines Testverfahrens nicht gewährleistet ist (man z.B. glaubt, logisches Denkvermögen zu erfassen, tatsächlich jedoch Verbalisierungsvermögen mißt), sind die gewonnenen Meßwerte nicht bzw. nur falsch interpretierbar, deren Erkenntniswert also gleich Null. Daher hat man der Gültigkeitsüberprüfung theoretisch und praktisch viele Überlegungen gewidmet, die in unterschiedliche Verfahren der Gültigkeitsprüfung einmündeten:

1. Expert-Validity

Die Expertenvalidität muß als eines der am wenigsten geeigneten Verfahren der Gültigkeitsüberprüfung gelten. Man legt einen Test einer Reihe von Experten vor, und diese überprüfen die Gültigkeit des Meßverfahrens aufgrund ihrer Informationen, ihrer wissenschaftlichen Erkenntnisse, ihrer Erfahrungen, aber natürlich auch aufgrund von Plausibilität und bestimmen dann mehr oder weniger dezisionistisch den Grad der Gültigkeit. Da bei diesem Verfahren subjektive Einflüsse nicht auszuschließen sind (z.B. durch eine nicht ausreichend große Zahl von befragten Experten), und im Regelfalle auch nicht kontrolliert werden können, muß diese Gültigkeitsprüfung als relativ unzuverlässig gelten.

2. Known-groups-Validity

Ein entwickelter Test wird auf eine Gruppe angewandt, von der man zu wissen glaubt, in welcher Verteilung die Merkmalsausprägungen des zu messenden Merkmals vorliegen. Stellt man eine Übereinstimmung der Meßergebnisse mit den als bekannt vorausgesetzten Informationen fest, so geht man davon aus, daß das Meßverfahren gültig ist. Auch hier muß jedoch kritisch eingewandt werden, daß die Informationen über die Merkmalsverteilung auf irgendeine empirische Art gewonnen werden mußten, die aber selbst mit Gültigkeitsproblemen behaftet sein kann.

3. Predictive-Validity

Bei der Vorhersage-Validität wird aufgrund des eingesetzten Meßverfahrens eine Prognose erstellt, die sich auf das gemessene Merkmal bezieht. Tritt diese Prognose dann in der Realität ein, geht man davon aus, daß das Meßverfahren valide war. (Prognostiziert man z.B. mittels eines Schuleignungstests oder Schulreifetests, daß das erste Schuljahr der Grundschule erfolgreich absolviert wird, so kann man zum Abschluß des Schuljahres mit dem festgestellten Schulerfolg (Versetzung oder Nichtversetzung), also mit Hilfe dieses Außenkriteriums, die Validierung des Tests vornehmen.)

4. Construct-Validity

Hypothetische Konstrukte (z.B. Intelligenz) sind in sich vielschichtige Phänomene und stehen mit anderen Phänomenen in vielfachem theoretischem und empirischem Zusammenhang. Solche Beziehungen zwischen Phänomenen auf theoretischer Ebene können hypothetisch erfaßt und mittels empirischer Relationen überprüft werden. Entsprechen nun die empirischen Untersuchungen in deren Resultaten den theoretischen Überlegungen, so kann man im Vergleich von hypothetischem Konstrukt und den empirischen Tatsachen auf Gültigkeit schließen. Eine solche Schlußfolgerung ist aus logischen Gründen nur dann möglich, wenn eine Übereinstimmung festgestellt wird. Aber selbst in diesem Fall ist sie kein endgültiger und gültiger Beweis für die Validität. Da die relationalen Zusammenhänge innerhalb und mit den theoretischen Konstrukten prinzipiell immer erweiterbar sind, ist eine Konstruktvalidierung nie endgültig und abgeschlossen; sie kann laufend durch weitere empirische Analysen ergänzt werden. (Gerade die Konstruktvalidierung ist ein Fall dafür, wo kein Koeffizient für das Ausmaß der Gültigkeit berechnet werden kann.)

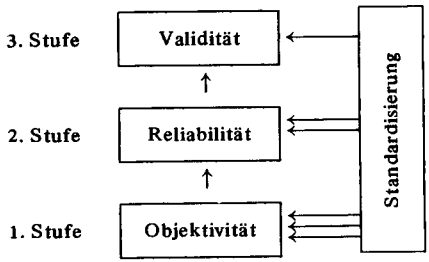
5. Criterion-Validity

Die Kriteriumsvalidität ergibt sich als Gültigkeit der Übereinstimmung zwischen den gemessenen Testresultaten und über andere Meßverfahren gewonnenen vergleichbaren Ergebnissen. (So könnte man z.B. den Hamburg-Wechsler-Intelligenztest für Erwachsene mit dem Stanford-Binet-Intelligenztest und vice versa validieren.) Die Übereinstimmungsvalidität kann als Oberbegriff für die Vorhersagegültigkeit und die Validitätsprüfung für bekannte Gruppen angesehen werden. Bei der Vorhersagevalidität wird eine Übereinstimmung der Testergebnisse mit einem später erfolgenden Außenkriterium verglichen, während bei der Known-Groups-Validity Übereinstimmung zwischen Meßergebnis des einen Tests mit vermuteten Merkmalsausprägungen oder solche, die über einen anderen Test erhoben wurden, geprüft wird.

Die in diesem Abschnitt entwickelten Gütekriterien, Objektivität, Reliabilität und Validität sind, obgleich sie analytisch getrennt vorgestellt wurden, nicht unabhängig voneinander zu sehen. War eine hohe Standardisierung der Erhebungssituation wie auch des Testverfahrens (Erhebungsinstrument) eine wichtige Vorbedingung dafür, daß die Objektivität auf den Ebenen der Durchführung, Auswertung und Interpretation gewährleistet war, so stellt *die Objektivität selbst eine notwendige, aber allein nicht hinreichende Voraussetzung für Zuverlässigkeit dar.* Objektive Untersuchungsdurchführung, Auswertung und Interpretation können bei Ungenauigkeit des Meßinstrumentes diese nicht kompensieren. Andererseits nützt das genaueste Meßinstrument nichts, wenn die Objektivität nicht gewährleistet werden kann.

Objektivität und Zuverlässigkeit sind notwendige, aber nicht hinreichende Bedingungen für die Gültigkeit eines Meßinstrumentes. Sollen Meßergebnisse valide sein, so müssen sie notwendig objektiv und zuverlässig gewonnen worden sein. Allerdings genügen Objektivität und Reliabilität alleine nicht, um Gültigkeit zu gewährleisten. Validität zeichnet sich zusätzlich dadurch aus, daß tatsächlich das gemessen wird, was gemessen werden soll, wofür Objektivität und Zuverlässigkeit keine Gewähr bieten. Die Standardisierung von Test, Testsituation, Auswertung und Interpretation der Testergebnisse, ist ein parallel laufendes Kriterium, das auf den Ebenen Objektivität, Reliabilität und Validität optimale Sorge dafür trägt, daß brauchbare Ergebnisse erzielt werden. Dies darf jedoch nicht so verstanden werden, daß hohe Standardisierung automatisch zu Gültigkeit, Zuverlässigkeit und Objektivität führt. Es kann der Einfluß der Standardisierung derart graduell abgestuft gesehen werden, daß sie die Objektivität stärker gewährleistet als Reliabilität und diese wieder stärker als Validität. Zur Illustration diene die folgende Darstellung:

Abb. 14. Der Zusammenhang der Gütekriterien:



3.1.3 Normierung und Eichung von Tests

Wird ein psychologischer Test unter den bisher entwickelten Voraussetzungen, also standardisiert, objektiv, reliabel und valide durchgeführt, so gelangt man je nach Testaufbau und Auswertungsvorschrift zu bestimmten Meßwerten, die als gültige Maßzahlen für das zu erhebende theoretische Merkmal angesehen werden können.

Allerdings sind solche rohen Testpunktwerte für die Interpretation der festgestellten Merkmalsausprägung relativ unbrauchbar, was das folgende Beispiel illustrieren soll:

Nehmen wir an, zwei Versuchspersonen A und B würden derselben Testsituation mit demselben Erhebungsinstrument unterzogen. Die Person A sei weiblich und 18 Jahre alt, die Person B männlich und 30jährig. Der Testrohwert für Person A betrage 90 Punkte, der für die Person B 110 Punkte.

Da die Gütekriterien des Tests erfüllt sind, kann man bei der Interpretation davon ausgehen, daß die Person B offensichtlich einen höheren Punktwert als die Person A erzielt hat. Weitere Informationen sind aus den Gesamtpunktwerten nicht abzuleiten. So wissen wir in der Tat nicht, ob die Person B einen höheren Punktwert erzielt hat, weil sie männlich oder weil sie älter ist als die Person A oder ob eventuell beide Variablen mit dazu beigetragen haben, den Punktwert zu erhöhen. Unterziehen wir eine dritte Person C, die männlich und 30 Jahre alt sei, denselben Testbedingungen; sie erziele einen Gesamtpunktwert von 120. Nun haben wir weiteres Vergleichsmaterial gewonnen, das zusätzliche Interpretationsmöglichkeiten eröffnet. Da B und C sich im Hinblick auf die herausgegriffenen Merkmale nicht unterscheiden, kann davon ausgegangen werden, daß die Differenz in den Rohpunktwerten von 110 zu 120 tatsächlich eine solche des gemessenen Merkmals ist und nicht durch äußere Faktoren, die kontrolliert worden sind, determiniert wurde. Hätten wir eine vierte Person D, die männlich und 18 Jahre alt ist, mit einem Gesamtpunktwert von 91, so könnten wir auch in Bezug auf die Person A weitere Erkenntnisse gewinnen, nämlich, daß sich hinsichtlich des Geschlechtes keine Einflüsse auf die im Test gemessenen Merkmale ergeben. Würden wir eine fünfte Person E, die weiblich und 30 Jahre alt ist und einen Rohpunktwert von 110 erzielt, einbeziehen, so könnten wir auch Aussagen darüber vornehmen, inwieweit das Alter die gemessene abhängige Variable beeinflußt.

Das obige Beispiel sollte gezeigt haben, daß zu einer gültigen und brauchbaren Interpretation von Testrohwerten immer Vergleichszahlen zur Verfügung stehen müssen. Nun wird es aus statistisch-wahrscheinlichkeitstheoretischen Gründen nicht genügen, einem gemessenen Wert jeweils nur einen Kontrollwert gegenüberzustellen. Man ist also gezwungen, eine Fülle von Testwerten zu sammeln, um zu sehen, wie sich die Testwerte auf eine bestimmte, ausgewählte Population (z.B. nur Männer, eine bestimmte Altersgruppe etc.) verteilen. Aus einer solchen populationsspezifischen Verteilung lassen sich bestimmte statistische Parameter angeben, die diese Verteilung charakterisieren (z.B. bei der Normalverteilung: arithmetisches Mittel und Standardabweichung; vgl. hierzu 1.6). Erst durch die Vielzahl von Meßwerten und die daraus berechneten Parameter der Verteilung werden wir in die Lage versetzt, eine gültige, relationale Beurteilung der Testwerte vorzunehmen. Würde man z.B. festgestellt haben, daß der durchschnittliche Testwert für Männer bei 100 Punkten liegt, so könnte das Meßergebnis der Personen D und C als über dem Durchschnitt liegend interpretiert werden.

Da dieser Durchschnitt statistisch über eine Häufigkeitsverteilung gewonnen wurde, gibt das arithmetische Mittel die statistische Norm ab. Hiermit wird Häufigkeit des Auftretens einer bestimmten Merkmalsausprägung mit Normalität gleichgesetzt. Die statistische Norm ist eine Realnorm, weil sie an der realen Verteilung der Merkmalsausprägungen gewonnen wurde. Man könne natürlich nicht nur jene Meßwerte, die exakt dem arithmetischen Mittel entsprechen, als normal interpretieren. Vielmehr wird man auch eine bestimmte Schwankungsbreite um das arithmetische Mittel herum als normal akzeptieren müssen. Diese Schwankungsbreite wird in Standardabweichungen angegeben.

Halten wir fest: Um Meßwerte von Tests gültig beurteilen zu können, müssen wir die Resultate auf vergleichbare Meßwertreihen einer gleichen Population beziehen. Der Test muß sozusagen an der statistischen Norm ausgerichtet werden. *Er wird normiert oder geeicht. Diese Eichung erfolgt an einer Normstichprobe, die im Hinblick auf die zu testenden Personen repräsentativ sein muß.*

Auch hierzu ein Beispiel: Wenn der Durchschnittswert eines Testresultates für die Gesamtbevölkerung der Bundesrepublik Deutschland bei 100 liegt, der für Männer bei 90 und der für Frauen bei 110, so macht es offensichtlich einen Unterschied aus, ob konkrete Meßwertergebnisse auf die Gesamtbevölkerung oder nur auf die Frauen bezogen werden. Läge der Testwert einer Frau bei 110, so würde man, auf die Bevölkerung bezogen, von einem überdurchschnittlichen, auf die Frauen in der Bundesrepublik Deutschland bezogen, von einem durchschnittlichen Testergebnis sprechen. Deswegen muß bei der Eichstichprobe darauf geachtet werden, daß jene Variablen, von denen man glaubt, daß sie einen Einfluß auf das zu messende Merkmal haben, in der Eichstichprobe in entsprechender Weise vertreten sind. (Bei einem Intelligenztest würde man z.B. das Alter mitberücksichtigen müssen, bei einem Interessentest vielleicht das Geschlecht und bei einem projektiven Persönlichkeitstest eventuell die Schulbildung (Verbalisierungsvermögen).)

Zweck der Normierung ist, eine Vergleichbarkeit der Meßwerte herzustellen. Die bisherige Argumentation lief jedoch darauf hinaus, daß eine solche Vergleichbarkeit auf gleichgeartete Gruppen bezogen werden und daß jeweils ein gleicher Test zugrunde liegen müßte. Tatsächlich jedoch gibt es unterschiedliche Tests, die vorgeben, dasselbe Merkmal zu messen (vgl. die Intelligenztests). Die bisherigen Überlegungen haben nur eine *testimmanente Vergleichbarkeit* produzieren können, keine Vergleichbarkeit zwischen verschiedenen Testverfahren. Es würde also relativ komplizierter Umrechnungen bedürfen, will man Testergebnisse verschiedener Verfahren miteinander vergleichen.

Es ergäbe sich z.B. bei dem Test X ein arithmetisches Mittel von 100, beim Test Y ein solches von 150; die Anwendung des Tests X bei der Person a erbrächte einen Wert von 120, die Applizierung von Y bei b einen Testrohwert von 170. Die empirisch gewonnenen Werte müssen jeweils auf das durchschnittliche Maß umgerechnet werden, um miteinander vergleichbar zu sein. Ohne diese Umrechnung ist nicht entscheidbar, ob a oder b das relativ bessere Resultat erzielt hat.

Um eine solche *Intertestvergleichbarkeit* herzustellen, verwendet man nicht die Rohpunktwerte eines Tests, sondern man standardisiert nach z.T. unterschiedlichen Verfahren die Testrohwerte so, daß für jeden Test z.B. ein gleiches arithmetisches Mittel als Bezugspunkt und Basis gilt.

Die Normierung oder Eichung eines Tests verfolgt die folgenden Ziele:

1. Schaffung von im Hinblick auf relevant erachtete Variable homogenen Vergleichsgruppen, um die Meßwerte richtig interpretieren zu können.

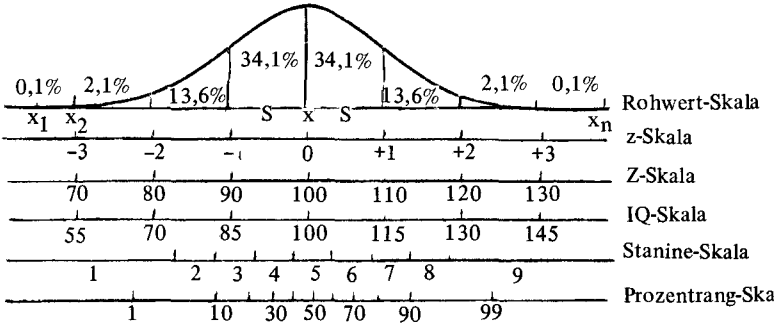
2. Beziehung der Meßwerte auf Gruppenwerte, die an der statistischen Norm orientiert sind.

3. Standardisierung der gemessenen Testwerte dahingehend, daß unterschiedliche Tests miteinander verglichen werden können.

Einige der Standardisierungsverfahren zum Zwecke der Normierung von Tests sollen im folgenden kurz besprochen werden, wobei die Abb. 15 eine Übersicht über die gebräuchlichsten Normskalen liefert. Die einfachste Form einer Normierung

stellt die *Prozentrangskala* dar. Die gewonnenen Testrohwerte der Eichstichprobe werden in Perzentile nach der Häufigkeit ihres Auftretens aufgeteilt. Jedem Testwert wird auf diese Weise ein Prozentrangwert zugeordnet. Hat z.B. eine Versuchsperson einen Prozentrangwert von 70 erreicht, dann kann dies dahingehend interpretiert werden, daß 70 % der Personen aus der Eichstichprobe einen niedrigeren und 30 % einen höheren Testwert erzielt haben. *Prozentrangwerte geben also die relative Position von Meßergebnissen im Vergleich zur Eichstichprobe an.* Da die Prozentrangwerte jedoch weniger differenziert sind als die Testrohwerte (so ist durchaus vorstellbar, daß einem Prozentrangwert von 70 Testrohwerte von 110 – 120 entsprechen), hat man mit ihrer Bildung einen Informationsverlust hingenommen, man hat das Meßniveau reduziert. Während die Testrohwerte noch auf Intervallksalenniveau lagen, mißt die Prozentrangskala nur auf ordinalem Meßniveau. Dieses hat zur Folge, daß die ursprünglich festgestellten größenmäßigen Unterschiede von Testleistungen (= Testrohwerte) in den Rangprozentwerten verschwinden, womit wegen dieser Meßniveaureduzierung Differenzenbildungen bei den Prozentrangwerten unsinnig und falsch wären. (Es kann nicht behauptet werden, daß eine Prozentrangwertdifferenz zwischen 60 und 70 genauso groß wäre wie die zwischen 70 und 80.)

Abb. 15. Übersicht über die wichtigsten Normskalen
(nach MICHEL, L., Allgemeine Grundlagen psychometrischer Tests, in: HEISS, R., (Hrsg.), Handbuch der Psychologie Bd. 6, Psychologische Diagnostik, Göttingen 1964, S. 29)



Ein Blick auf die vergleichende Tabelle der Normskalen zeigt auch, daß Prozentrangnormen ein relativ „verzerrendes" Bild der Tests insoweit liefern, als sich die Einheiten „von der Mitte nach außen" verändern. Da Prozentrangskalen wegen ihres ordinalen Meßniveaus diesen Nachteil aufweisen, der durch andere Normskalen vermeidbar ist, werden meist andere, insbesondere die *Standardnormskala* (z-Skala) angewandt.

Kann davon ausgegangen werden, daß die Häufigkeiten der Meßwerte für einen Test in der Eichstichprobe in etwa normalverteilt sind, so können diese Rohwerte in Standardnormwerte transformiert werden. Nimmt man nämlich Normalverteilung an, so können arithmetisches Mittel und Standardabweichung als deskriptive

Parameter dieser Verteilung zur Standardisierung benutzt werden. Während in der Rohwertverteilung die Abstände vom arithmetischen Mittel noch in Rohpunktwerten angegeben wurden (was aber eine Vergleichbarkeit mit anderen Tests verunmöglichte), wird das arithmetische Mittel nach einer Standardisierung den Wert 0 haben, und die einzelnen Abweichungen vom arithmetischen Mittel werden in Standardabweichungen ausgedrückt, wobei durch die Standardisierung eine Standardabweichung von 1 erzielt wird. Die Standardisierung wird nach der folgenden Formel vorgenommen: $z = \frac{x_i - \bar{x}}{\sigma}$, wobei $\bar{x}$ den Mittelwert, σ die Standardabweichung der Rohwertverteilung und x_i den jeweils einzelnen Testrohwert symbolisieren.

Eine Eichstichprobe habe ein arithmetisches Mittel von 120 und eine Standardabweichung von 10 ergeben. Ein außerhalb dieser Eichstichprobe getesteter Proband erreiche einen Testrohpunktwert von 130. Bei der Transformation in die z-Skala ergibt sich:

$$z = \frac{x_i - \bar{x}}{\sigma} = \frac{130 - 120}{10} = 1$$

Der standardisierte Meßwert liegt also eine Standardabweichung über dem arithmetischen Mittel.

Da Standardabweichungen sowohl nach rechts (also in den positiven Bereich) als auch nach links (in den negativen Bereich) gehen können, muß man das Vorzeichen beachten. Weil in der Berechnung Dezimalzahlen (z.B. 1,23 Standardabweichungen) auftreten können, verwendet man häufig nicht die Standardnormskala mit dem Mittelwert 0 und der Standardabweichung 1, sondern eine solche mit dem Mittelwert 100 und der Standardabweichung 10. Eine solche Standardnormskala wird als *Z-Skala* bezeichnet. (Die bekannte Abweichungsskala des Intelligenzquotienten arbeitet mit einer Normierung, die auf ein arithmetisches Mittel von 100 und eine Standardabweichung von 15 hinausläuft.)

Als letzte der bekannten Normskalen soll die sog. *Stanine-Skala* kurz besprochen werden. Stanine ist dabei zusammengesetzt aus „Standard" und „nine". Durch die Zahl 9 wird angedeutet, daß die Skala die Grenzen 1 und 9 aufweist, durch Standard wird hervorgehoben, daß es sich um eine standardisierte Verteilung, also um eine Transformation der Testrohwerte handelt. Die Stanine-Skala hat das arithmetische Mittel 5. Mit der Transformation einer Rohwertverteilung mittels der Stanine-Skala wird eine Normalisierung der Meßwerte erreicht. Wie aus der vergleichenden Abbildung der Normskalen ersichtlich, gilt für die Stanine-Skala, was für die Prozentrangskala ausgeführt wurde, nämlich daß den einzelnen Skalenwerten unterschiedliche Abstände zukommen. (Während z.B. innerhalb von zwei Standardabweichungen jeder Stanine-Wert von jedem benachbarten eine halbe Standardabweichung entfernt ist, gilt dieses für die äußeren Werte (1 und 9 zu 2 und 8 nicht mehr).)

Prozentrangskala und Stanine-Skala kann man zu den *Grobskalen* rechnen, während die Z-Skalen Feinskalen sind. Bedenkt man das fehlerbehaftete Messen im Test (vgl. z.B. Reliabilität etc.), so spiegeln Feinskalen eine Meßgenauigkeit vor,

die dem tatsächlichen Meßvorgang nicht entspricht. Daher könnte man davon ausgehen, daß für den Normalfall eine Grobskala zum Zwecke der vergleichenden Interpretation der Meßergebnisse durchaus ausreichend ist. Zwar werden in den meisten Tests Rohpunktwerte summierend gewonnen, sodaß das Meßniveau intervallskaliert ist, doch ist dieses hohe Meßniveau eher artifiziell als methodologisch begründet zustande gekommen. (Man vergleiche hierzu z.B. die Zeugnisnoten, die ja eindeutig auf ordinalem Meßniveau liegen, die aber allzu häufig so behandelt werden, als wären sie intervallskaliert.) Berücksichtigt man solche Überlegungen, dann ist die Prozentrangskala als verteilungsfrei und auf dem ordinalen Meßniveau liegend den meisten Tests wohl eher adäquat als die Standardnormskala. Der häufig vorgetragene Einwand, daß die Prozentrangskala im mittleren Meßbereich (um das arithmetische Mittel herum), dort also, wo die häufigsten Meßwerte der Rohwertverteilung zu finden sind, zu stark differenziert (was sich in den engen Abständen der Prozentränge niederschlägt) und an den unteren und oberen Endpunkten erhebliche Differenzen wegen der geringen Häufigkeiten nicht erfaßt, ist stichhaltig. Allerdings kann diesem Argument auch entgegengehalten werden, daß es gerade dort, wo bestimmte Merkmalswerte häufig auftreten, auf genauere Differenzierung ankommen kann.

Abschließend sei noch darauf hingewiesen, daß eine Standardisierung der Testrohwerte nur unter der Voraussetzung durchgeführt werden kann, daß diese Rohwerte sich annähernd normal verteilen. Die häufig geübte Praxis, unabhängig von der Rohwertverteilung eine Standardisierung vorzunehmen, ist ungültig. Zwar wird durch eine solche Standardisierung die Verteilung als solche nicht verändert (es ergibt sich nur eine Verschiebung des Mittelwertes und/oder eine Streckung oder Stauchung der Standardabweichung), doch sind die transformierten Werte, weil die Anwendungsbedingungen (Gauß'sche Normalverteilung) nicht erfüllt sind, keine Standardwerte.) Auch sei darauf hingewiesen, daß z.B. auf ordinalem Meßniveau erhobene Testwerte durch deren Standardisierung keine höhere Meßqualität erfahren, obwohl die Standardwerte ein intervallskaliertes Meßniveau vorspiegeln.

Die bisher vorgestellten Standardisierungsverfahren führen zu sog. *Variabilitätsnormen.* Der Begriff der Variabilität rührt daher, daß als Basis für die Beurteilung und Interpretation der individuell gemessenen Teilleistung die Variabilität der Meßwerte einer für die Versuchsperson adäquaten Vergleichsgruppe entstammen. Neben diesen Variabilitätsnormen gibt es *Äquivalenznormen,* wie etwa das Intelligenzalter, das Entwicklungsalter etc. Am Beispiel des Intelligenzalters können die Konstruktionen solcher Äquivalenznormen wie auch die Probleme, die damit verbunden sind, illustriert werden.

Das Intelligenzalter ist eine Maßzahl für den erreichten Intelligenzstand, wobei diese Maßzahl aus dem Vergleich mit dem durchschnittlichen Intelligenzniveau eines bestimmten Lebensalters gewonnen wird. So sagt man z.B., ein zehnjähriger Junge habe die Intelligenz eines siebenjährigen (Lebensalter = 10, Intelligenzalter = 7). Theoretische und methodologische Einwände sind gegen diese Äquivalenznorm erhoben worden: So glaubt man theoretisch belegen zu können, daß der beispielhaft angeführte Zehnjährige mit dem Intelligenzalter von 7 Jahren eben nicht die durchschnittliche Intelligenz eines Siebenjährigen besitzt, sondern eher vergleichbar ist mit einem äußerst minderbegabten Zehnjährigen. Im übrigen beinhalte ein solcher Vergleich eventuell diskriminierende Elemente, die nicht beabsichtigt und unerwünscht sind. Der

methodologische Einwand bezieht sich darauf, daß die Intelligenzalterseinheiten nicht konstant sind. Man weiß eben, daß nicht von Lebensjahr zu Lebensjahr ein gleicher Intelligenzzuwachs erzielt wird, was aber aufgrund der Zuordnung zu einem bestimmten Intelligenzalter, das ja in kardinalen Zahlen angegeben wird, also auf intervallskaliertem Meßniveau liegt, verschleiert wird. Weil unterschiedliche Intelligenzeinheiten zugrundeliegen, hat W. STERN vorgeschlagen, das Intelligenzalter auf das Lebensalter zu beziehen und diese Proportion den Intelligenzquotienten zu nennen. Wird dieser Quotient mit 100 multipliziert (womit Dezimalstellen vermieden werden), so hat man jene Maßzahl für die Intelligenz, die mit einer weiteren Modifikation als Abweichungs-IQ heute am häufigsten angewandt wird.

3.2 Anwendung und Beispiel

Nehmen wir an, wir würden beauftragt werden, einen Test für die Vorprüfung in medizinischer Psychologie zu entwerfen. Dieses Beispiel mag als Paradigma einer Testentwicklung trivial, ja geradezu primitiv sein, doch es eignet sich didaktisch, die Testkonstruktion zu demonstrieren. Der Nachvollzug einer Testkonstruktion von schon ausgearbeiteten, vorhandenen Tests bietet wegen deren Komplexität und Differenziertheit nur geringere Möglichkeiten der Verständniserleichterung. Andererseits hat ein Vorprüfungstest in der medizinischen Psychologie – wie er hier verkürzt und vereinfachend entwickelt werden soll – den Nachteil, daß seine Wiederholbarkeit an der gleichen Population durch Wiederholungseinflüsse ausgeschlossen ist, die Durchführung an einer anderen Population jedoch auch wegen des Examenscharakters der Fragen nicht beliebig oft vorgenommen werden kann. Das Beispiel erfüllt also nur bedingt die Definitionskriterien des Tests. (Als ausgezeichnetes und umfassendes Beispiel einer Testentwicklung sei auf Helmut BELSER, Testentwicklung, Weinheim und Basel 1975 verwiesen, wo der Frankfurter Analogietest in aller Ausführlichkeit und in allen Untersuchungsschritten behandelt wird.) Bei dem von uns zu entwickelnden Test muß auch in Kauf genommen werden, daß das theoretische Anspruchsniveau äußerst gering ist und deswegen einige erhebliche Probleme, wie sie für Tests charakteristisch sind (vgl. insbesondere die theoretische Darstellung zur Gültigkeit) nicht oder nicht in dem üblichen Ausmaße auftreten werden. Da häufig aber gerade theoretisch banale Fragestellungen, wie gerade in unserem Falle, auch existentielle Bedeutung haben können, erscheint – im Zusammenhang mit den didaktischen Chancen – das gewählte Beispiel durchaus nützlich.

3.2.1 Der Testentwurf (Instrumentarium)

Obgleich der zu konstruierende Test in medizinischer Psychologie keine erheblichen theoretischen Anforderungen stellt, muß vor Inangriffnahme konkreter Testentwicklungsarbeiten theoretisch abgeklärt werden, inwieweit der Untersuchungsgegenstand inhaltlich eingrenzbar erscheint. Gegenstand des Tests soll die Überprüfung des Wissens auf dem Gebiet der medizinischen Psychologie sein. Während normalerweise das zu messende Merkmal von theoretischen Vorüberlegungen, Annahmen und Bedingungen konstituiert wird (vgl. z.B. Aufmerksamkeitstests, Lerntests, Intelligenztests etc.), was mit Schwierigkeiten in der Operationalisierung ver-

bunden sein kann, scheint die hier zu bearbeitende Fragestellung kaum Probleme aufzuwerfen. Die Termini des Wissens, der Kenntnisse beziehen sich offensichtlich auf die Reproduzierbarkeit von Sachverhalten durch Gedächtnisleistungen (je nach Testform durch Erinnerungs- oder Leistungen des Wiedererkennens), wobei sich diese Sachverhalte auf das Gebiet der medizinischen Psychologie beziehen. Es müssen daher Methoden gefunden werden, die in der Lage sind, das Wissen abzufragen und eine theoretische Definition dessen geben, welches Wissen oder welche Sachverhalte zum Gebiet der medizinischen Psychologie gehören. Die letzte Fragestellung läßt sich nominalistisch unter Zuhilfenahme des Gegenstandskatalogs für das Medizinstudium klären. Alle jene Elemente, die dort unter dem Kapitel Medizinische Psychologie gesammelt sind, können Gegenstand einer Prüfung zu diesem Objektbereich werden. Eine solch nominalistische und vereinfachende Auffassung (nämlich medizinische Psychologie soll das sein, was im Gegenstandskatalog als solche bezeichnet ist), vermeidet Operationalisierungsprobleme, schafft aber die Gültigkeitsprobleme nicht aus der Welt. So wäre sicher vorstellbar, daß manche Autoren bestimmte Sachverhalte eher der medizinischen Soziologie als der medizinischen Psychologie zuordnen würden, obgleich dies im Gegenstandskatalog nicht so geschehen ist. Aber schon allein aus Praktikabilitätsgründen und solchen der Konsistenz (im Sinne der studentischen Erwartungen) scheint die nominalistische Definition des Objektbereichs angemessen.

Nun wird man davon ausgehen müssen, daß nicht in jeder Prüfung alle einzelnen Elemente des Gegenstandskataloges abgefragt werden können, sodaß für die Zwekke der Prüfung eine Auswahl vorgenommen werden muß, die in etwa dem Gesamtinhalt des Gegenstandskatalogs adäquat und repräsentativ ist.

Die Variable „Wissen" bedarf ebenfalls einer Operationalisierung. Auch hier kann nominalistisch entschieden werden, doch wird man dagegen erhebliche theoretische Einwände haben können. So kann durchaus darüber diskutiert werden, ob das Wissen als Gedächtnisleistung in der Form des Erinnerungsvermögens reproduziert werden soll, oder ob das Wissen durch Wiedererkennen abgefragt werden soll. Im letzteren Falle würde es sich technisch gesehen um geschlossene Fragen (vgl. 4.1.5), bei denen die richtige Antwortvorgabe herauszufinden ist, bei ersterem würde es sich um offene Fragen ohne Antwortvorgaben handeln, wo der Proband die Antworten selbst frei verbalisieren und formulieren muß. (Auf den medizinischen Bereich übertragen, wäre zu prüfen, ob ein Arzt in der Lage sein muß, alle Symptome einer bestimmten Krankheit abstrakt und theoretisch aufzuzählen, oder ob es nicht vielmehr darum geht, bei Auftreten dieser Symptome im konkreten Falle auf eine bestimmte Indikation schließen zu können.) Nehmen wir daher an (insbesondere auch wegen der Fülle des Stoffs zur medizinischen Vorprüfung), daß das Wiedererkennen die bessere Form der Wissensüberprüfung darstellt.

Ein zweites theoretisches Problem müßte in der Operationalisierung des Begriffs des Wissens noch geklärt werden. Einige werden davon ausgehen, daß der Test tatsächlich das akkumulierte Wissen überprüfen soll, wobei Wissen als Reproduzierbarkeit von Erlerntem verstanden wird. Andere hingegen mögen in diesem Zu-

sammenhang eher eine Wissensüberprüfung im Sinne von einem Nachweis des Verständnisses des zugrundeliegenden Sachverhaltes (selbstverständlich sind auch Mischformen denkbar) vorziehen. Sind all diese theoretischen Vorfragen geklärt, so ergeben sich erste Vorentscheidungen über Form und Auswahl der Items, die in den Test eingehen sollen.

Da sich die abzuprüfenden Inhalte aus dem Gegenstandskatalog ergeben, wird man diesen zur Basis der Itemformulierung wählen. Alle in ihm aufgeführten medizinisch-psychologischen Begriffe werden quasi als Schablone auf die medizinisch-psychologischen Lehrbücher aufgelegt, um die relevanten Fragestellungen und die richtigen Antworten zu ermitteln. Formal könnte die Itemformulierung so aussehen, daß auf jeden Fall eine geschlossene Frageformulierung gewählt wird. Wie bekannt, werden in den medizinischen Vorprüfungen unterschiedliche, formale Fragestellungen in jeweils geschlossener Frageform vorgegeben (vgl. hierzu RATHGEBER, W. (Hrsg.), Examensfragen zur medizinischen Soziologie, München 1975 und RATHGEBER, W. (Hrsg.), Examensfragen zur medizinischen Psychologie, München 1975).

Mit diesen theoretischen Entscheidungen ist zwar eine Eingrenzung dessen vorgenommen worden, was als Fragestellung in den Test eingehen kann, jedoch sind damit noch keine Angaben gemacht, wie darüber hinaus Fragen- und Antwortvorgaben zustandekommen. Dieses ist eine originär-kreative Tätigkeit, sie sich von Testkontrukteur zu Testkonstrukteur unterscheiden wird und nicht in generalisierende Aussagen gefaßt werden kann.

Will man auch fragetechnische Einflüsse in der Testkonstruktion berücksichtigen und kontrollieren, so empfiehlt es sich, unterschiedliche Techniken anzuwenden, wie z.B. Falschantwortaufgabe, Richtigantwortaufgabe, Zuordnungsaufgabe etc. Die folgenden 7 Fragen zur medizinischen Psychologie, die als Basis für unseren Test dienen sollen, entstammen der ärztlichen Vorprüfung vom März 1976:

1. Welche Antwort trifft zu?

Ein Patient erzielt in einem Intelligenztest einen Prozentrang-Wert von 50. Das bedeutet, daß von 100 vergleichbaren Personen einer Altersstichprobe etwa die Hälfte ein schlechteres Ergebnis erzielen würde. Der entsprechende Wert auf der IQ-Skala lautet:

(A) 50
(B) 75
(C) 100
(D) 125
(E) 150

2. Welche Aussage trifft zu?

Die Validität eines Tests gibt das Ausmaß an:

(A) in dem einzelnen Versuchspersonen bei wiederholter Testvorgabe zu übereinstimmenden Ergebnissen kommen

(B) in dem die einzelnen Untertests dasselbe messen

(C) in dem der Test zwischen verschiedenen Patientengruppen differenziert

(D) in dem ein Test zeitsparend ist

(E) in dem der Test das Merkmal, das er zu messen vorgibt, auch tatsächlich mißt

3. Welche Aussage trifft zu?

Der wesentlichste Einwand gegen projektive Testverfahren (z.B. Rorschach, TAT) ist:

(A) das Fehlen von Testkriterien

(B) die Offenheit der Interpretation

(C) der hohe Aufwand

(D) die hohe Bandbreite

(E) die Undurchschaubarkeit für den Probanden

4. Welche Aussage trifft n i c h t zu?

Folgende Begriffe beschreiben einen systematischen Beurteilungsfehler:

(A) Placebo-Effekt

(B) Einstellungsfehler

(C) Rosenthal-Effekt

(D) Halo-Effekt

(E) logischer Fehler

5. Welche Aussage trifft n i c h t zu?

Intelligenztests werden angewandt zur Einschätzung

(A) spezifischer Begabungen

(B) der Schulreife

(C) der Allgemeinbegabung

(D) der Kreativität

(E) hirnorganisch bedingter Leistungsstörungen

6. Auf welchen der folgenden Skalentypen sind Einstellungen erfaßbar?

(1) Intervallskala

(2) Rationalskala

(3) Nominalskala

(4) Ordinalskala

(A) nur 1 ist richtig

(B) nur 1 und 2 sind richtig

(C) nur 1, 2 und 3 sind richtig

(D) nur 1, 3 und 4 sind richtig

(E) 1 – 4 = alle sind richtig

7. Mangelnde Zuverlässigkeit eines psychologischen Tests kann verursacht sein durch

(1) Zeitinstabilität des Tests

(2) Zeitinstabilität des gemessenen Merkmals

(3) geringe Validität

(4) geringe Objektivität

(A) nur 4 ist richtig

(B) nur 1 und 2 sind richtig

(C) nur 3 und 4 sind richtig

(D) nur 1, 2 und 4 sind richtig

(E) nur 1, 3 und 4 sind richtig

Für die endgültige Testkonstruktion müßten weitere, theoretische Überlegungen Platz greifen. So könnte man einen solchen Test als *Speedtest* oder als *Niveautest* konstruieren. Man müßte die Testlänge auf die durch die Prüfungsordnung vorgegebene Zeit beziehen und prüfen, ob keine Überforderung vorliegt. Es müßte auch analysiert werden, ob die Zielgruppe, für die der Test konstruiert wird, in sich so homogen ist, wie die formale Definition als „Medizinstudierender" es vorzugeben scheint. So wäre durchaus denkbar, daß Ausbildungsniveauunterschiede von Universität zu Universität (z.B. durch Nichtbesetzung von Stellen etc.) gegeben sind, sodaß das minimalste Ausbildungsniveau berücksichtigt, also der kleinste gemeinsame Nenner gesucht werden müßte. Auch müßte vorab überlegt werden, ob alle Fragen einen gleichen Schwierigkeitsgrad aufweisen sollten, oder ob Unterschiede bei der Auswertung des Tests gewichtet werden sollen etc. Um die Sachlage nicht über Gebühr zu komplizieren, wollen wir zunächst davon ausgehen, daß der Schwierigkeitsgrad der einzelnen Fragen annähernd gleich ist und somit eine gleichgewichtige Bewertung angemessen erscheint.

3.2.2 Datenerhebung an einer Stichprobe

Ist der Test in der oben entwickelten vorläufigen Form erstellt, so geht man daran, eine erste Testprüfung durchzuführen. Der Test wird einigen Experten vorgelegt, um diese beurteilen zu lassen, wie die einzelnen Testitems inhaltlich und formal zu bewerten sind. Man wird weiter einige Tests an Versuchspersonen durchführen, diese auswerten und die Versuchspersonen danach befragen, inwieweit sie Probleme und Schwierigkeiten entdeckt und empfunden haben. Gerade im Hinblick auf die Itemformulierung bzw. die Abfassung von Antwortvorgaben, erlebt der Testkonstrukteur häufig Überraschungen, weil er glaubt, seine Fragen und Antworten wären eindeutig und interpersonell synonym formuliert, obgleich sich durch Befragung der Versuchspersonen dann herausstellt, daß Verständnis- und Interpretationsprobleme nicht ausgeschlossen sind.

Nach einer solchen vorläufigen Testprüfung, die noch eher qualitativen als quantitativen Charakter hat, kann man eine erste Rohfassung des Tests formulieren, wobei diese bereits eine erste Modifizierung der allerersten Überlegungen darstellen wird. Mit dieser Rohfassung geht man dann in ein größeres Untersuchungsfeld, um den Test auch quantitativ überprüfen und einigermaßen gesicherte Analysen und Aussagen vornehmen zu können. Aus diesem zweiten Testversuch heraus läßt sich dann das endgültige Instrumentarium entwickeln. Der auf dieser Basis formulierte und konstruierte Test wird einer repräsentativen Stichprobe der Population vorgelegt, auf die der Test appliziert werden soll. Diese repräsentative Stichprobe dient

dann als *Eichstichprobe, um den Test normieren zu können.* Diese drei Entwicklungsstufen in der Testkonstruktion sollen in unserem Beispiel auf eine Stufe reduziert werden. Nach Konstruktion der allerersten Fassung des Tests (s. hierzu die oben formulierten Fragen) wird diese einer Stichprobe von n = 200 Medizinstudierenden, die sich der ärztlichen Vorprüfung unterziehen wollen, vorgelegt. Aus der Auswertung dieses vorläufigen Tests soll die endgültige Testform hervorgehen, indem eine auf statistischem Wege vorzunehmende Aufgabenanalyse erste Erkenntnisse darüber liefern soll, welche Items „brauchbar" sind und welche aus dem endgültigen Test ausgeschlossen werden sollen.

Die Durchführung des ersten Tests erbrachte die folgende Verteilung richtiger Antworten auf die einzelnen Fragen.

Tab. 22. Verteilung der richtigen Testantworten

Frage	Zahl der richtigen Antworten bei n = 200
1	160
2	150
3	180
4	170
5	190
6	40
7	60

Ohne auf die methodologischen Gütekriterien von Reliabilität und Validität eingehen zu müssen, indiziert die (fiktive) Verteilung bereits einen unterschiedlichen Schwierigkeitsgrad der Fragen, wie der Vergleich zwischen Frage 5 und Frage 6 belegt. Während erstere in 190 Fällen richtig beantwortet wurde, kann dies nur für 40 Fälle bei letzterer behauptet werden. Daß eine solch große Differenz nicht zufällig entstanden sein kann, versteht sich von selbst.

Zunächst wird man wohl daran gehen, einige deskriptive Maßzahlen für diese Verteilung anzugeben. So wird man das arithmetische Mittel und die Standardabweichung berechnen wollen:

$$\bar{x} = \frac{\Sigma x_i}{n} = 135 \; ; \quad \sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}} = 60{,}2$$

Im Durchschnitt wird also jede Frage von 135 Personen richtig beantwortet, wobei eine doch erhebliche Streuung von etwa 60 Fragen zu verzeichnen ist. Der Modus der Verteilung (häufigster Wert) liegt bei 190 richtigen Antworten. Solche einfachen statistischen Maßzahlen liefern zwar Kurzinformationen über die gesamte Verteilung, reduzieren auch die Fülle von Informationen, die in der Verteilung selbst stecken, auf nur wenige Maßzahlen, womit sie zunächst für eine differenziertere Analyse nicht sinnvoll zu verwerten sind. (Zwar wird man sagen können, daß arithmetische Mittel, die näher bei dem maximal möglichen Wert von 200 liegen, auf einen leichteren Test hinweisen und geringere Standardabweichungen auf einen im Schwierigkeitsgrad relativ homogenen Test deuten, doch kann mit solchen Informationen nicht das einzelne Item quantifiziert werden.)

Durch die eigentliche *Itemanalyse* soll festgestellt werden, welche empirischen Informationen dazu herangezogen werden können, die Güte der formulierten Items zu prüfen. Eine erste Möglichkeit ist die Feststellung der Schwierigkeit der einzelnen Items. Mit einem relativ einfachen Verfahren wird ein *Schwierigkeitsindex*

konstruiert, der mit Hilfe der folgenden Formel berechnet werden kann:

$$I_{Schw} = \frac{\text{Zahl der richtigen Lösungen}}{\text{Zahl der Probanden}} \cdot 100$$

Die Anwendung dieses Schwierigkeitsindexes auf die einzelnen Fragen erbringt die folgenden Ergebnisse:

Tab. 23. Schwierigkeitsindex der Fragen

Frage	Index
1	80
2	75
3	90
4	85
5	95
6	20
7	30

Dieser einfache Index hat im Grunde genommen nicht mehr Information geliefert als die deskriptive Häufigkeitsverteilung der richtigen Antworten bereits enthalten hatte. Die einzige Veränderung besteht darin, daß nun die Meßwerte überschaubarer und vergleichbarer gemacht werden, weil sie alle auf die Basis 100 bezogen sind. (Bei diesem einfachen Schwierigkeitsindex ist allerdings zu beachten, daß vorausgesetzt wird, daß alle Versuchspersonen jede Aufgabe beantwortet haben. Für die Fälle, wo nicht alle Fragen von allen beantwortet worden sind, muß die obige Formel korrigiert und modifiziert werden. Hierzu sei jedoch auf die Spezialliteratur verwiesen, z.B. LIENERT, Testaufbau und Testanalyse, Weinheim 1976, S. 88ff.) Höhere Werte des Schwierigkeitsindex deuten auf leichte, niedrigere Werte auf schwierigere Aufgaben. Deshalb wäre es sinnvoller, den Schwierigkeitsindex als Leichtigkeitsindex zu bezeichnen, oder wie er exakter und wahrscheinlichkeitstheoretisch richtiger verstanden und gebraucht werden sollte, als *„Lösungswahrscheinlichkeit"*.

Der Schwierigkeitsindex allein ist nur eine deskriptive Maßzahl, die keinerlei Aussagen darüber macht, weshalb eine Frage schwieriger und eine andere leichter ist. So kann in dem obigen Beispiel die höhere Schwierigkeit der Frage 6 (und auch der Frage 7) an der unterschiedlichen „technischen" Fragestellung liegen, denn darin unterscheiden sich beide Fragen offensichtlich von den vorausgehenden. Es wäre aber auch denkbar, daß die Frageform selbst keinen Einfluß hat, die Inhalte jedoch, aus welchen Gründen auch immer, von den Probanden als schwieriger empfunden wurden. Daher wird man bei einem solch einfachen Index nicht stehenbleiben können; weitere Verfahren der Itemanalyse müssen herangezogen werden.

Die Berechnung eines *Trennschärfekoeffizienten* für die einzelnen Testaufgaben bietet eine erweiterte Erkenntnismöglichkeit und wird auch häufig vorgenommen. *Der Trennschärfekoeffizient ist ein Korrelationskoeffizient* (der also Werte zwischen − 1 und + 1 annehmen kann), *der angibt, wie stark der Zusammenhang der Lösung einer Aufgabe mit den Lösungen aller Aufgaben des Tests ist.* Ein hoher Trennschärfekoeffizient gibt für ein bestimmtes Item an, daß dieses häufiger von jenen Personen richtig beantwortet wurde, die auch die meisten anderen Aufgaben des Tests richtig gelöst haben. Ein negativer Trennschärfekoeffizient drückt aus, daß die Versuchspersonen ein Item häufiger gelöst haben, die ansonsten im

gesamten Test schlechter abschnitten. Ein solch „verkehrtes" Ergebnis wird z.B. durch mißverständliche Frageformulierung und ähnliche Fehler provoziert. Bestimmen wir daher den Trennschärfekoeffizienten für die Frage 6, die offensichtlich den höchsten Schwierigkeitsgrad aufweist.

Tab. 24. Berechnung des Trennschärfekoeffizienten für Frage 6 (fiktive Daten):

Zahl richtiger Antworten im Test	davon Zahl der bei Frage 6 richtigen Antworten	Rangreihe der richtigen Antworten insges.	Rangreihe der Zahlen der richtigen Antworten bei Frage 6	d_i	d_i^2
7	8	1	6	− 5	25
6	9	2	5	− 3	9
5	10	3	3,5	− 0,5	0,25
4	10	4	3,5	+ 0,5	0,25
3	14	5	1	+ 4	16
2	12	6	2	+ 4	16
					$\Sigma d_i^2 =$ 66,5

Die Tabelle 24 zeigt, daß in 8 Fällen alle Fragen richtig beantwortet wurden, während z.B. die Zahl der Fälle, bei denen 4 Fragen, zugleich aber auch die Frage 6 richtig beantwortet wurde, 10 beträgt. Aus der Verteilung dieser Daten kann schon entnommen werden, daß eine geringere Zahl richtig beantworteter Fragen mit größerer Häufigkeit dafür einhergeht, daß zugleich die Frage 6 richtig beantwortet wurde. Um diesen Sachverhalt noch übersichtlicher zu gestalten, wurde die Tabelle in zwei Rangreihen transformiert, die den negativen Zusammenhang deutlich indizieren. Mit der Herstellung zweier Rangreihen haben wir zwar das Meßniveau auf ein ordinales reduziert (nehmen also einen gewissen Informationsverlust hin), doch soll es für unsere Zwecke genügen, einen auf ordinalem Meßniveau liegenden Trennschärfekoeffizienten, nämlich den Korrelationskoeffizienten ρ zu berechnen.

$$\rho = 1 - \frac{6\,\Sigma d_i^2}{n^2(n-1)} = 1 - \frac{6 \cdot 66{,}5}{6^2(6-1)} = -0{,}9$$

Die Berechnung zeigt einen sehr hohen negativen Zusammenhang von − 0,9, womit indiziert wird, daß die im Test schwächeren Studenten die Frage 6 häufiger richtig gelöst haben als die allgemein besseren Studenten, obgleich die Frage 6 eine weitaus geringere Lösungswahrscheinlichkeit aufwies. Prüft man, ob dieses Ergebnis zufällig entstanden sein kann (hierfür gibt es in den Statistiklehrbüchern für kleine n (in unserem Fall n = 6) für ein vorzugebendes Signifikanzniveau von z.B. α = 0,05 bestimmte Grenzwerte von ρ, die mindestens erreicht werden müssen, um ein zufälliges Entstehen des Koeffizienten ausschließen zu können. Der Blick in eine solche Tabelle zeigt uns, daß unser Ergebnis signifikant ist, weil ein ρ von mindestens |0,83| hätte erzielt werden müssen, unser Wert jedoch größer ist.

Die Trennschärfeanalyse für die Frage 7 könnte ein ähnliches Resultat liefern. So wie wir oben einen Rangkorrelationskoeffizienten für die Trennschärfe berechnet haben, hätte man aber auch einen auf höherem Meßniveau liegenden anwenden können, wie man auch andere Metho-

den der Trennschärfebestimmung heranziehen kann. Eine häufig praktizierte Trennschärfebestimmung (z.B. bei der Likertskalierung) besteht darin, daß Extremgruppen im Hinblick auf ein bestimmtes Item mittels t-Test miteinander verglichen werden. So sucht man sich z.B. jene 25% der Versuchspersonen aus, die die höchsten Testwerte erreicht haben, und vergleicht sie mit jenem Viertel, die die niedrigsten Testwerte erzielten im Hinblick auf ein bestimmtes Item. Je größer der Unterschied zwischen den beiden Gruppen bei einer Frage ist, desto besser ist die Trennschärfe dieser Frage.

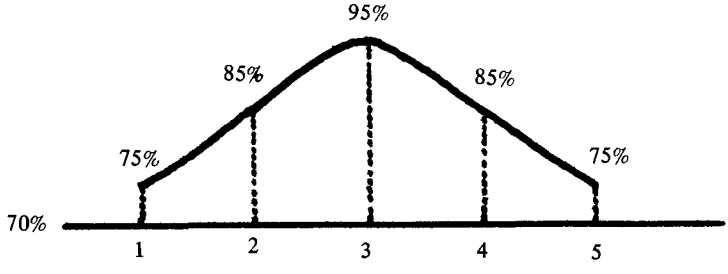
Die Durchführung des t-Tests für Item 7 des entwickelten Tests soll Anhaltspunkte dafür ergeben haben, daß es wegen mangelnder Trennschärfe sinnvoll sein könnte, diese Frage aus dem endgültigen Test auszuschließen. Wir entscheiden uns daher dafür, die Fragen 6 und 7 ersatzlos zu streichen, sodaß sich die endgültige Testform aus den Fragen 1 – 5 konstituiert. Da der Test nur durch Streichung modifiziert wurde und keine neuen Items hinzukommen, ist es nicht erforderlich, den Test an einer erneuten Stichprobe anzuwenden, da die in den Test gelangten Items 1 – 5 bei der ursprünglichen Stichprobe von n = 200, die repräsentativ gezogen worden war, schon erhoben wurden. In die weitere Analyse und Auswertung gehen daher nur die Daten ein, die in der Tab. 25 dargestellt sind:

Tab. 25: Ergebnisse des Tests für 5 Fragen bei n = 200 (richtige Antworten)

Richtige Antworten	Zahl der richtigen Antworten	relative Zahl der richtigen Antworten
5	150	75 %
4	170	85 %
3	190	95 %
2	170	85 %
1	150	75 %

Diese Tabelle fiktiver Daten kann auch grafisch gestaltet werden, sodaß die Verteilung der richtigen Antworten (natürlich auch der falschen Antworten) illustrativ erkennbar wird:

Abb. 16. Ergebnisse des Tests für 5 Fragen bei n = 200 (richtige Antworten grafisch und prozentuiert)



3.2.3 Testprüfung und Testnormierung

Die an den Daten gewonnene, annähernd empirische Normalverteilung verleitet natürlich dazu, den Test so zu akzeptieren, wie er in die Datenerhebung eingegangen ist, weil eine solche idealisierte Verteilung als frei von Fehlern und Verzerrungen perzipiert wird. Tatsächlich jedoch darf sie uns keinesfalls davon abhalten, die theoretisch entwickelten Gütekriterien auf den hier entwickelten Test anzuwenden.

Da der Test hochstandardisiert ist, indem die Fragen und Antworten schriftlich formuliert, somit also inhaltlich und in der Reihenfolge identisch sind, die Durchführungssituation standardisiert ist (allen Versuchspersonen wird unter gleichen äußeren Bedingungen der Fragebogen vorgelegt), und die Auswertung nach Vorgabe einer Musterlösung durch Auflegen von Schablonen mechanisch und ohne systematische Verzerrungen vorgenommen werden kann, steht außer Zweifel, daß Objektivität des Testverfahrens gewährleistet ist. Dies umso mehr, als auch die Interpretationsobjektivität der Auswertungsergebnisse subjektiven Einflüssen entzogen ist, wenn bestimmte Grenzwerte als allgemein gültig festgelegt werden: Soll z.B. nur entschieden werden, ob die Prüfung bestanden oder nicht bestanden ist, so könnte man mindestens drei richtige Antworten als notwendig für das erfolgreiche Absolvieren des Testes ansehen. Auch differenziertere Vorgehensweisen, indem Notenabstufungen gegeben werden, sind interindividuell zuverlässig vorzunehmen und jeder subjektiven Verzerrung entzogen (was nicht ausschließt, daß bewußte Fälschungen vorkommen könnten), wenn die Zuordnung der Fehlerzahlen (oder Richtigantworten) zu den einzelnen Notenstufen allgemein fixiert wird.

Nachdem die Frage nach der Objektivität positiv beantwortet werden konnte, ist die Zuverlässigkeit des Tests zu überprüfen. Hierfür stehen, wie oben ausgeführt, unterschiedliche Verfahren zur Verfügung: Die Test-Retest-Methode wird sich bei einem solch spezifischen Leistungstest verbieten. Wählt man nämlich die Zeitspanne zwischen erstem und zweitem Test zu groß, besteht die erhebliche Gefahr, daß der zu messende Objektbereich sich verändert, d.h. also, daß das fürs Examen angelernte Wissen nicht mehr verfügbar ist. Wählt man die Zeitspanne zu kurz, ist das Erinnerungsvermögen an die einzelnen Fragen in einem solchen Ausmaße gegeben (insbesondere wenn nach dem ersten Test versucht wird, anhand der Literatur die richtigen Antworten herauszufinden), daß verzerrende Einflüsse auf den Reliabilitätskoeffizienten nicht ausgeschlossen werden können. Die Methode der Testhalbierung kann in unserem einfachen Beispiel auch nicht angewendet werden, weil die Zahl der Fragen zu gering ist, als daß zwei Tests konstruiert werden könnten, die in Inhalt und Schwierigkeitsgrad einander ähnlich sind (hierzu wären, um einen groben Anhaltspunkt zu geben, mindestens 20 Items notwendig, die in zwei Tests zu je 10 Items einfließen könnten). Die Zuverlässigkeitsprüfung müßte daher über etwa vorhandene Paralleltests erfolgen. Nehmen wir daher an, es existierte ein paralleler Test, der den gleichen Objektbereich messe. Wir würden also diesen Test unserer Versuchsstichprobe von n = 200 vorlegen und würden nach Auswertung des Paralleltests eine Korrelation aufstellen zwischen den Ergebnissen des hier

konstruierten Tests mit den Resultaten des Paralleltests. Aus Vereinfachungsgründen führen wir diesen Test nicht für alle 200 Versuchspersonen durch, sondern geben für 10 Versuchspersonen die Punktwerte der beiden Tests im Vergleich an:

Tab. 26. Zuverlässigkeitsprüfung mit einem Paralleltest

V_p	Test 1 (x_i)	Paralleltest (y_i)
A	5	10
B	5	8
C	4	9
D	4	9
E	4	8
F	3	7
G	3	6
H	2	5
I	2	5
J	1	4
Σ	33	71

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{33}{10} = 3{,}3$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = 1{,}34$$

$$\bar{y} = 7{,}1$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n}} = 2{,}02$$

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \; \Sigma(y_i - \bar{y})^2}}$$

$$r = \frac{22{,}6}{\sqrt{16{,}1 \cdot 36{,}9}} = \frac{22{,}6}{24{,}37} = 0{,}93$$

Die Berechnung des Korrelationskoeffizienten r nach BRAVAIS-PEARSON (der intervallskalierte Daten zur Voraussetzung hat, die in der Tat auch vorliegen), erbringt einen Reliabilitätskoeffizienten von + 0,93. Da wir davon ausgegangen waren, daß Koeffizienten der Größe 0,9 und höher als ausreichend angesehen werden können, dürfen wir von unserem entwickelten Test behaupten, er sei zuverlässig.

Nachdem der Test als objektiv und zuverlässig ausgewiesen ist, muß nun die wichtigste, theoretische Ebene, die Gültigkeit geprüft werden. Eine dieser Gültigkeitsprüfungen war bereits angesprochen worden, indem darauf hingewiesen wurde, daß der Test Experten zur Gültigkeitsbeurteilung vorgelegen hat. Wie bereits im theoretischen Teil ausgeführt, ist dieses eine sehr oberflächliche und problematische Gültigkeitsprüfung. Eine brauchbarere Methode wäre jedoch die Known-Group-Validity, wenn man als Known-Groups jene Experten bezeichnet, die sich professionell mit dem Gebiet der medizinischen Psychologie beschäftigen. Würde man ihnen den Test vorlegen (unterstellt natürlich, sie hätten ein umfassendes und in der Testsituation reproduzierbares Wissen), so müßten eigentlich alle Fragen beantwortet werden können. Treten von diesem Idealzustand Abweichungen auf, so könnte dies ein Indiz dafür sein, daß der Test nicht das erfaßt, was er eigentlich messen soll; es könnte aber nicht angegeben werden, welches die Gründe dafür sind, daß der Test nicht gültig ist. So könnte man sich im Extremfall durchaus vorstellen, daß der Test sehr wohl hohe Gültigkeit besitzt, daß aber die Experten (da

nicht gewohnt, sich einer solchen Form von Prüfung zu stellen), in der Testsituation psychische Belastungen entwickeln, die die Reproduzierbarkeit des Wissens erschweren und somit ein falsches Gültigkeitsbild vermitteln.

Auch die Vorhersagevalidität ist in unserem Beispielfalle nur äußerst schwer zu überprüfen; wählte man als Außenkriterium für die Prognose, ob jemand ein guter oder ein schlechter Mediziner wird (und würden wir unterstellen, daß solche Zuordnungen relativ einfach vorzunehmen wären), so müßte theoretisch hypostasiert werden, daß zwischen einer guten Testleistung in medizinischer Psychologie und der später entwickelten Qualität im Beruf eine hohe Assoziation besteht. Selbst wenn man davon abstrahiert, daß eine solche Validierung aus Praktikabilitätsgründen in unserem Beispiel ausscheidet (Faktor Zeit), so ist unmittelbar plausibel und einsichtig, daß bei Vorliegen einer solchen Korrelation und bei Verwendung dieser Validierungsmethode der Test in medizinischer Psychologie nur schwer zu validieren ist (so könnte er eher ärztliche Qualitäten messen, oder die Fähigkeit, sich im Berufsleben in entsprechender Weise zu exponieren etc.).

Bei der Überprüfung der Gültigkeit des Tests zur medizinischen Psychologie kann unter Berücksichtigung der einschränkenden theoretischen Bedingungen eine Expertenvalidierung durchaus ausreichen, indem man durch einen herbeizuführenden Konsens den Objektbereich gemäß Konvention eingrenzt und jeweils für die Testitems entscheidet, ob der Objektbereich durch sie repräsentiert wird. Diese konventionalistisch-nominalistische Validitätsprüfung kann theoretisch verantwortet werden, weil der hier konstruierte Test unmittelbar und direkt das mißt, was durch die Operationalisierung angestrebt wird: Kenntnisse in medizinischer Psychologie. Bedeutend schwieriger gestaltet sich die Validitätsprüfung in all jenen Fällen, wo eine indirekte Messung vorgenommen werden muß, weil die zu erfassenden hypothetischen Konstrukte nicht unmittelbar operationalisier- und beobachtbar erscheinen. (Kreativität in einem Testverfahren zu operationalisieren dürfte deutlich schwieriger sein, womit sich die Gültigkeitsprüfung in erheblichem Ausmaß verschlechtert.)

Da Gültigkeitsüberprüfungen – methodologisch gesehen – häufig zirkulär sind (vgl. hierzu die theoretischen Ausführungen), bewegen sich die Feststellungen zur Gültigkeit mehr oder weniger auf argumentativ-theoretischer Ebene, indem durch Plausibilisierungsversuche der Nachweis angestrebt wird, das entwickelte Erhebungsinstrument erfasse tatsächlich das theoretisch Gemeinte. Unter solchen Bedingungen kann der theoretisch äußerst voraussetzungslose, hier entwickelte Test als gültig akzeptiert werden (obgleich konzediert werden müßte, daß eine erhebliche Erweiterung der Items notwendig wäre; denn man kann einen doch sehr weiten Objektbereich wie den der medizinischen Psychologie nicht adäquat durch 5 Items repräsentieren).

Nach diesen theoretischen Vorüberlegungen, die notwendige Vorbedingung für das weitere Vorgehen sind, können wir die *Normierung des Tests* in Angriff nehmen. Zunächst ist dabei die Frage zu klären, welches die *Normierungsparameter* sind,

die für unseren Test in Frage kommen. Die Eingrenzung auf eine bestimmte Population, die dem Test unterzogen werden soll, ergibt sich quasi automatisch aus dem Untersuchungsgegenstand: Der Objektbereich der medizinischen Psychologie ist Prüfungsgegenstand der ärztlichen Vorprüfung, womit sich alle Medizinstudierende dem entwickelten Test zu unterziehen haben. Sind jedoch evtl. weitere Normierungsparameter zu berücksichtigen?

Geschlechtsspezifische Differenzierungen scheiden ebenso aus wie altersspezifische, weil eine Gleichbehandlung aller Prüfungskandidaten gefordert ist. Somit bedarf es für diese beiden Variablen keiner spezifischen Normgruppe. Denkbar wäre jedoch, daß hinsichtlich des Ausbildungsstandes an den einzelnen Universitäten aus Gründen der personellen Kapazität Gewichtungsfaktoren erarbeitet werden sollen, die eine ungleiche Ausbildungsmöglichkeit kompensieren. Da jedoch die Prüfung eine bundeseinheitliche ist und solche Gewichtungsfaktoren nur schwer theoretisch wie praktisch legitimierbar erscheinen, fällt auch diese Variable als Normierungsgröße aus. Auf diese Weise könnte man sozusagen negativ ausgrenzend alle nur denkbaren Variablen überprüfen, die in einem theoretisch-hypothetisch vermuteten Zusammenhang zum Testergebnis stehen könnten. Im Regelfalle wird jedoch eine solche Ausgrenzung positiv vorgenommen, d.h. in unserem Beispiel kann man durch die Vorgabe bestimmter rechtlicher Bedingungen spezifische Normgruppen ausschließen, sodaß die ursprüngliche Stichprobe von n = 200 Medizinstudierenden, die repräsentativ gewonnen wurde, als *Eichstichprobe* gelten kann. Die für die Eichstichprobe gewonnenen Informationen wurden bereits in der Häufigkeitsverteilung angegeben, wobei sich eine annähernde Normalverteilung ergab (vgl. Tab. 25 und Abb. 16). Eine solche Normalverteilung ist durch zwei Verteilungsparameter gekennzeichnet, das arithmetische Mittel und die Standardabweichung. Diese beiden Meßgrößen wollen wir im folgenden berechnen (vgl. Tab. 25.)

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad \bar{x} = 3$$

$$s^2 = \frac{\Sigma(x_i - x)^2}{n} \qquad s^2 = 1{,}85$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} \qquad s = 1{,}36$$

Wir ermitteln also ein arithmetisches Mittel von 3 und eine Standardabweichung von 1,36. Wir wissen nun, daß das durchschnittliche Ergebnis bei 3 richtigen Antworten liegt und können abschätzen, ob jemand besser oder schlechter als der Durchschnitt ist. Dieses gilt jedoch nur immanent für den hier vorgestellten Test. Normierungsparameter sollten jedoch auch dazu dienen, Testergebnisse unterschiedlicher Verfahren miteinander vergleichbar zu machen. Eine solche Vergleichbarkeit wird erzielt, indem eine Standardisierung vorgenommen wird, sodaß verschiedene Mittelwerte verschiedener Tests gleichgesetzt werden. Um die hiermit

verfolgten Absichten zu verdeutlichen, sei noch einmal auf das Beispiel der Reliabilitätsprüfung zurückgegangen. Dort wurden arithmetische Mittelwerte von 3,3 bzw. 7,1 für den Paralleltest erzielt (vgl. Tab. 26). Zwar sind diese beiden Werte miteinander direkt vergleichbar, weil sie eben die Durchschnitte der Häufigkeitsverteilung darstellen, jedoch können andere Merkmalsausprägungen der beiden Verteilungen nicht unmittelbar im Vergleich beurteilt werden. So ist nicht entscheidbar, ob ein erzielter Punktwert von 2 in dem Test mit dem arithmetischen Mittel 3,3 besser oder schlechter ist als ein Punktwert von 5 im Paralleltest mit dem arithmetischen Mittel 7,1. Mit Hilfe der Standardisierung jedoch gelingt es, eine solche Vergleichbarkeit und damit bessere und genauere Interpretierbarkeit von Testergebnissen herbeizuführen. Wir erstellen daher für unseren Test Normwerte auf der Basis der an der Stichprobe gewonnenen Testrohwerte – die sich als Summierung der richtigen Antworten ergeben hatten – mit Hilfe der zu entwickelnden Standardnormskala. Wir transformieren also unsere gewonnenen Untersuchungsergebnisse in eine Verteilung mit dem arithmetischen Mittel von 0 und der Standardabweichung von 1. Dies geschieht über die im theoretischen Teil schon angegebene Formel:

$$z = \frac{x_i - \bar{x}}{s}$$

Hierbei gibt z eine Maßzahl dafür an, inwieweit empirisch gewonnene Testwerte bei positivem Vorzeichen rechts, bei negativem Vorzeichen links vom arithmetischen Mittel liegen; die absolute Größe von z gibt die Zahl der Standardabweichungen an, die der Testwert vom arithmetischen Mittel entfernt liegt.

Waren wir ursprünglich davon ausgegangen, daß bei der Rohwertinterpretation, d.h. der Festlegung ob bestanden oder nicht bestanden, die Testrohpunktwerte herangezogen wurden, so sind wir jetzt in der Lage, gleiche Maßstäbe auf unterschiedliche Tests anzuwenden, indem wir die festzulegenden Grenzen bei der Interpretation der Meßergebnisse durch den Multiplikator der gleichen Standardabweichungen angeben können. So könnte man festlegen, daß alle Resultate, die mehr als eine Standardabweichung unter dem arithmetischen Mittel liegen, als ungenügende Testleistung zu interpretieren sind. Die Standardnormskala erlaubt es uns auch, die oben apostrophierten unterschiedlichen Meßergebnisse der beiden Tests miteinander zu vergleichen:

$$z_1 = \frac{x_i - \bar{x}}{s_x} \qquad z_2 = \frac{y_i - \bar{y}}{s_y}$$

$$z_1 = \frac{2 - 3{,}3}{1{,}34} \qquad z_2 = \frac{5 - 7{,}1}{2{,}02}$$

$$z_1 = -0{,}97 \qquad z_2 = -1{,}04$$

Für die beiden Beispiele, die bei der Reliabilitätsprüfung genannt wurden, gilt also, daß der Proband, der im Test 1 zwei richtige Ergebnisse erzielt hat, 0,97 Standard-

abweichungen schlechter als der Durchschnitt ist, während der Proband, der im Paralleltest 5 Richtige erzielt hat, um 1,04 Standardabweichungen schlechter als das arithmetische Mittel liegt. Der Proband in Test 1 ist also geringfügig besser als der im Paralleltest.

3.2.4 Interpretation von Meßwerten

Je nach dem, welche Interpretationsregeln zu einem Test angegeben werden, ergeben sich qualitative oder quantitative Deutungsmöglichkeiten. Im einfachen diskreten Falle, wo es nur darum geht, zwischen bestanden und nicht bestanden zu differenzieren, würde es genügen, eine untere Grenze der Standardnormskala anzugeben, beispielsweise 1,5 Standardabweichungen links vom arithmetischen Mittel als Grenzwert. (Solche Angaben sind natürlich relativ sinnlos, wenn nur wenige Items einen Test konstituieren. So müßte jemand bei einem arithmetischen Mittel von 3 und einer Standardabweichung von 1,36 in unserem Test einen Rohpunktwert von 0,96 erzielen, um 1,5 Standardabweichungen unterhalb des arithmetischen Mittels zu liegen. Ein solcher Wert kann wegen ganzzahliger Punktewertung natürlich nicht auftreten. Hat man hingegen 30 Fragen mit einem arithmetischen Mittel von 15 und einer Standardabweichung von 4, so ergibt sich die erforderliche Rohpunktzahl, die zum erfolgreichen Teilnehmen erreicht werden muß, als sinnvoller Wert, nämlich als 9.)

Wären abgestufte Noten zu vergeben, so könnte man diese ebenfalls über die Standardabweichungen der Standardnormskala definieren und so Interpretationen eindeutig zu machen. Will man z.B. eine annähernde Gleichbehandlung aller Prüfungskandidaten über die einzelnen Prüfungstermine hinweg erzielen, wobei in jedem Prüfungstermin unterschiedliche Tests praktiziert werden, so ist die Möglichkeit einer Notendefinition mittels der Standardabweichungen einer Standardnormskala optimal zu bestimmen, insbesondere dann, wenn die Tests nicht nur inhaltlich, sondern auch quantitativ (auf die Anzahl der Items bezogen) variieren. Andererseits werfen solche Festlegungen jedoch auch normative Fragen nach Durchfallquoten etc. auf, die nicht durch statistische Operationen entscheidbar wären. Mit Hilfe der Standardisierung soll es auch nur gelingen, von einer absoluten und testimmanenten Interpretation von Untersuchungsergebnissen zu relativen, aussagekräftigeren, testübergreifenden Informationen zu kommen.

Ergänzende und vertiefende Literatur:

BELSER, H., Testentwicklung, Weinheim 1972

FISCHER, G.H., Einführung in die Theorie psychologischen Testens, Bern 1974

HEISS, R. (Hrsg.), Handbuch der Psychologie, Bd. 6, Psychologische Diagnostik, Göttingen 1964(2)

LIENERT, G.A., Testaufbau und Testanalyse, Weinheim 1967(2)

4 INTERVIEW UND SCHRIFTLICHE BEFRAGUNG

Vorab sei eine terminologische Unschärfe, bzw. Inkonsistenz in der methodischen Literatur für dieses Kapitel ausgeräumt. Einige Autoren fassen Interview als Oberbegriff für jede Form einer kommunikativen Datenerhebung auf, während andere das Interview als mündliche Befragung der schriftlichen Befragung gegenüberstellen. Wir wollen im weiteren davon ausgehen, daß Befragung als Oberbegriff für die gleichrangigen Begriffe der schriftlichen und der mündlichen Befragung (= Interview) steht.

Ob man das Interview als den „Königsweg der empirischen Sozialforschung" ansieht (KÖNIG, Praktische Sozialforschung, Bd. 1, Das Interview 1966, S. 27), oder ob man der Befragung als Methode kritisch gegenübersteht (vgl. hierzu ROEDE H., Befrager und Befragte, Probleme der Durchführung des soziologischen Interviews, Berlin 1965), grundsätzlich dürfte unbestritten sein, weil empirisch bestätigt, daß die Befragung die am häufigsten angewandte Methode der Datenerhebung ist, und daß diese Methode einen erheblichen Beitrag zum gegenwärtigen, wissenschaftlich abgesicherten Wissen beigetragen hat (vgl. hierzu SCHEUCH E.K., Das Interview in der Sozialforschung in: R. KÖNIG, Hrsg. Handbuch der empirischen Sozialforschung Bd. 2, Grundlegende Methoden und Techniken der empirischen Sozialforschung, 1. Teil, Stuttgart 1973, S. 66). Diese bevorzugte Stellung der Befragung ist jedoch keineswegs darauf zurückzuführen, daß sie methodische oder methodologische Vorzüge vor allen anderen Techniken hat; vielmehr wird sich zeigen, daß sie erhebliche meßtheoretische Nachteile gegenüber anderen Methoden aufweist. Die gleichwohl hohe Beliebtheit und die enorme Einsatzhäufigkeit dürften mit darauf zurückzuführen sein, daß Gesichtspunkte der Praktikabilität und der materiellen Ressourcen eine Entscheidung zugunsten der Befragung provozieren. Während das Experiment mehr oder weniger nur als Instrument der Hypothesenprüfung eingesetzt und der Test zum Zwecke der individuellen Datensammlung benutzt wurde, eignen sich Befragungen (sowohl mündliche, wie schriftliche) zu explorativen Erhebungen, um relevante Dimensionen eines Objektbereiches in Erfahrung zu bringen, um also zunächst einmal die Hypothesen formulieren zu können, als Instrument einer quantitativ umfangreichen und preiswerten Datensammlung, als Methode mit deren Hilfe Forschungsergebnisse, die mit anderen Techniken gewonnen wurden, kontrolliert und nicht zuletzt auch als eine Technik, mit deren Hilfe Hypothesen überprüft werden können.

Wenn man von den folgenden 4 Phasen im Ablauf eines empirischen Forschungsprojektes ausgeht:

1. **Theoretische Vorüberlegungen zum Projekt (Entwicklung von und Stoffsammlung zu den Fragestellungen, Analyse der potentiellen Objektdimensionen zum Zwecke der Hypothesenformulierung)**
2. **Entwicklung der Untersuchungsinstrumente (Auswahl der möglichen Techniken, Operationalisierung der Fragestellungen und Begriffe, Konstruktion des Erhebungsinstrumentes)**
3. **Vorbereitung und Durchführung der Datensammlung (z.B. Ziehung der Stichprobe, Durchführung und Kontrolle der Datenerhebung etc.)**

so wird sich im weiteren die Darstellung der Befragung als Methode im wesentlichen auf die beiden mittleren Phasen (also Entwicklung der Untersuchungsinstrumente, Vorbereitung und Durchführung der Erhebung) beschränken. Wenn hier also die überwiegend technische Durchführung der Befragungstechniken zur Diskussion steht, so sollte dadurch nicht der Eindruck entstehen, daß die theoretische Vorbereitung einer Erhebung, einen untergeordneten Stellenwert einnehme. Tatsächlich ist die theoretische Vorbereitung Grundvoraussetzung jeder empirischen Untersuchung.

4.1 Methode und Logik der Befragung

Die Befragung kann als eine besondere Art einer Kommunikation verstanden werden: Ein Kommunikator, nämlich der Forscher, verkehrt verbal mit vielen Kommunikatoren, nämlich den Befragten, wobei die einzelnen Fragen einer Befragung als Kommunikationsreize anzusehen sind. Somit determiniert die jeweilige Kommunikationssituation die Art und Weise der Befragung, wie die Befragung als erhebungstechnisches Instrument durch die allgemeinen Bedingungen der Kommunikation bestimmt wird.

In der empirischen Sozialforschung hat man sich bisher bemüht, mögliche Unterschiede in der Kommunikationssituation bei der wissenschaftlichen Befragung zu minimieren, um eine Standardisierung der Kommunikationssituationen zu erreichen, womit angestrebt wird, die verbalen Reaktionen der Befragten als gültige Antworten auf die gestellten Fragen (die vorgegebenen Reize) zu messen. Gerade durch diesen Standardisierungsversuch soll der wissenschaftliche, objektive, reliable und valide Charakter der Methode der Befragung erzielt werden. Weil nun rein äußerlich eine starke Ähnlichkeit zwischen der Befragung als Erhebungsinstrument und dem alltäglichen sozialen Verhalten zu bestehen scheint, fühlen sich viele dazu verführt, ohne methodologische und methodische Kenntnisse, eine Befragung durchzuführen. Es wird im weiteren zu zeigen sein, daß eine solche Fülle von Problemen mit der Befragung verbunden sind, daß Laien im Regelfalle nicht a priori geeignet sind, eine Befragung von den theoretischen Ausgangsüberlegungen bis hin zur Auswertung konzipieren zu können. Eine geradezu basale Differenz zwischen Befragung und Alltagskommunikation determiniert nämlich die methodischen Überlegungen zur Befragung so weitgehend, daß eine Applizierung der Erfahrungen der Alltagskommunikation auf die wissenschaftliche Befragungssituation durch Laien nicht vornehmbar erscheint.

Während Alltagskommunikationen im Regelfalle symmetrisch verlaufen, (Fragen und Antworten wechseln sich in einem permanenten Rollentausch ab) und jede verbale Reaktion ein Eingehen auf die verbale Aktion des Kommunikationspartners darstellt (sofern keine krankhaften Veränderungen vorliegen), *verläuft die Kommunikationssituation eines wissenschaftlichen Interviews absolut asymme-*

trisch (insbesondere bei hoher Standardisierung): Der Interviewer stellt nur Fragen, der Proband gibt nur Antworten. Der Interviewer reagiert nicht auf die Antworten des Befragten, sondern notiert diese „quasi unabhängig" von den Antworten des Befragten. Sofern man von Filterfragen und Gabelungsfragen absieht, stellt der Interviewer eine Frage nach der anderen, „so als ob keine Antworten gegeben worden wären". Diese für den Interviewer und den Befragten ungewohnte Problematik muß durch die Konstruktion des Erhebungsinstruments aufgefangen werden. Es genügt eben nicht, daß man ein paar Fragen formuliert, diese aneinanderreiht und damit eine wissenschaftlich abgesicherte Datenerhebung durchgeführt werden könnte. Vielmehr muß bewußt werden, daß das Interview als methodisch-wissenschaftlich eingesetztes Instrumentarium eine sehr spezielle, gewissermaßen unnatürliche Art der Kommunikation darstellt, die ebenso erlernt wie die asymmetrische Kommunikation mit ihren Problemen bei der Konstruktion des Erhebungsinstrumentes antizipiert werden muß (SCHEUCH).

4.1.1 Definition der Befragung

Sieht man von der Vielfalt der Verwendungsmöglichkeiten der Befragung ab und betrachtet man stattdessen die Befragung abstrakt als wissenschaftliches Instrument der Datenerhebung, dann bietet sich die Definition von SCHEUCH an: Unter Befragung versteht man „ein planmäßiges Vorgehen mit wissenschaftlicher Zielsetzung bei dem die Versuchsperson durch eine Reihe gezielter Fragen oder mitgeteilter Stimuli zu verbalen Informationen veranlaßt werden soll (SCHEUCH, S. 70).

Die Ähnlichkeit des Interviews mit dem Experiment scheint sich darin anzudeuten, daß auch bei der Befragung bestimmte Stimuli vorgegeben werden, die bestimmte Reaktionen hervorbringen sollen. Sie ist jedoch nur eine scheinbare. Während bei der Befragung allen Probanden die gleichen Stimuli vorgegeben werden, wurden die Reize beim Experiment variiert. Neben diesem gibt es noch einen weiteren methodologischen Unterschied: Während im Experiment eine unabhängige Variable variiert wurde und deren Wirkungen auf eine abhängige Variable festgestellt werden sollte, also die Relationen zwischen den Variablen Gegenstand der Untersuchung waren, dienen die bei der Befragung vorgegebenen Stimuli (Fragen) ausschließlich dazu, Variablen (seien sie im theoretischen Kontext unabhängig oder abhängig) zu erheben. Die Reize sind also nur Operationalisierungen von theoretischen Begriffen oder Variablen. Durch sie soll nicht gemessen werden, ob sie bestimmte Veränderungen provozieren; sie sollen die Variablen nur deskriptiv erfassen. Dies schließt jedoch nicht aus, daß relationale Hypothesen durch korrelative Kombination der einzelnen, deskriptiv erhobenen Variablen überprüft werden.

Zwischen Test und Befragung ergibt sich eine unmittelbare Verwandtschaft insoweit, als im Test das Instrument der Befragung eingesetzt werden kann und ihm auch in den meisten Fällen zugrundeliegt. Gleichwohl unterscheidet sich die Befragung von den Testmethoden dadurch, daß der Test eine hohe Standardisierung aufweist und die Absicht verfolgt, zu einem Routineverfahren entwickelt zu werden, das mehrfach in der gleichen, unveränderten Form angewandt werden kann.

Auch wenn Befragungen hochstandardisiert sein sollten, verbietet sich bei ihnen jedoch im Regelfalle die wiederholte Anwendung, weil sie meist auf augenblickliche Zustandssituationen abgestellt sind und eine generalisierende Anwendbarkeit nicht angestrebt wird. Auch in der Zielsetzung und der Zielverfolgung unterscheiden sich Experiment, Test und Befragung voneinander. Während das Experiment eine originär „kausale" Analyse darstellt, wo relationale Hypothesen überprüft werden und auf allgemeine Gesetzmäßigkeiten abgestellt wird, geht es beim Test darum, individuelle Daten als deskriptive Informationen zu erhalten. *Die Befragung ist nun eine Methode, die (in der Regel) auf akkumulierte Häufigkeitsdaten abstellt, deskriptive Informationen erhebt, die aber durch korrelative Analyse auch zur Überprüfung relationaler Hypothesen herangezogen werden können.*

Die oben gegebene Definition der Befragung täuscht einen Grad der Einheitlichkeit vor, der für diese Methode geradezu atypisch ist: Die Befragung erscheint realiter in so vielen Formen und Modifikationen, daß es schwerfällt, die einzelnen Befragungstypen unter die obige Definition zu subsumieren. So gibt es eine Fülle von Einteilungsgesichtspunkten, die untereinander kombiniert, eine Vielzahl von spezifischen Befragungsformen ergeben. Einige dieser Einteilungsgesichtspunkte seien in den folgenden Abschnitten erarbeitet.

4.1.2 Die Standardisierung von Reizen

Eine wissenschaftliche Befragung kann durchaus in der Form einer Alltagskommunikation durchgeführt werden, wobei Fragen und Antworten sich gegenseitig bedingen und ein mehr oder weniger symmetrischer, bilateraler, gleichgewichtiger Gesprächsablauf zu verzeichnen ist. Die Befragung kann aber auch dann, wenn es um die Erzielung von quantitativen Ergebnissen geht, äußerst asymmetrisch konstituiert sein, indem die Fragen als vorzugebende Reize voll vorgeschrieben sind und sozusagen unabhängig von den Antworten des Probanden ablaufen. Berücksichtigt man ausschließlich den Gesichtspunkt, inwieweit eine Befragung standardisiert ist, so lassen sich im wesentlichen 3 Typen von Befragungen herauskristallisieren:

Bei der standardisierten Befragung wird ein detailliert ausgearbeiteter Fragebogen verwendet, in dem sowohl die Formulierung der einzelnen Fragen, wie auch die Reihenfolge der Fragen fixiert ist. Jedes Abweichen davon ist unzulässig und würde die Reizstandardisierung und damit die Akkumulierung der Daten und die Generalisierung der Dateninterpretation problematisch machen. Solche standardisierten Befragungen werden auch als *strukturierte oder gelenkte Befragungen* bezeichnet.

Da die Standardisierung voraussetzt, daß die mittels der Befragung zu erhebenden Sachverhalte vorher voll bekannt sein müssen (denn es kann nur das für alle Probanden erhoben werden, was als theoretisch wichtig und relevant erkannt wurde), kann eine standardisierte Befragung nur dort eingesetzt werden, wo ein Problemkreis bereits durch mehrere Untersuchungen oder Voruntersuchungen (= *Pretests*)

abgeklärt wurde und man die relevanten Dimensionen des Problems, das bearbeitet werden soll, zu kennen glaubt. Erst unter dieser Voraussetzung ist ja die Operationalisierung von Begriffen und Variablen möglich. Da bei der standardisierten Befragung eine optimale Vergleichbarkeit sowohl der Reizsituation, wie auch der Reaktionen auf diese Reize (Antworten auf die Fragen) gegeben ist, sind standardisierte Befragungen (schon auch aus Kostengründen), immer dort einzusetzen, wo umfangreicheres Datenmaterial gewonnen und aufbereitet werden soll.

Der standardisierten Befragung als Extremtypus steht auf der anderen Seite die *nichtstandardisierte, (freie, ungelenkte, qualitative oder unstrukturierte) Befragung* gegenüber. *Sie wird ohne Fragebogen oder festes Frageschema durchgeführt. Weder die Formulierung der einzelnen Fragen, noch der gesamte Ablauf der Befragung ist vorab festgelegt.* Ausschließlich die Kenntnis von Absicht und Zweck der Befragung sind notwendige Vorbedingung für deren Durchführung. In der Regel wird dem Befragten ein Rahmenthema vorgegeben, über das man sich mit dem Interviewer frei unterhält, wobei letzterer lediglich durch Zwischenfragen weiterhilft, zur Präzisierung auffordert usw. Die nichtstandardisierte Befragung erfüllt eher *explorative Funktionen*. Diese Form der Befragung erlaubt es, mögliche Dimensionen, Problemfelder und Hypothesen des Untersuchungsgegenstandes zu ergründen. Man versucht dabei noch unbewußte Motivationen, Einstellungen, Vorstellungen etc. zu erfahren, wobei darauf geachtet wird, daß der Interviewer unabhängig sondieren kann und nicht an der Oberfläche einer, durch eine standardisierte Frage verursachten Reizsituation verweilt. Diese Informationsammlungsfunktion der unstrukturierten Befragung kann natürlich nur dann fruchtbar sein, wenn der Interviewer Geschicklichkeit und Sachkompetenz entwickelt. Geschicklichkeit setzt lange Erfahrung im Interviewen voraus und Sachkompetenz erfordert die genaue Kenntnis des Forschungszieles und der theoretisch-hypothetischen Überlegungen hierzu. Wegen dieser hohen Qualifikationsanforderungen werden die nichtstandardisierten Interviews häufig von den Forschern selbst durchgeführt. Die unstrukturierten Interviews sind meistens auch Vorstufen für standardisierte Befragungen, die quantifizierende Ziele verfolgen.

Zwischen diesen beiden Extremtypen läßt sich die *halbstandardisierte Befragung* lokalisieren. *Hier wird durch den Forscher ein „Interviewerleitfaden", also ein festes Frageschema vorgegeben, es wird jedoch dem Interviewer überlassen, Reihenfolge und Formulierung der Fragen selbst zu bestimmen.* Gegebenenfalls können auch Zusatzfragen gestellt werden, die dem Interviewer im individuellen Fall besonders nutzbringend erscheinen (so können Nachfragen zur Präzisierung unscharfer Antworten, zum Zwecke einer tiefergehenden Analyse etc. gestellt werden). Dieser Technik der halbstandardisierten Befragung bedient sich insbesondere das *Intensiv- oder Tiefeninterview*, dessen Vorteil in der hohen Flexibilität und Anpassung des Vorgehens an die einzelne Persönlichkeit des Befragten und an die jeweilige Befragungssituation besteht. Vergleichbarkeit und Quantifizierbarkeit der Ergebnisse sind jedoch bei der halbstandardisierten und der unstrukturierten Befragung äußerst problematisch. Quantifizierungen erfordern meist eine hohe Standardisierung.

Während eine standardisierte Befragung praktisch sowohl schriftlich wie auch mündlich durchgeführt werden kann, (vgl. Abschn. 4.1.3.) ergibt sich aus dem Wesen der nichtstandardisierten und der halbstandardisierten Befragung, daß diese nur in der mündlichen Form praktiziert werden können. Methodologisch gesehen, ergeben sich zwischen standardisierter und nichtstandardisierter Befragung auch im Hinblick auf sprachsoziologische Probleme Unterschiede: Während die standardisierte Befragung versucht, die Problematik von *Sprachbarrieren und unterschiedlicher Ausdrucksfähigkeit* der Probanden durch leicht verständliche Fragen mit im Regelfall vorgegebenen Antwortkategorien zu lösen, berücksichtig das nichtstandardisierte Interview eher die Problematik der *Bedeutungsäquivalenz*, wonach für gleiche Bedeutungen durchaus unterschiedliche Reize und Bedingungen angegeben werden können (vgl. hierzu 1.3.).

In der Literatur ist für standardisierte und nichtstandardisierte Befragungen eine Fülle von Vor- und Nachteile aufgezählt worden. Eine Gegenüberstellung der wichtigsten Vorteile der einzelnen Befragungsformen sollte als positive Ausgrenzung und Anwendungsmöglichkeit genügen:

1. Standardisierte Befragung
   - Vergleichbarkeit der Antworten
   - Höhere Zuverlässigkeit
   - Reduktion von Fehlern beim Wortlaut von Fragen auf ein Minimum (z.B. unterschiedliche Fragestellung, Reihenfolge der Fragen etc.)
   - Einfachere Durchführung des Interviews
   - Schnelle und preiswerte Analyse der Antworten

2. Nichtstandardisierte Befragung
   - Eher Standardisierung von Bedeutungen als die der oberflächlichen Aspekte der Reizsituation (Bedeutungsäquivalenz der Fragen)
   - Ermutigung zu lebensnäheren Antworten, da der alltäglichen Gesprächssituation mehr angepaßt
   - Flexibler in der Durchführung

Von der jeweiligen Vorteilen Befragungsformen läßt sich auf deren unterschiedliche Funktionen schließen, die schon implizit angedeutet wurden:
Das standardisierte Interview dient vorwiegend der Messung von relevanten Merkmalen (zumeist im Endstadium einer Untersuchung). Das nichtstandardisierte Interview verfolgt im besonderen die Exploration von Sachverhalten und Bezugssystemen der Befragten im Anfangsstadium einer Untersuchung. Prinzipiell läßt sich jedoch nicht sagen, daß die eine oder andere Technik die bessere sei. Vielmehr sind deren Anwendungsbereiche verschieden und der Einsatz der Befragungsform wird entscheidend durch Absicht, Untersuchungsziel und Objektbereich des Forschers im jeweiligen Einzelfalle bestimmt.

4.1.3 Die Präsentierung von Reizen

Neben der Standardisierung der Reize als Einteilungsgesichtspunkt für verschiedene Befragungsformen, spielt die Präsentierung der Reize in einer Befragung, eine theoretisch wie praktisch wichtige Rolle. Fragt man danach, in welcher Kommunikationsform die entsprechenden Reize präsentiert werden, so kann man im Hinblick auf diese Kommunikationssituation *schriftliche oder mündliche Darbietun-*

gen der Fragen unterscheiden. Dies ist eine Differenzierung instrumenteller Art.

Bei der mündlichen Befragung geschieht die Stimulusübermittlung und die Registrierung der Reaktionen des Probanden (Antworten) über den Interviewer. Dieses Vermittlungsinstrument zwischen Forscher und Befragunsperson fällt bei der schriftlichen Befragung weg, weil dort der Befragte den Fragebogen selbständig ausfüllt. Bei der schriftlichen Befragung können zwei Hauptformen unterschieden werden: die sog. *postalische Befragung und die paper-and-pencil-Methode.* Bei der postalischen Befragung werden die Probanden angeschrieben und mit Hilfe eines Begleitschreibens wird ihnen der Zweck der Untersuchung erläutert; der Fragebogen soll durch den Probanden selbständig ausgefüllt und mittels beiliegendem Freiumschlag an den Umfrageträger zurückgeschickt werden. Bei der paper-and-pencil-Methode handelt es sich meist um sog. *Klassenzimmer-Interviews*; eine bestimmte Anzahl von Probanden sind zur gleichen Zeit am selben Ort versammelt, wo ihnen der Fragebogen überreicht und mündlich durch den Untersuchungsleiter Sinn und Zweck der Befragungsaktion mitgeteilt wird. Auch hier füllen die Probanden den Fragebogen selbständig aus, jedoch steht der Untersuchungsleiter für eventuell auftretende Fragen zur Verfügung.

Wegen der Bedeutsamkeit der Differezierung zwischen schriftlicher und mündlicher Befragung (Interview), seien die in der Literatur hierzu genannten jeweiligen Vor- und Nachteile angeführt. Deren gedanklicher Nachvollzug im Versuch ihrer Begründung sollte auch ohne zusätzliche Erläuterungen möglich sein. (Vgl. hierzu H.J. RICHTER, Die Strategie schriftlicher Massenbefragungen, Bad Harzburg 1970).

Vorteile der schriftlichen Befragung

- Geringere Kosten
- Geringer Zeitaufwand und schnellere Durchführung.
- Befragung geographisch weitstreuender Personenkreise leichter möglich.
- Interviewereinflüsse werden ausgeschaltet (Vgl. hierzu Abschn. 4.1.7.).
- Der Befragte hat mehr zeit für jede Frage und kann daher die Fragen besser durchdenken (dies kann allerdings auch zum Nachteil gereichen).
- Größere Anonymität, daher ehrlichere (= gültigere) Antworten.

Nachteile der schriftlichen Befragung

- Es besteht keine Kontrolle der Befragungssituation (die Identität der Befragungssituation ist eine entscheidende Voraussetzung für die Gleichheit der Stimuliwirkung; so ist z.B. auch nicht kontrolliert, ob tatsächlich die gemeinte Befragungsperson den Fragebogen selbst ausfüllt).
- Es kann praktisch keine Befragungshilfe bei schwierigen Fragen geben werden, weshalb man auf solche Fragen von vorne herein verzichten sollte.
- Durch die fehlende physische Anwesenheit des Interviewers besteht eine zu geringe Stimulierung. Antwortbereitschaft und Motivation sind häufig zu gering, sodaß sich geringe und nur sehr niedrige Rücklaufquoten erzielen lassen
- Aus den genannten Nachteilen (insbesondere dem Zuletztgenannten) ergeben sich z.T. unzureichende Gültigkeit und Zuverlässigkeit der Ergebnisse, weil die Art der Ausfälle (systematisch oder zufällig) nicht bekannt ist.

Vorteile der mündlichen Befragung

- Es können schwierigere Fragen erhoben werden.

- Der Fragebogen kann länger sein als bei der schriftlichen Befragung.
- Es sind höhere Stichprobenausschöpfungen möglich als bei der schriftlichen Befragung (aufgrund neuerer Erkenntnisse ist es jedoch auch bei der schriftlichen Befragung möglich, die Rücklaufquoten entscheidend zu steigern).
- Es können im Gegensatz zur schriftlichen Befragung Fragen erhoben werden, die Spontanreaktionen erfordern.

Nachteile der mündlichen Befragung

- Verzerrungseffekte durch mangelnde Neutralität des Interviewers und Anonymität des Befragten (vgl. hierzu Abschn. 4.1.7.)
- Hohe Durchführungskosten und erheblicher organisatorischer Aufwand
- Keine genaue Kontrolle der Befragungssituation (z.B. Anwesenheit von Drittpersonen, fingierte Interviews, Lerneffekt des Interviewers und ähnliche Verzerrungsfehler).

Aus diesen Vor- und Nachteilen scheint sich insgesamt ableiten zu lassen, daß die Anwendungsbreite der schriftlichen Befragungen beschränkter ist. Insbesondere auch die geringen Rücklaufquoten waren bisher mit dafür ausschlaggebend, daß die kommerziellen Institute nach wie vor mit mündlichen Befragungen – allerdings meist mit Quotaverfahren und keinen reinen Zufallsauswahlen – arbeiten. Es konnte jedoch gezeigt werden (vgl. hierzu RICHTER H.J., Die Strategie schriftlicher Massenbefragungen), daß es durchaus möglich ist, diesen Nachteil nicht zur Geltung kommen zu lassen und äußerst hohe Rücklaufquoten zu erzielen, wenn die schriftliche Befragung auf relativ *homogene Untersuchungspopulationen* angewandt wird, sodaß in einem zu verfassenden Begleitschreiben, eine hohe Antwortbereitschaft stimuliert werden kann, weil der Appell zu Mitwirkung wegen der Homogenität sehr spezifisch und wirksam konstruiert werden kann (z.B. eine Befragung bei bestimmten Berufsgruppen). Letztendlich wird man mit schriftlichen Befragungen immer dann arbeiten müssen, wenn keine andere Methode, als die der Befragung möglich erscheint, aus Zeit und Kostengründen aber Interviews ausgeschlossen sind.

Befragungen können auch im Hinblick auf die *Struktur der Rezipienten* klassifiziert werden. Die Struktur kann als *Einzel- oder als Gruppenbefragung* charakterisiert werden. Die schon erwähnten paper-and-pencil-Befragungen in Klassenzimmern sind typische Gruppenbefragungen, wobei deren Vorteil darin besteht, daß die äußere Situation für jeden Probanden konstant ist. Gleichwohl beantwortet jeder einzelne für sich und unbeeinflußt von seiner personellen Umgebung die gestellten Fragen. Neben der Tatsache der Konstanz der äußeren Situation, die als Argument für den Einsatz der Gruppenbefragung herangezogen wird, spielen die reduzierten Kosten dieser Befragunsform für deren Einsatz eine entscheidende Rolle. Trotzdem überwiegen quantitativ die Einzelbefragungen, was daraus resultiert, daß die Untersuchungseinheiten in der Regel Individuen sind, die repräsentativ ausgewählt werden und somit nicht ohne z.T. erhebliche Schwierigkeiten in einer Gruppe lokal und zeitlich zusammengefaßt werden können.

Befragungen können sich auch im *Stil der Reizpräsentierung* unterscheiden. Mit den unterschiedlichen Stilen werden dabei verschiedene Taktiken des Vorgehens diskutiert. Beim sog. *harten Interview* geht man davon aus, daß der Befragte nur widerstrebend antworten wird und oft durch unrichtige Antworten ausweicht.

Der Fragebogen wird nun so konstruiert, daß der Interviewer gleichsam als Autorität, wie in einem Verhör, sehr massiv vorgeht, um Antworten zu erhalten.

Beim *weichen Interview* geht der Forscher von ähnlichen Annahmen über die Befragten aus, versucht jedoch das sympathisierende Verständnis des Interviewers für die spezielle Situation der Befragten zum Ausdruck zu bringen, um dadurch die widerstrebende Haltung des Befragten abzubauen.

Die *neutrale Vorgehensweise* schließlich ist nicht nur jene, die am häufigsten eingesetzt wird, sondern sie ist auch die einzige, der es gelingt, die Probleme von Gültigkeit und Zuverlässigkeit in den Griff zu bekommen. Nur bei einer neutralen Vorgehensweise ist die intersubjektive Nachprüfbarkeit der Daten gewährleistet. Bei dem harten oder weichen Interview ist der Einfluß des Interviewers nicht zu kontrollieren.

4.1.4 Die Intentionen von Befragungen

Mit der Durchführung von Befragungen können unterschiedliche Absichten verbunden werden. Einleitend wurde als Ziel von Befragungen die Informationsgewinnung über bestimmte, ausgewählte Objektbereiche formuliert. Diese Absicht liegt den meisten Interviews zugrunde. Die potentiellen Befragungspersonen werden hier als Informanden perzipiert. Eine, einem solchen Zweck dienende Befragung wird als *ermittelnde Befragung* bezeichnet (vgl. hierzu van KOOLWIJK, J. Die Befragungsmethode, in: van KOOLWIJK, J., und WECKEN-MAYSER, M. (Hrsg.), Techniken der empirischen Sozialforschung, Bd. 4, Erhebungsmethoden: Die Befragung, München 1974, S. 15f.) Hiervon setzt sich das *vermittelnde Interview* ab, nach dem die Probanden einerseits zwar Informationen liefern, aber andererseits durch die Befragung die bewußte und gezielte Intention einer Veränderung der Befragungspersonen bzw. des Objektbereichs realisiert werden soll. Solche Formen der Sozialforschung werden in der Literatur als *Aktions- oder Handlungsforschung* bezeichnet.

Das ermittelnde Interview läßt sich im Hinblick auf die Absicht der Befragung wiederum in drei Formen klassifizieren: das informatorische, das analytische und das diagnostische Interview. Während bei dem *informatorischen Interview* die Befragung dazu dient, bestimmte ausgewählte Tatsachen als solche festzustellen, um sie deskriptiv beschreiben zu können, geht die *analytische Befragung* darüber hinaus, indem durch sie versucht wird, die theoretischen Überlegungen (meist in Form relationaler Hypothesen) an der Realität zu überprüfen. Beim informatorischen Interview wird die Befragungsperson sozusagen als Experte angesehen, der über bestimmte Sachverhalte Informationen weitergibt. Beim analytischen Interview ist die Befragungsperson selbst Untersuchungsgegenstand und es sollen über die Akkumulation der individuellen Informationen, generelle Sachverhalte überprüft und allgemeine Aussagen vorgenommen werden. Es versteht sich von selbst, daß das analytische Interview notwendigerweise als Vorstufe das informatorische Interview mit einschließt, denn schließlich müssen vor jeder Analyse von miteinander in

Beziehung stehenden Sachverhalten diese selbst deskriptiv erhoben werden. Die *diagnostische Befragung* ist die dritte Form des ermittelnden Interviews. Beim diagnostischen Interview wird ein Set von Merkmalen und Merkmalsausprägungen deskriptiv und individuell erhoben, um Aussagen über den individuellen Zustand der Befragungsperson und eventuelle Behandlungsmöglichkeiten vornehmen zu können. Das diagnostische Interview setzt dabei natürlich voraus, daß die zu erhebenden Sachverhalte theoretisch ausreichend abgeklärt, bekannt und in konsistente und bereits überprüfte Theorien Eingang gefunden haben, sodaß unter Zuhilfenahme der mittels diagnostischem Interview gewonnenen individuellen Informationen und der zugrundeliegenden wissenschaftlichen Theorie eine gültige Diagnose vorgenommen werden kann, die in eine Therapie technologisch transformiert werden kann. So gehört z.B. das systematische, standardisierte aber auch das völlig unstrukturierte Abfragen von bestimmten Krankheitssymptomen dazu, bestimmte Diagnosen stellen, um daraus resultierende Therapien einleiten zu können.

Aus der knappen Differenzierung der drei Formen der ermittelnden Befragung, läßt sich unschwer ableiten, daß es sich hierbei um idealtypische handelt. Realiter wird es nicht immer gelingen, diese drei Typen streng voneinander zu trennen. So enthält das diagnostische Interview notwendigerweise Elemente der informatorischen, wie auch solche, der analytischen Befragung.

Zu den vermittelnden Interview rechnet man insbesondere die *therapeutischen Interviews*, obgleich auch sie zunächst einmal ermitteln müssen. Therapeutische Interviews gehen von der Überlegung aus, daß durch die Durchführung einer Befragung durch einen entsprechend qualitativ vorbelasteten Interviewer bestimmte therapeutische Prozesse in Gang gesetzt werden. So kann die Visite im Krankenhaus, die persönliche Zuwendung des Arztes zum Patienten, das Erkundigen nach dem persönlichen Befinden etc. als unstrukturiertes Interview mit partiell therapeutischer Funktion verstanden werden. Einige Behandlungsformen psychischer Erkrankungen machen sich die Erkenntnis zunutze, daß das Bewußtwerden bestimmter Sachverhalte bereits einen ersten Schritt zum Erkennen und Lösen der spezifischen Probleme darstellt. So kann das Gespräch zwischen Psychoanalytiker, Psychiater oder Psychotherapeut und Patienten erste therapeutische Funktionen übernehmen.

Eine weitere Unterscheidungsmöglichkeit von Befragungsformen kann in der Subdimension der *Interpretationsabsicht* gesehen werden: Man differenziert zwischen *objektbezogenen, indikatororientierten oder inhaltsbezogenen* auf der einen und *personenbezogenen, subjektzentierten oder psychologischen Befragungen* auf der anderen Seite (vgl. KRAPP, A., PRELL, S., Studienhefte zur Erziehungswissenschaft, Heft 5, Empirische Forschungsmethoden, München 1975, S. 119f). Die schon bekannten Testverfahren gehören zur zweitgenannten Version von Befragungen, denn bei ihnen ging es darum, aus den Antworten der Befragungspersonen bestimmte Merkmale dieser Personen festzustellen. Bei den objektbezogenen Befragungen dagegen wird versucht, aus den Antworten der Probanden auf infrage-

stehende Sachverhalte zu schließen: Die Frage des Kinderarztes nach bestimmten Krankheitssymptomen, die an die Mutter gerichtet ist, dient keineswegs dazu, die Beobachtungsfähigkeit der Mutter zu testen. Vielmehr soll sie Aufschluß über die Symptome bei dem nicht selbst befragten Kind liefern. (Objektorientierte Befragung) Die Durchführung eines Intelligenztests jedoch, dient der Ermittlung eines individuellen Merkmals und ist somit personenzentriert.

Beabsichtigt der Forscher bestimmte *Veränderungen an seinem Objektbereich im Zeitablauf* festzustellen, so hat er unter Einsatz der Befragungsmethoden zwei Möglichkeiten: Er zieht zum Zeitpunkt t_1 eine Stichprobe, die für die ihn interessierende Bevölkerungsgruppe repräsentativ ist und befragt diese. Zu einem Zeitpunkt t_2 wird erneut eine repräsentativ Stichprobe gewonnen (die sich aber von der ersten unterscheidet), und dieser der gleiche Fragebogen vorgelegt. Aus der Differenz der gegebenen Antworten lassen sich dann quantitative Veränderungen angeben, wobei sich diese Verändrungen aber ausschließlich auf die kumulierten Daten beziehen. Über Veränderungen bei den Einzelpersonen kann, wegen der unterschiedlichen Stichproben, keine Aussage gemacht werden. (*Trendbefragung*)

Das *Panel-Verfahren* versucht individuelle Veränderungen (und mit deren Aggregation natürlich auch globale Veränderungen) an einer bestimmten Population festzustellen, d.h. *ein und derselben Population werden sowohl zum Zeitpunkt t_1, als auch zum Zeitpunkt t_2 dieselben Fragen gestellt;* aus den unterschiedlichen Antworten lassen sich die Veränderungsprozesse ablesen. Diese einfache Form des Panels kann natürlich über mehrere Befragungen hinweg fortgesetzt werden. Allerdings ist dabei zu bedenken, daß die sog. *Panel-Sterblichkeit* auftritt, d.h., daß nicht zu jedem Befragungszeitpunkt immer wieder dieselben Personen befragt werden können, weil manche z.B. verstorben, verzogen, nicht erreichbar etc. sind. Man versucht dann die ausgefallenen Probanden durch solche zu ersetzen, die ähnliche Merkmalskonstellationen aufweisen. Für die Auswertung der Panel-Verfahren ist zu berücksichtigen, daß nur bestimmte statistische Verfahren für sie angemessen sind, weil es sich um korrelierende, abhängige Stichproben handelt.

4.1.5 Die Frageformulierung

Bevor einige Grundregeln der Frageformulierungen und einige typische Frageformen besprochen werden, sollte vor einem häufig gemachten grundlegenden Irrtum gewarnt werden: Allzu häufig findet man, wenn Laien eine Fragebogenaktion durchführen, eine Vermengung der *Programm- oder Beweisfragen mit den Ermittlungs- oder Interviewfragen.* (vgl. ROEDE, H., Befrager und Befragte, Berlin 1965, S. 18) Da am Beginn einer empirischen Untersuchung immer theoretische Hypothesen stehen, ist der Laie allzu sehr geneigt, seine theoretischen Überlegungen unmittelbar als Erhebungsfragen zu formulieren. Nun gibt es aber oft keine direkte Beziehung zwischen den Programm- oder Beweisfragen und den im Fragebogen zu stellenden Fragen. (So könnte z.B. eine Beweisfrage die Absicht haben, festzustellen, ob eine Person neurotisch ist. Es dürfte unmittelbar einsichtig sein, daß die direkte Erhebungsfrage: „Sind Sie neurotisch" unsinnig ist und keinerlei Erkenntnis-

wert besitzt.). Es geht also bei der Frageformulierung darum, eine theoretisch relevante Beweisfrage in eine erhebungstechnisch und psychologisch, wie auch methodologisch sinnvolle Fragestellung zu transformieren (vgl. hierzu insbesondere NOELLE/NEUMANN, E., Umfragen in der Massengesellschaft, Reinbek 1963, S. 54f).

> Die Programmfragen, „auf die es ankommt, sind oft im im ganzen Bogen nicht zu finden. Dafür aber lange Frageserien, deren Sinn nicht in wenigen Worten zu erklären ist... Zugespitzt: je besser ein Fragebogenentwurf ist, desto weniger überzeugend wirkt er auf den Leser. Die Überzeugungskraft sollte man daher nicht vom Fragebogen mit seinen Testfragen erwarten, sondern von der späteren Analyse, dem Ergebnisbericht... . Indem dies betont wird, wird vielleicht die Zahl der naiven Fragen eingedämmt, in denen plump die Programmfragen ohne jede Übersetzung aneinandergereiht sind. In keiner Phase einer statistisch-repräsentativen Erhebung drängen sich die Laien so eifrig zur Mitwirkung vor, wie beim „Schmieden" des Fragebogens. Und doch ist gerade dies die empfindlichste, komplizierteste Phase, von deren Bearbeitung vor allem anderen Niveau und Ertrag einer Erhebung abhängen" (NOELLE/NEUMANN, S. 59).

Daraus lassen sich zwei grundsätzliche Sachverhalte ableiten:

1. Erst wenn es gelingt, eine Programmfrage in eine gültige und zuverlässige Erhebungsfrage zu transformieren, was erfahrungsgemäß mit zu den schwierigsten Aufgaben bei der Befragung als erhebungstechnischer Methode gehört, kann angenommen werden, daß die Erhebungsergebnisse als gültige und zuverlässige Daten gelten können.

2. Häufig wird es unmöglich sein, eine Hypothese in nur eine Fragestellung zu kleiden. Vielmehr gilt, daß je komplexer ein Problem und eine Hypothese sind, desto mehr Fragen sind für deren Operationalisierung erforderlich.

Grundregeln der Frageformulierung

Die Formulierung von Fragen ist auch heute noch der Abschnitt einer Befragung, in dem persönliches Geschick und Erfahrung, und weniger methodische Schulung entscheidend sind (SCHEUCH, E.K., Das Interview in der Sozialforschung, S. 141). Dies muß unbestritten bleiben, weil die bisher vorliegenden theoretischen Ansätze zur Frageformulierung und Fragebogenkonstruktion eher auf Erfahrungswerten, denn auf theoretisch abgesicherten Aussagen basieren. Aus solchen Erfahrungswerten heraus kann man einige Faustregeln der Frageformulierung angeben, deren Kenntnis hilft, gravierende Fehler zu vermeiden. Nach wie vor gibt es jedoch keine „Gesetze" darüber, wie eine optimale Operationalisierung eines theoretischen Begriffes in einem Fragebogen auszusehen hat. (Zwar hat HOLM eine Theorie der Frage und eine solche der Fragebatterie aufgestellt, die auf faktorenanalytischen Überlegungen aufbauen, jedoch sind auch diese fortschrittlichen Entwicklungen noch nicht ausreichend, um die Gültigkeitsprobleme der Operationalisierungen voll in den Griff zu bekommen. (vgl. hierzu HOLM, K., Theorie der Frage, bzw.

Theorie der Fragebatterie in: Kölner-Zeitschrift für Soziologie und Sozialpsychologie Heft 1 und 2 1974). Als erste Regel soll festgehalten werden, daß *die Fragen grundsätzlich so einfach formuliert sein sollen, wie es gerade noch mit dem sachlichen Zweck der Fragestellung vereinbar ist,* d.h., daß die Fragen kurz sein sollen, daß grammatikalisch schwierige Konstruktionen vermieden werden und daß sie auf den Bezugsrahmen des Befragten bezogen sein müssen, damit dieser sich schon beim ersten Anhören der Frage, auf dieselbe einstellen kann. Dabei geht es sicherlich nicht darum, wie NOELLE eindrucksvoll zeigt, die Fragen druckreif zu formulieren. Vielmehr kann es sogar erforderlich werden, eine bewußt fehlerhafte Grammatik in der Formulierung zu wählen, um die Verständlichkeit einer Frage zu erhöhen (NOELLE, S. 71). Diese einfachste aller Regeln wird häufig vernachlässigt, insbesondere, wenn man davon ausgeht, daß die zu befragende Population ein relativ hohes intellektuelles Niveau aufweist. Tatsächlich jedoch zeigen Untersuchungen, daß gerade auch bei diesem Personenkreis äußerster Wert (z.B. schon aus Zeitgründen) darauf gelegt wird, daß auch komplizierte Sachverhalte in relativ einfache Formulierungen gefaßt werden.

Die Fragen sollten zweitens so formuliert werden, daß sie möglichst von allen Befragten nur in einem, und zwar in dem von dem Forscher gemeinten Sinne verstanden werden. Denn wenn gleiche Frageformulierungen in unterschiedlichem Sinne von den Befragten aufgefaßt werden können, dann sind die gegebenen Antworten nicht mehr vergleichbar, weil sie auf unterschiedlich verstandene Fragen bezugnehmen.

FRIEDRICHS (Jürgen FRIEDRICHS, Methoden der empirischen Sozialforschung, Reinbek 1973. S. 196) gibt für mehrdeutige und mehrdimensionale Fragen ein sehr plastisches Beispiel: So kann man auf die Frage „Was haben Sie für Nachbarn?" in den Dimensionen Schichtzugehörigkeit, Einkommen, Religion, Sympathie und Alter geantwortet werden: Als Reaktionen hätte man erhalten können: „Hier wohnen Arbeiter wie wir". „Meistens reiche Leute". „Die meisten sind Katholiken". „Ich finde, recht nette Leute". „Junge Ehepaare mit vielen Kindern" usw. Daß diese Antworten nicht mehr miteinander vergleichbar sind, ist offenkundig.

Da aber insbesondere in standardisierten Befragungen kaum die Möglichkeit besteht, nachzufragen, Sondierungen vorzunehmen, bzw. Präzisierungen der Antworten zu fordern, dürfen die Fragen auf keinen Fall mehrdimensional gestellt sein, will man klare, eindeutige und miteinander vergleichbare Antworten erhalten.

Als dritte Regel wäre zu nennen, *daß auf alle Fälle eine intellektuelle Überforderung des Befragten vermieden werden muß.* So wird mangelnde Eindeutigkeit der Fragen, bzw. der Antworten nicht nur durch unbekannte Termini oder schwierige Formulierungen bewirkt. Häufig kommt es vor, daß zwar die Frage durchaus verstanden wird, daß aber die Fragestellung so schwierig und so komplex erscheint, daß der Proband darauf keine vernünftige Antwort geben kann. Hypothetische Fragen gehören zu solchen, den Befragten meist überfordernden Fragestellungen. Auch Erinnerungsleistungen zu Sachverhalten, die für den einzelnen eine relativ untergeordnete Bedeutung haben, überfordern die Befragungsperson. (So wird kaum jemand in der Lage sein, die Frage nach der Häufigkeit der Kinobesuche im letzten Jahr konkret und exakt beantworten können.)

Mit den angegebenen drei Regeln ist man freilich noch nicht in der Lage, die Erhebungsfragen eines Fragebogens in jedem Falle als gültige Operationalisierungen für ein theoretisches Phänomen zu formulieren. Drei zugegebenermaßen eklektizistisch herausgegriffene aber besonders wichtige Probleme sollen diese Aussage verdeutlichen helfen:

Die *Bedeutungsäquivalenz von Fragen* ist zwar theoretisch längst nicht ausreichend abgeklärt, doch durch einige empirische Erfahrungen ausreichend abgesichert. Mit der Standardisierung einer Frage ist noch lange nicht ihre Bedeutung für alle Befragten standardisiert. Bei den meisten Befragungen hat man es nämlich mit verschiedenen Sprachtypen von Befragten zu tun, d.h., mit sprachlich heterogenen Befragungspopulationen. Kann man nun keinen einheitlichen Sprachtypus bei den Befragten voraussetzen, dann ist sicherlich vorstellbar, daß ein und dasselbe Wort für zwei Befragte von inhaltlich unterschiedlicher Bedeutung ist, bzw. zwei verschiedene Worte für Befragungspersonen gleiche inhaltliche Bedeutung besitzen. Es wäre daher in jedem Einzelfalle zu entscheiden, ob verschiedene aber bedeutungsgleiche Wörter benutzt werden sollen, oder ob nicht vielmehr zwei oder gar mehrere Versionen einer Frage für verschiedene Typen von Befragten formuliert werden sollen. Für diesen Sachverhalt oft zitiert sind die Befragungen von KINSEY zum sexuellen Verhalten. KINSEY konnte in der Oberschicht durchaus nach der Häufigkeit des „Koitus" fragen, während er in der Unterschicht von „Geschlechtsverkehr" oder von „Ficken" sprechen mußte. Er hat also offensichtlich versucht die *Bedeutungen zu standardisieren,* indem er unterschiedliche Formulierungen zum selben Sachverhalt wählte.

Gerade bei hochstandardisierten Befragungen einer Vielzahl von Personen wird sich jedoch eine solche unterschiedliche Formulierung als Standardisierung der Bedeutungsäquivalenz relativ schwer realisieren lassen. In solchen Fällen kann die Frageformulierung in dem Sprachtypus der niedrigsten Schicht vorgenommen werden, denn dieser wird im Regelfalle auch von den höheren Schichten verstanden (dies stellt im übrigen mit ein Argument für möglichst einfache Frageformulierung dar). Berücksichtigt man diese letzte Regel, so wird man annehmen dürfen, daß die Fehlerwahrscheinlichkeit bei einer Standardisierung der Frageformulierung in Bezug auf die Bedeutungsäquivalenz für die verschiedenen Befragungspersonen doch relativ gering ist.

Ein weiteres theoretisches Problem der Frageformulierung, das empirisch in seinen Auswirkungen nur schwer abschätzbar ist, stellen die *Suggestivfragen* dar. Bei einer suggestiven Frage wird dem Befragten die Antwort in den Mund gelegt, d.h. *die Frageformulierung beeinflußt die zu gebende Antwort in einer ganz bestimmten Richtung.* Solche Beeinflussungen sind natürlich im Sinne einer objektiven Faktenerhebung abzulehnen und können leider im konkreten Falle einer bestimmten Frage nicht quantitativ abgeschätzt werden. So ist es relativ einfach, Definitionen darüber zu liefern, wann Fragen neutral, oder wann sie suggestiv sind. Es ist jedoch nicht anzugeben, wann eine konkrete Frage in welchem Ausmaße suggestiv wirkt. Man wird zwar mit bestimmten Erfahrungswerten argumentieren und eine

suggestive Wirkung aufzeigen können, jedoch in den seltensten Fällen in der Lage sein, diese Wirkung zu quantifizieren. Nur bei einigen methodologisch orientierten Untersuchungen, wo verschiedene Frageformulierungen zum selben Gegenstand auf deren Wirkung hin untersucht werden sollen, gelang es, solche Suggestiveffekte in z.T. erheblichem Ausmaße nachzuweisen.

So wird man unschwer feststellen können, daß die Frageformulierung: „Sind sie *auch* der Auffassung, daß das gegenwärtige Bildungssystem wenige bevorzugt und viele benachteiligt?", suggestiv ist, doch man wird ohne eine entsprechende Kontrollfrage nicht in der Lage sein, zu ermitteln, wieviele Personen sich durch die suggestive Formulierung zu einer positiven Antwort haben beeinflussen lassen. Als Regel wird man daher nur daraus ableiten können, daß jeder Forscher versuchen sollte, so neutral wie möglich zu formulieren, womit natürlich unbewußt suggestive Formulierungen nicht ausgeschlossen sind.

In vielen Fällen will der Sozialforscher Dinge von dem Befragten wissen, die dieser als so intim empfindet, daß es ihm schwerfällt, eine richtige Antwort darauf zu geben. Der Befragte ist in solchen Fällen in seiner Antwort gehemmt und neigt zu unwahren Angaben. Solche Antworthemmungen werden üblicherweise durch Fragen nach der politischen Einstellung, zu tabuisierten Sachverhalten ganz allgemein, insbesondere zu sexuellem Verhalten usw. provoziert. Fragestellungen mit Inhalten, die gesellschaftlich negativ sanktioniert, tabuisiert oder emotional tangierend sind, beschäftigen sich mit *heiklen Themen und provozieren Antwortwidersände.* Um solche Widerstände abzubauen, sind unterschiedliche Techniken entwickelt worden, die aber selbst nicht immer unproblematisch sind. Entschärfend, verharmlosend, überrumpelnd oder auf den Mitläufereffekt abstellend versuchen solche Fragen, heiklen Themen ihre Problematik zu nehmen. Das folgende Beispiel gibt 8 graduell unterschiedliche Möglichkeiten der Frageformulierung bei heiklen Themen in geradezu persiflierender Form an:

Persiflierende Darstellung von möglichen Frageformen bei heiklen Fragen, aus Public Opinion Quarterly, 22/1958, S. 67f.

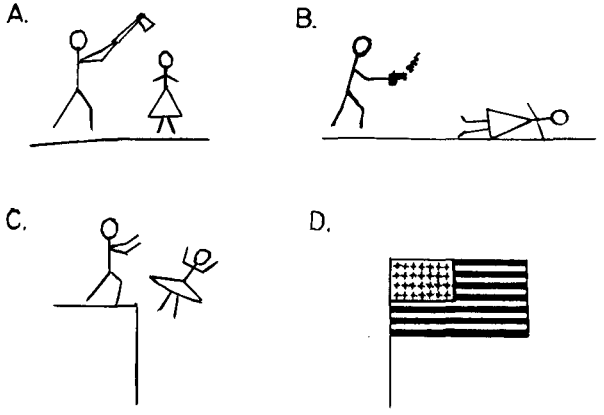
Asking the Embarrassing Question

BY ALLEN H. BARTON
University of Chicago

The POLLSTER's greatest ingenuity has been devoted to finding ways to ask embarrassing questions in non-embarrassing ways. We give here examples of a number of these techniques, as applied to the question, „Did you kill your wife?"

1. The Casual Approach:
„Do you happen to have murdered your wife?"
2. The Numbered Card:
Would you please read off the number on this card which corresponds to what became of your wife?" (HAND CARD TO RESPONDENT)
1. Natural death
2. I killed her
3. Other (What?)
(GET CARD BACK FROM RESPONDENT BEFORE PROCEEDING!)
3. The Everybody Approach:
„As you know, many people have been killing their wives these days.
Do you happened to have killed yours?"

4. The „Other people" Approach:
   (a) „Do you know any people who have murdered their wives?"
   (b) „How about yourself?"
5. The Sealed Ballot Technique:
   In this version you explain that the survey respects people's right to anonymity in respect to their marital relations, and that they themselves are to fill out the answer to the question, seal it in an envelope, and drop it in a box conspicuously labelled „Sealed Ballot Box" carried by the interviewer.



6. The Projective Technique:
   „What thoughts come to mind as you look at the following pictures?"

   (Note: The relevant responses will be evinced by picture D.)
7. The KINSEY Technique:
   Stare firmly into respondent's eyes and ask in simple, clearcut language such as that to which the respondent is accustomed, and with an air of assuming that everyone has done everything,
   „Did you ever kill your wife?"
8. Putting the question at the end of the interview.

Ob solche Methoden tatsächlich dazu in der Lage sind, heikle Themen in den Griff zu bekommen, kann nur im Einzelfalle und mehr oder weniger plausibel-argumentativ entschieden werden.

Fragetypen:

Neben diesen theoretischen Problemstellungen gibt es einige erhebungstechnische – in der Konsequenz wieder theoretisch determinierte – formale Möglichkeiten, Fragen zu formulieren, deren Kenntnis für das Verständnis und die praktische Tätigkeit von Bedeutung ist. Eines der pragmatisch und theoretisch wichtigsten Differenzierungskriterien von Fragen, stellt deren *Strukturiertheit* dar. So gehörte zu den hauptsächlichsten Diskussionsgegenständen lange Zeit die Frage, ob man *offene* oder *geschlossene Fragen* verwenden soll. *Unter offenen Fragen sind dabei solche zu verstehen, die keine Antwortmöglichkeiten vorgeben. Demnach sind geschlossene Fragen solche, die die Antwortvorgaben bereits enthalten.* Da der Grad der Strukturiertheit einer Frage, sofern sie einen geschlossenen Charakter aufweist, durchaus variabel sein kann, unterscheidet man bei den geschlossenen Fragen weitere Fragetypen:

Die Alternativfrage (Gegensatzfrage oder dichotomische Frage) läßt nur zwei Antwortalternativen zu. Eine Frage kann mit ja oder mit nein, positiv oder negativ beantwortet werden. Dieser Fragetypus ist der einfachste Fall einer geschlossenen Frage.

Die Auswahlfrage ist eine geschlossene Frage, die die dichotomische Frage erweitert, indem *mehrere Alternativen als Antwortmöglichkeiten vorgegeben* sind. Das einfache Gegensatzpaaar der Alternativfrage wird also ergänzt.

Durch das Einführen von Zwischenkategorien kann eine ursprüngliche Gegensatzfrage in eine Auswahlfrage der elementaren Form einer *Rangordnungsfrage* (Ordinalskala) übergeführt werden. (So kann die Frage „Haben Sie schon einmal ihre Frau verprügelt?" (Antwort: ja/nein) erweitert werden: „Wie häufig verprügeln Sie ihre Frau?" (Antwort: häufig, selten oder nie).

Eine Gegensatzfrage kann aber auch durch die Vorgabe eines *Katalogs von Antwortmöglichkeiten* erweitert werden, wobei diesen Antwortmöglichkeiten kein Kontinuum im Sinne einer Ordinalskala zugeordnet werden kann. Man spricht in solchen Fällen von *Cafeterieafragen.* Diese Frageform wird wohl am häufigsten in der Sozialforschung verwendet. Hierbei hat die Versuchsperson aus den vorgegebenen Antwortalternativen eine (bei Einfachnennungen) oder mehrere (bei der Zulassung von Mehrfachnennungen) anzugeben.

Bei diesen hier als Auswahlfragen bezeichneten Fragen (Cafeteria- oder Multiple-Choice-Fragen), kann man weiterhin danach differenzieren, ob der Interviewer die Antwortkategorie dem Befragten lediglich vorliest, oder ob er (z.B. wegen des Umfangs der Antwortkategorien) diese dem Befragten in schriftlicher Form überreicht. Erfolgt eine schriftliche Vorlage, so können wiederum zwei Hauptarten gegenübergestellt werden: Die *Listenfrage* in der die Antwortmöglichkeiten als Katalog oder Liste der Antwortalternativen vermerkt sind und das *„Kartenspiel"*, bei dem für jede Antwortkategorie eine eigene Karte verwandt wird.

Die Differenzierung nach einzelnen Fragetypen ist kein theoretisches und terminologisches Spielchen; vielmehr kann gezeigt werden, daß je nach Frageform von den Befragten unterschiedliche Leistungen verlangt und erbracht werden. Im Hinblick auf die grobe Unterscheidung nach offenen und geschlossenen Fragen läßt sich folgende Feststellung treffen: „Es fällt auf, daß offene Fragen in diesem Sinne vom Befragten verlangen, sich an etwas zu *erinnern* – und dies dann spontan darzustellen. Geschlossene Fragen verlangen dagegen vom Befragten etwas *wiederzuerkennen*. Die Literatur über das Gedächtnis durch Erinnerung und Wiedererkennen zeigt, daß mehr wiedererkannt als erinnert wird und dies erweist sich auch bei der Gegenüberstellung von offenen und geschlossenen Fragen". (MACCOBY, in: KÖNIG, R. (Hrsg.) Praktische Sozialforschung, Bd. 1, Die Befragung, Köln, 1966, S. 50). Aus diesen unterschiedlichen, vom Befragten zu verlangenden Leistungen geht hervor, daß der Einsatz von offenen oder geschlossenen Fragen auch theoretisch legitimiert werden muß. So verlangt die offene Frage, daß der Befragte seine Meinung selbst in Worte faßt, während bei der geschlossenen Frage, sich die Befragungsperson nur für eine der vorgegebenen Antwortalternativen zu entscheiden braucht. Je nach Verbalisierungsvermögen der potentiellen Befragungspersonen schließen sich daher bestimmte Frageformen aus.

Die Vor- und Nachteile der einzelnen Frageformen lassen sich gegenüberstellen und können als Entscheidungshilfen für deren jeweiligen Einsatz dienen:

Vorteile der geschlossenen Frageform

- Geschlossene Fragen helfen, eine Antwort in dem vorgegebenen, theoretischen Bezugsrahmen einer Untersuchung zu erhalten und zwar auch in der Form, daß die Antwort statistisch aufbereitet werden kann.
- Geschlossene Fragen machen den Frageinhalt deutlicher. Der Sinn der Frage wird durch die vorgegebenen Antwortkategorien zusätzlich erhellt.
- Geschlossene Fragen sind durch den Interviewer, wie auch bei der schriftlichen Befragung durch den Probanden, einfacher zu handhaben.
- Geschlossene Fragen erleichtern in erheblichem Maße die Auswertung.
- Daraus ergibt sich auch, daß die geschlossene Frageform normalerweise billiger und zeitsparender eingesetzt werden kann.

Nachteile der geschlossenen Frageform

- Ein erheblicher theoretischer Nachteil besteht darin, daß durch die Vorgabe einer bestimmten Antwortkategorie möglicherweise eine Antwort suggeriert wird, die vorher als Meinung des Probanden nicht gegeben war.
- Auf ähnlicher Ebene liegt der Nachteil, daß eine bestimmte Antwortnennung nicht notwendigerweise mit der tatsächlichen Meinung übereinstimmt. Die Vorgabe von Antwortkategorien provoziert Antworten, die dem Befragten die Möglichkeit eröffnen, sich „aus der Affäre zu ziehen".
- Sind die vorgegebenen Antwortkategorien im Hinblick auf die zu erhebende Fragestellung nicht vollständig und umfassend, so führt das Fehlen einer wichtigen Antwortkategorie zu erheblichen Daten- und Interpretationsfehlern, weil die Antworten sich auf die Vorgaben beschränken müssen.

Vorteile der offenen Frageform

- Offene Fragen helfen, unklare Frageformulierungen und unerwartete Bezugssysteme bei den Befragten durch deren Rückfragen zu entdecken.
- Offene Fragen suggerieren dem Befragten keine spezifische Antwort, wie dies durch die Vorgabe von Antwortkategorien geschieht.
- Offene Fragen erinnern mehr an eine alltägliche Kommunikationssituation, fördern da-

her den Kontakt zwischen Interviewer und Befragten.
- Offene Fragen liefern ein getreueres Abbild der Meinungen, da die Befragten mit ihren eigenen Worten formulieren, was sie meinen.

Nachteile der offenen Frage

- Es werden hohe Anforderungen an den Befragten im Hinblick auf sein Sprach- und Ausdrucksvermögen gestellt, was häufig – insbesondere bei Repräsentativbefragungen – nicht gegeben ist.
- Der Aufwand für die Auswertung und Analyse erhöht sich in erheblichem Maße, wobei Aufwand und Erkenntnisgewinn nicht in jedem Falle in einem ausballancierten Verhältnis stehen.
- Die Qualität, der auf offene Fragen gegebenen Antworten, hängt nicht zuletzt von deren richtiger Wiedergabe durch die Notizen des Interviewers ab, sodaß bei offenen Fragen auch erhöhte Anforderungen an den Interviewerstab zu stellen sind.

Aus diesen Vor- und Nachteilen der offenen und geschlossenen Frageform, lassen sich die unterschiedlichen Funktionen, die diesen im Rahmen einer Untersuchung zukommt, summarisch ableiten:

Geschlossene Fragen sind dann einzusetzen, wenn

- die Antwortalternativen bekannt sind,
- die Antwortvorgaben in Bezug auf die Frage innerhalb eines einheitlichen, eindimensionalen Bezugsystems liegen,
- eine bestimmte Zahl von Antworten sinnvoll vorgegeben werden kann,
- die Antwortalternativen trennscharf formuliert sind.

Offene Fragen werden gebraucht,

- wenn ein bestimmter Sachverhalt zunächst exploriert werden soll und die relevanten Dimensionen noch unbekannt sind,
- wenn es darum geht, gerade die noch unbekannten Bezugsysteme bei den Befragten selbst zu erkennen.

Geschlossene und offene Frageform bilden eine gewisse Analogie zur standardisierten, bzw. nichtstandardisierten Befragung. So wie das standardisierte Interview nur in der Endphase einer Untersuchung vorwiegend zum quantitativen Messen eingesetzt werden soll, gilt für die geschlossene Frage, daß sie nur nach genügender Kenntnis und vorherigen Exploration wirkungsvoll angewandt werden kann. Die Analogie von Standardisierung und Offenheit bzw. Geschlossenheit ergibt sich auch daraus, daß bei standardisierten Befragungen geschlossene Fragen häufiger eingesetzt werden als offene. Generell gilt jedoch, daß jede gute Befragung (schon aus psychologischen Gründen des Interviewablaufs) – unabhängig von dem Grad der Standardisierung – sowohl mit geschlossenen, wie auch mit offenen Fragen zu arbeiten hat.

Eine weitere Differenzierung von Fragetypen, ist von Relevanz: die Unterscheidung nach *direkter* oder *indirekter Fragestellung.* Schon bei der Behandlung der heiklen Fragen waren wir auf die Möglichkeit eingegangen, daß ein Sachverhalt nicht unmittelbar und direkt, sondern indirekt über Umwege erfragt wird. *Unter indirekten Fragestellungen versteht man solche Fragen, deren offenbare Bedeu-*

tung für den Befragten von dem eigentlich vom Forscher gemeinten Sachverhalt abweicht, die es jedoch ermöglichen, auf diesen Sachverhalt zu schließen. Zu solchen indirekten Fragen greift man dann, wenn die Befragten nicht fähig oder nicht willens sind, die erbetenen Informationen unmittelbar und direkt anzugeben. Der Unterschied zwischen direkter und indirekter Frage besteht also wesentlich darin, daß bei der direkten Frage für den Probanden Inhalt und Absicht der Fragestellung durchschaubar sind, während bei der indirekten Fragestellung eine für den Probanden scheinbar andere als die vom Forscher gemeinte Zielsetzung verfolgt wird. Die indirekte Fragestellung unterstellt, daß Sachverhalte, die – aus welchen Gründen auch immer – direkt erfragt, nicht zu gültig auswertbaren Antworten führen, mittels der indirekten Fragestellung durchaus zu erheben sind. Für die Frage, ob man mittels indirekter Fragestellung Erkenntnisse gewinnen kann, die durch eine direkte Frage nicht erhebbar gewesen wären, gibt es ein sehr plastisches Beispiel aus der Marktforschung, die Nescafé-Untersuchung (vgl. SCHEUCH, S. 148): Es ging darum, die Widerstände gegen den Gebrauch von löslichem Kaffee zu ermitteln, nachdem die üblichen Befragungsformen zu keinem vernünftigen Ergebnis geführt hatten. Man legte zwei vergleichbaren Gruppen je einen fiktiven Einkaufszettel mit der Aufforderung vor, Vermutungen über den Charakter der Person anzustellen, von der der jeweilige Einkaufszettel stammte. Beide Einkaufszettel waren mit einer Ausnahme völlig gleich: sie unterschieden sich nur darin, daß in einem Falle Pulverkaffee angeführt war, im anderen Bohnenkaffee. Als Ergebnis konnte festgehalten werden, daß ein erheblicher Teil der Befragten sich die Hausfrau, die löslichen Kaffee kaufen wollte, als bequem und nachlässig vorgestellt hat, und nur eine verschwindende Minderheit diese als modern und fortschrittlich perzipierte. An diesem relativ einfachen, indirekten Instrument las man jene Vorstellungen gegenüber dem neuen, löslichen Kaffee ab, die selbst zu formulieren den Befragten offensichtlich unmöglich war, bzw. deren Inhalt und Bedeutung ihnen nicht bewußt gewesen waren.

Die schon angesprochenen, projektiven Frageformen (wie Wortassoziationen, Satzergänzungen, Interpretation von diffusen Stimuli etc.) sind ein Spezialfall der indirekten Frage. Für die projektive Frage gilt wie für die indirekte, daß deren Validität nicht in jedem Falle gesichert ist. Da man bei *projektiven Fragen annimmt, daß der Proband sich selbst an die Stelle der Person setzen wird, auf welche die Frage Bezug nimmt, sodaß die gegebene Antwort in Wirklichkeit seine eigene Einstellung widerspiegelt,* werden solche Fragestellungen dann falsch, wenn diese Annahme nicht zutrifft. Auch hierfür ein illustratives Beispiel (Nach MACCOBY, in: KÖNIG, Praktische Sozialforschung Bd. 1 Die Befragung, Köln 1966, S. 53):

In einigen Fällen scheinen die Antworten tatsächlich die eigenen Einstellungen des Befragten wiederzugeben. In anderen hingegen, scheinen sie etwas sehr verschiedenes darzustellen, nämlich die realistische Bewertung der Einstellungen des in der projektiven Frage vorgeschobenen anderen. So wurde z.B. eine junge Frau über ihre Einstellung zur Berufssituation befragt: „Wie denken in ihrem Betrieb die meisten Mädchen über den Meister?" diese indirekte Frage sollte Aufschlüsse über ihre eigene persönliche Auffassung geben. Sie antwortete: „Sie meinen, er ist groß-

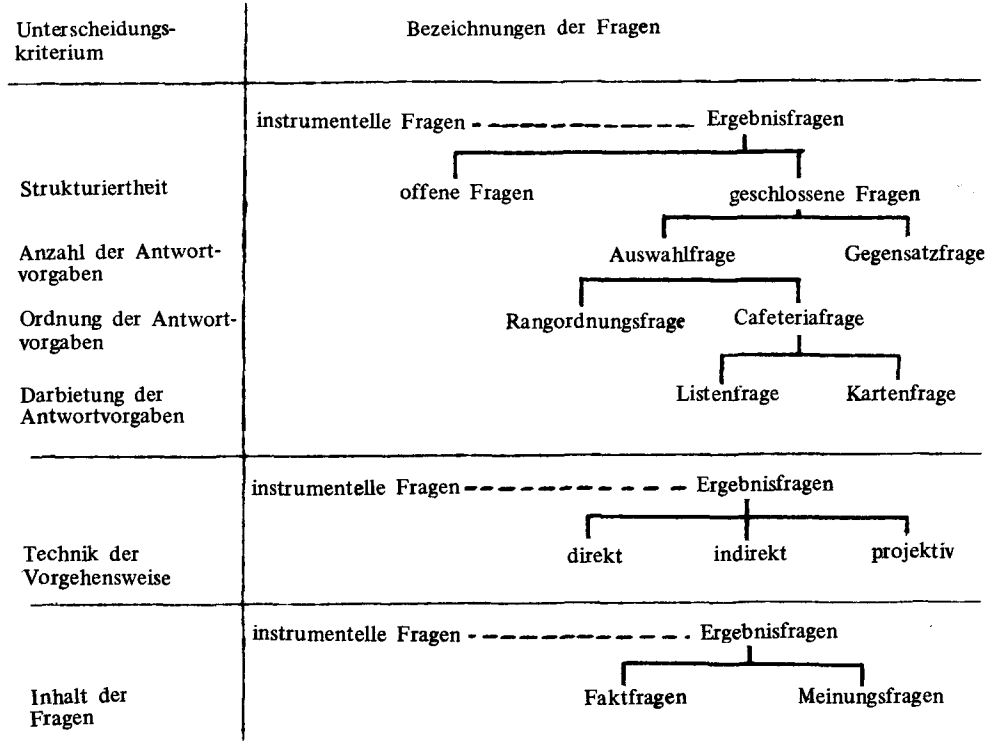
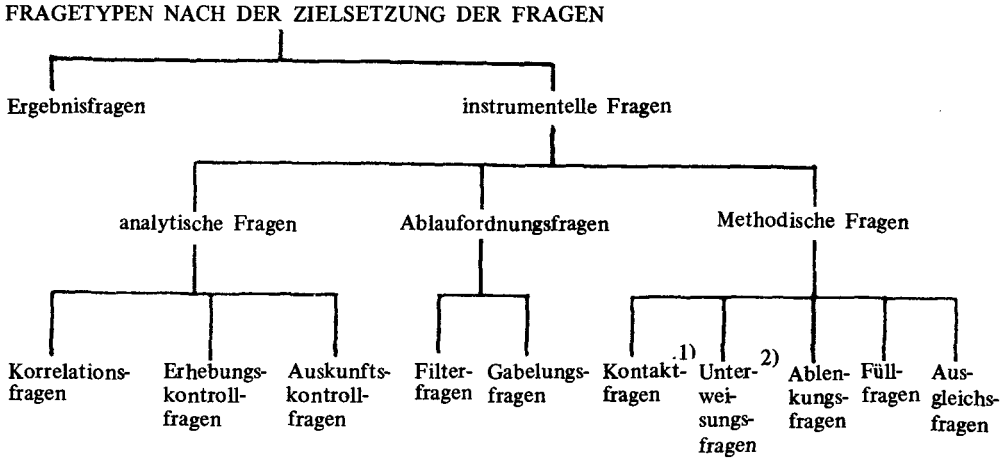


Abbildung 17: Übersicht über Arten von Ergebnisfragen.
Abbildung 17 und 18 nach STROSCHEIN, F.-R., Die Befragungstaktik in der Marktforschung, Wiesbaden 1965



1) Kontaktfragen werden auch als Eisbrecher- oder Einleitungsfragen bezeichnet.
2) Unterweisungsfragen werden auch als Trainings- oder Lehrfragen bezeichnet.

Abbildung 18: Übersicht über Arten von instrumentellen Fragen

artig, sie würden alles für ihn tun". An dieser Stelle nun folgte die direkte Frage zur Kontrolle der indirekten: „Und wie steht es mit Ihnen, was halten Sie von ihrem Meister?" und die Frau erwiderte: „Ich finde ihn abscheulich, ich versuche gerade, von hier wegversetzt zu werden". Hier hat man ein typisches Beispiel dafür, daß die Gültigkeit von indirekten und projektiven Fragestellungen nicht in jedem Falle gewährleistet ist. Hätte man zur Kontrolle die direkte Frage nicht formuliert, wäre man vermutlich zu völlig falschen Interpretationen gelangt.

Die bisher vorgestellten unterschiedlichen Fragetypen bezogen sich im wesentlichen auf *Ergebnisfragen, d.h. Fragen die sich auf inhaltlich-begründete und aus dem theoretischen Bezugsrahmen ableitbare Aspekte bezogen.* Abbildung 17 gibt einen summarischen Überblick über die Typen von Ergebnisfragen.

Neben den Ergebnisfragen sind jedoch auch *instrumentelle Fragen* zu berücksichtigen. Sie dienen dazu, einen Fragebogen technisch so zu gestalten, daß die Probleme von Gültigkeit und Zuverlässigkeit, von Antwortverweigerung und Rücklaufquoten etc. reduziert werden. Ein Fragebogen wird nicht in jedem Falle so zu konstruieren sein, daß nur solche Fragen aufgenommen werden, die dem inhaltlichen Interesse des Forschers, bzw. der Problemstellung selbst entspringen. Erfahrungswerte sprechen dafür, auch *methodische Fragen* in den Fragebogen aufzunehmen, *die die Funktion übernehmen, einen reibungslosen Ablauf des Interviews und der Befragung zu gewährleisten.* Solche Fragen, die der eigentlichen Auswertung nicht zugeführt werden, die aber für den Ablauf einer Befragung notwendig sein können, können je nach Zweck entsprechend klassifiziert werden. Hierzu die Übersicht in Abbildung 18.

Kontaktfragen dienen dazu, einen ersten Kontakt mit der Befragungsperson zu schaffen und eine Motivationslage herzustellen, die die Antwortbereitschaft bei den Befragten weckt.

Bei den *Unterweisungsfragen* handelt es sich um solche, die selbst nicht analysiert werden, die dem Befragten aber die Möglichkeit geben, sich in die Technik der Fragestellung einzuarbeiten. Sie sind besonders dann notwendig, wenn sich Fragestellung und Beantwortung als kompliziert erweisen. (z.B. skalenartige Fragen, Polaritätsprofile, etc.)

Ablenkungsfragen dienen dazu, den Probanden entweder vom eigentlichen Gegenstand zunächst abzulenken, damit er seine Antworten nicht kontrollieren und manipulieren kann, oder um Beeinflussungseffekte von einer vorhergehenden Frage auf eine nachfolgende auszuschalten (Haloeffekt).

Füllfragen leiten von einem Gegenstand zum anderen über, um einen zu abrupten Wechsel zu vermeiden, der bei den Probanden eventuell zu einem möglichen Abbruch der Befragung führen könnte. Füllfragen können aber auch eingesetzt werden, um eine entsprechende Antwortmotivation über einen längeren Fragebogen hinweg aufrechtzuerhalten oder zu steigern.

Ausgleichfragen werden eingefügt, um möglicherweise eingetretene Frustrationen bei den Probanden zu kompensieren. Hat z.B. eine bestimmte Frage Ärger hervorgerufen, so kann man versuchen, mit der nächsten Frage den Ärger abzubauen.

Zu den *Ablaufordnungsfragen* gehören die *Filter- und Gabelungsfragen*, die man auch als *Sondierungsfragen* bezeichnet. Die Filterfrage geht meist als Einleitungsfrage voraus und hat den Sinn, die Gesamtheit der Befragten in bestimmte Untergruppen zu zerlegen, die bezüglich eines bestimmten Merkmals in sich homogen sind, sich jedoch bezüglich eben dieses Merkmals von den anderen Untergruppen unterscheiden. Dieses trifft auch für die Gabelungsfragen zu. In der weiteren Fragenabfolge unterscheiden sich jedoch Gabelungs- und Filterfragen. Während bei den Gabelungsfragen jeder Untergruppe weitere, unterschiedliche Fragen gestellt werden, werden bei einem Filter nur ein oder zwei unterschiedliche Fragen für die ausgefilterte Gruppe gestellt, während alle weiteren Fragen für alle Befragungsgruppen identisch sind.

Bei den *analytischen Fragen* unterscheidet man zwischen *Korrelations-, Erhebungskontroll- und Auskunftskontrollfragen*. Bei den *Korrelationsfragen* kommt es weniger auf die Frage als solche, als vielmehr auf die Kombination dieser mit anderen Fragen an. *Erhebungskontrollfragen* beabsichtigen zu kontrollieren, ob in der Erhebung bestimmte Fehler aufgetreten sind. (So muß man die Interviewer im Hinblick auf Fälschungen etc. kontrollieren.) *Auskunftskontrollfragen* dienen den Zweck zu überprüfen, inwieweit die Probanden konsistent antworten und sich nicht innerhalb eines Fragebogens selbst widersprechen.

Die analytisch ausgearbeiteten Typen von Fragen schließen sich natürlich nicht gegenseitig aus. Vielmehr ergeben sich in der Realität erhebliche Überschneidungen. So können z.B. Korrelationsfragen sowohl instrumentellen- wie Ergebniszwecken dienen, sie können zugleich auch Filter- oder Gabelungsfragen, Cafeteriafragen oder offene Fragen, direkt oder indirekt formuliert sein.

4.1.6 Die Konstruktion eines Fragebogens

Nicht weniger bedeutsam als die Formulierung der einzelnen Frage ist ihre Stellung innerhalb des Fragebogens. Zu oft wird auch heute noch bei der Interpretation einer einzelnen Frage so vorgegangen, als ob es sich um einen isoliert dargebotenen Stimulus und nicht um einen spezifischen Punkt im Laufe einer längeren Kommunikationssituation gehandelt habe. Mit zunehmendem Erkenntnisfortschritt hat man jedoch der Fragenabfolge mehr Aufmerksamkeit geschenkt.

Folgende, allgemeine Gesichtspunkte müssen bei der Konstruktion eines Fragebogens berücksichtigt werden, wobei auch hierbei keine explizit theoretisch-empirische Absicherung zugrundeliegt, sondern vielmehr von Erfahrungssätzen ausgegangen wird:

1. Der Fragebogen soll so aufgebaut werden, daß ein *Interessenszuwachs* bei den Probanden von den ersten bis zu den letzten Fragen zu verzeichnen ist.

2. Der Fragebogen soll mit *einfachen Fragen beginnen* und zu komplizierteren Fragestellungen überleiten.
3. *Zunächst sind allgemeine, emotional nicht tangierende* und in alltäglichen Kommunikationssituationen durchaus auch zu stellende *Fragen zu formulieren.* Erst ab einem gewissen Zeitpunkt können schwierigere, tabuisierte, den intimen Persönlichkeitsbereich betreffende Fragen angesprochen werden.
4. Der Fragebogen soll so konstruiert werden, daß man von einem *Bezugsystem zum anderen überleiten kann.*
5. Ähnlich, wie es bei der Formulierung der einzelnen Frage nicht um die grammatikalisch korrekte, sondern um die umgangssprachlich leicht verständlichen Frageformulierung geht, fordert man als Grundsatz für die Konstruktion von Fragebogen, eine *psychologisch richtige*, statt einer logisch richtigen *Frageabfolge.* Hierbei wird versucht, den Gedankengang und die Übergänge von einem Bezugsystem zum anderen dem alltäglichen Denken anzupassen. Allerdings ist vor dem Irrtum zu warnen, daß sich logische und psychologische Fragefolge ausschließen würden. Häufig fallen sie sogar zusammen.

Diese Regeln kann man annähernd in eine grafische Darstellung bringen, die die Vorgehensweise bei der Fragebogenkonstruktion illustrieren soll. (Abb. 19.)

In der Darstellung ist senkrecht der Grad der Allgemeinheit der Fragestellung, der Schwierigkeitsgrad und der der Tabuisierung abgetragen, d.h. mit zunehmender Dauer des Interviews können Schwierigkeit, Tabuisierung und Spezifizierung zunehmen. Dabei wird davon ausgegangen, daß die assoziativen Blöcke (= Fragenkomplexe) immer kürzer werden, weil schwierigere Fragestellungen nicht über eine bestimmte Dauer hinaus beibehalten werden können. Dieses Ablaufschema einer Befragung erfährt für jede Untersuchung eine spezifische Ausprägung. Es stellt nur einen Anhaltspunkt und eine grobe Richtlinie dar.

Wie schon erwähnt, kann es zu sehr verzerrten Ergebnissen führen, wenn man versucht, einen bestimmten Sachverhalt nur mittels einer einzigen Frage zu erheben. Wir wissen aufgrund der Theorie der Frage, daß jede Frage neben ihrem intendierten Sinn durch in ihr enthaltenen Begriffe auch eine Fülle von Nebenbedeutungen entwickelt, die nicht auf die Untersuchungsthematik selbst bezogen sind. Um solche Nebenbedeutungen unter Kontrolle zu halten, werden zu einem Sachverhalt sinnvollerweise mehrere Fragen formuliert. Dies ist eine Begründung für den zunehmenden Einsatz von *Fragebatterien*, die man als *einen Satz von Fragen, die den gleichen Gegenstand durch verschiedene Formulierungen zu erheben versuchen,* (vgl. SCHEUCH, S. 149f.) bezeichnen kann.
(In Analogie hierzu kann das Diagnosebemühen des Arztes gesehen werden, der durch verschiedene Fragen die Symptome einer Krankheit zu klären sucht.) Solche Fragebatterien können zu Fragekomplexen verdichtet oder über den gesamten Fragebogen verstreut werden.

Faßt man Fragebatterien zu einzelnen Komplexen zusammen, so erleichtert dies

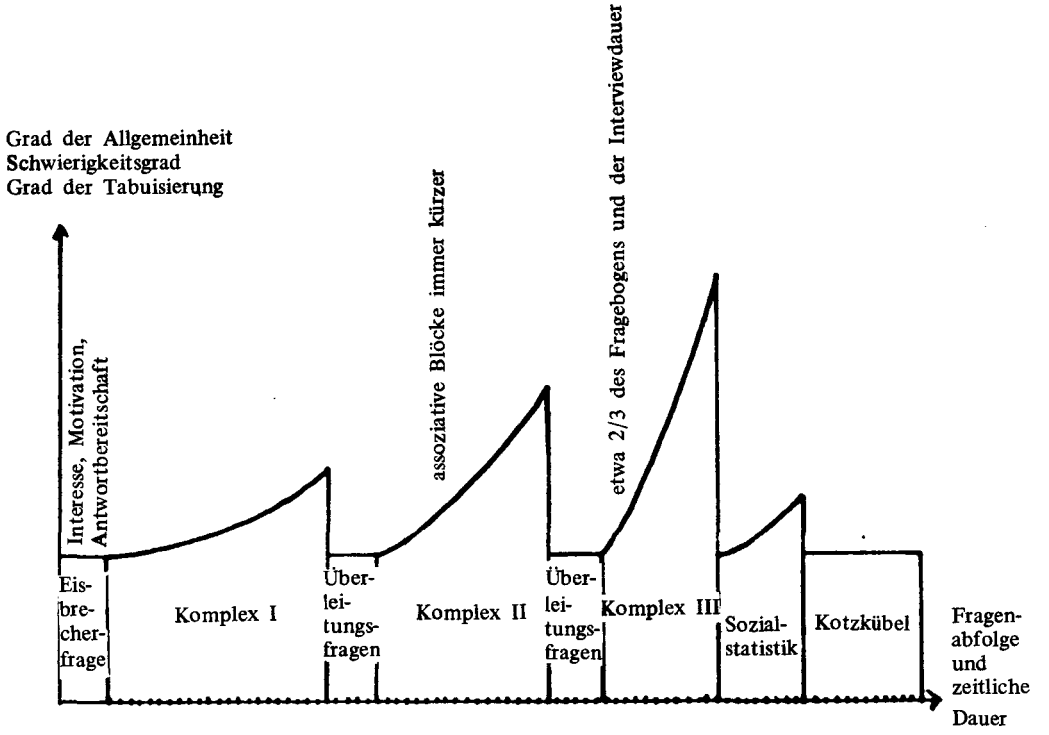


Abbildung 19: Darstellung des Fragebogenaufbaus.

eine differenzierte Stellungnahme durch den Probanden. Andererseits besteht bei dieser Vorgehensweise die Gefahr einer bewußt konsistenten Verzerrung der Antworten. Solche Wirkungen treten allgemein auf, wenn *benachbarte Fragen in der Beantwortung durch den Befragten Einfluß aufeinander ausüben.* Man bezeichnet dieses Phänomen als *Ausstrahlungs-* oder *Haloeffekt.* Diese Wirkung ist darauf zurückzuführen, daß jede Frage einen Bezugsrahmen für die folgenden Fragen setzt, innerhalb dessen der Befragte bestrebt ist, seine Antworten konsistent abzustimmen. Will man solche Ausstrahlungseffekte verhindern, so kann man sog. *Pufferfragen* dazwischensetzen oder die Fragebatterien auflösen und die einzelnen Fragen an verschiedenen „unverfänglichen" Stellen des Fragebogens placieren.

In der Fragebogenkonstruktion werden zwei unterschiedliche Möglichkeiten der Fragenabfolge unterschieden, der *positive und der negative Trichter* (umgekehrter Trichter). Bei positivem Fragetrichter werden zum gleichen Thema zuerst sehr allgemeine und nachfolgend immer spezifischere Fragen gestellt. Die Folgefragen dienen zur Präzisierung der zunächst allgemein gehaltenen, meist mehrdimensionalen Reaktion. Die logische Fortführung von allgemeinen zu spezifischeren Fragestellungen, ermöglicht auch, bestimmte Antwortbarrieren (z.B. bei heiklen Themen) abzubauen. Der umgekehrte Trichter, der weit weniger häufig angewandt wird, geht von spezifischen Fragestellungen aus und führt zu immer allgemeineren.

Wie z.T. schon ausgeführt, dienen die *Eisbrecher- oder Eröffnungsfragen* dazu, Interesse, Motivation und Antwortbereitschaft bei den Befragten zu schaffen. Zwischen den einzelnen Fragekomplexen sollten *Überleitungsfragen oder Erholungsfragen* gestellt werden, die einerseits eine „Regeneration" des Befragten ermöglichen, andererseits jedoch auch eine psychologische und logische Abfolge der Fragen bei dem Probanden suggerieren sollen. Nachdem die sozialstatistischen Merkmale erhoben worden sind (gelegentlich findet man die Demographie auch zu Beginn eines Fragebogens), schließt sich der „Kotzkübel" an. Er soll dem Probanden die Möglichkeit geben, seine Auffassungen, die bisher durch hohe Standardisierung und geschlossene Fragen in ein Korsett gezwängt waren, in seinen eigenen Formulierungen offen und breit als ein Ventil für die „erduldeten" Restriktionen mitzuteilen. Eine solche Kotzkübelfrage könnte z.B. lauten: „Welche Aspekte, glauben Sie, haben wir bei unserer Befragung übersehen?" oder „Welche Anregungen würden Sie geben?" etc.

Bei der Fragebogenkonstruktion ist auch die Praktikabilität des Fragebogens zu berücksichtigen, insbesondere der Gesichtspunkt, wie lang ein Fragebogen sein darf und sein kann, bevor er bei den Befragten Unmut und evtl. Abbrüche hervorruft. Auch hierfür lassen sich nur schwer allgemeine Regeln angeben. Sicher jedoch kann man davon ausgehen, daß schriftliche Befragungen normalerweise kürzer sein müssen als mündliche. Man kann auch generell sagen, daß es nicht so sehr auf die Zahl der Fragen ankommt, als vielmehr auf die für die Beantwortung der Fragen erforderliche Zeit (So ist die Frage: „Welchen Familienstand haben Sie", sicher schneller zu beantworten, als die nach der Einstellung zur Abtreibung).

Was nun die Interviewdauer betrifft, die als akzeptabel angesehen wird, muß wieder darauf hingewiesen werden, daß nicht so sehr die absolute Zeitdauer sondern vielmehr ein psychologisch akzeptables Maß entscheidend ist. Ein Interview, das 2 Stunden dauert, kann subjektiv als viel kürzer und interessanter empfunden werden, als eines mit einer Dauer von 20 Minuten. Normale und durchschnittliche Interviews werden sich jedoch innerhalb der Zeitspanne von einer Stunde bewegen. Alles, was darüberhinausgeht, muß als problematisch beurteilt werden und alle Interviews, die länger als eine halbe Stunde dauern, sollten in besonderer Weise psychologisch vorbereitet und getestet sein.

4.1.7 Fehlerquellen der Befragung

Die in diesem Abschnitt zu besprechenden Fehlerquellen beziehen sich auf die mündliche Befragung, weil dort zusätzlich zu den verzerrenden Effekten die sowohl bei der schriftlichen Befragung, wie auch beim Interview, z.B. durch fehlerhafte Frageformulierung, Fragebogenkonstruktion etc. Interviewereinflüsse auftreten können. „Kaum eine Phase der mit Hilfe von Interviews durchgeführten Datenermittlung enthält so viele ungelöste Probleme, wie der Prozeß der Befragung selbst. Nur wenig von dem, was sich im Augenblick der Befragung zwischen Interviewer und Befragtem über das reine Fragen und Antworten hinaus ereignet, und einen nur schwer zu durchschauenden Einfluß auf die Gültigkeit und Zuverlässigkeit der Datenermittlung ausübt, ist im Sinne der Objektivitätskriterien kontrollierbar." (MAYNTZ, R.u.a. Einführung in die Methoden der empirischen Soziologie, Köln und Opladen 1969, S. 114). Man weiß jedoch aus einer Fülle von Untersuchungen, daß selbst unter sonst gleichen äußeren Bedingungen, verschiedene Interviewer unterschiedliche Antwortverteilungen erhalten haben. Auf der Suche nach den Ursachen hierfür, wurden methodologisch orientierte Analysen angeschlossen, die einige wichtige Erkenntnisse zum Interviewereinfluß liefern konnten:

Ein erstes Ergebnis bestand darin, daß die *Einstellungen des Interviewers* zu den von ihm zu stellenden Fragen das Antwortverhalten der zu Befragenden beeinflussen können. Hierbei muß es keineswegs um eine bewußte Beeinflussung durch Offenlegung der Ansichten des Interviewers kommen; tatsächlich muß man davon ausgehen, daß unbewußte Beeinflussungsprozesse stattfinden, weil der Interviewer seine Einstellungen – trotz des Versuchs – nicht voll verbergen kann und der Befragte solche Einstellungen wahrzunehmen glaubt. Insbesondere durch subtile, auch nonverbale Verhaltensweisen des Interviewers, die dieser nicht immer kontrollieren kann, (Veränderung des Tonfalls, Räuspern, Mimik, Gestik etc.) greifen solche Beeinflussungsprozesse Platz. (So wird man bei Wahlstudien, wenn die Interviewer der SPD nahestehen, häufiger SPD-freundliche Antworten bekommen, als wenn diese eine andere Partei favorisieren würden). Solche Einstellungseffekte des Interviewers konnten empirisch bestätigt werden. Allerdings ist der Grad der Beeinflussung meist nicht exakt angebbar, außer wenn gezielt versucht wird, solche Beeinflussungsfaktoren zu kontrollieren.

Die Einstellungen des Interviewers determinieren in zweifacher Weise die Inter-

viewresultate: Einmal werden die Antworten der Probanden durch sie eventuell mitbeeeinflußt, andererseits kann es aber auch sein, daß in der *Registrierung der Antworten* subjektive Verzerrungen durch die Einstellungen des Interviewers wirksam werden. Dieses gilt natürlich für offene Fragen stärker als für geschlosssene.

So wie der Befragte den Interviewer in ganz bestimmter Weise perzipiert, und ihn bestimmten sozialen Kategorien zuordnet, so tut dieses der Interviewer auch mit dem zu Befragenden. Auch beim Interviewer entwickeln sich Erwartungen darüber, welche Antworten der Proband gemäß seiner sozialen Gruppenzugehörigkeit wohl geben wird. (So wird er von Frauen andere Antworten erwarten als von Männern, von Intellektuellen andere als von Berufslosen, von Hausbesitzern andere als von in Untermiete Wohnenden). Die im Verlaufe des Interviews sich herauskristallisierenden sozialen Merkmale führen als *Verhaltenserwartungen* dazu, daß der Interviewer *selektiv hört.* In der Erwartung bestimmter Antworten, die im übrigen auch durch früher durchgeführte Interviews mitdeterminiert sind, werden nun mehr solche Antworten wahrgenommen, die mit den Erwartungshaltungen kompatibel sind. Davon abweichende Stellungnahmen werden verdrängt, nicht wahrgenommen, oder nicht notiert.

In diesem Zusammenhang ist auch die *Konsistenzerwartung* des Interviewers von erheblicher Bedeutung. Unterstellt er nämlich eine einheitliche und logisch stringente Einstellungsstruktur aufgrund von Antworten, die der Proband zu früher gestellten Fragen gegeben hat, so werden auf ähnlich gelagerte Fragen, ähnliche, konsistente Antworten erwartet und die tatsächlich gegebenen Antworten selektiv im Sinne dieser Konsistenzerwartung perzipiert, oder im Extremfall werden diese Fragen gar nicht mehr gestellt und der Interviewer füllt sie im Sinne der Konsistenz selbst aus.

Ebenfalls experimentell abgesichert und belegt ist der Einfluß von *sozialen Merkmalen des Interviewers* auf das Antwortverhalten der Befragten. Bestimmte äußere soziale Merkmale werden durch den Befragten perzipiert und definiert und erscheinen neben den eigentlich interessierenden Stimuli als Variable, die das Antwortverhalten beeinflussen. So konnte man nachweisen, daß eine unterschiedliche Antwortverteilung erzielt wurde, je nach dem, ob Männer oder Frauen von Männern oder Frauen befragt, ob Neger oder Weiße von Negern oder Weißen interviewt wurden usw. Diese Erkenntnisse können noch dahingehend präzisiert werden, daß solche äußeren Merkmale des Interviewers einen besonders gravierenden, systematischen Verzerrungseinfluß ausüben, die in einem Zusammenhang zu den inhaltlichen Fragen des Interviews stehen (So dürfte es unerheblich sein, ob ein 30-jähriger, verheirateter, männlicher Proband von einer 50-Jährigen oder einem 30-jährigen Interviewer nach seinem Familienstand gefragt wird. Hingegen könnte von verzerrender Relevanz sein, wenn die Frage nach dem Familienstand von einer äußerst attraktiven 25-jährigen Frau gestellt wird). Es gilt also festzuhalten, daß soziale Merkmale des Interviewers das Antwortverhalten der Befragten beeinflussen können, daß jedoch eine solche Beeinflussung nicht generell, für alle Fragen und alle Befragungen gelten wird, sondern daß je nach Merkmal und je nach Frageinhalt

spezifische und unterschiedliche Wirkungen zu erwarten sind.

Die Interviewermüdigkeit ist eine weitere Fehlerquelle. Darunter versteht man nicht die allgemein physische Ermüdung des Interviewers, sondern man meint die Tatsache, daß mit zunehmender Zahl der durchgeführten Interviews, aber auch mit zunehmender Länge des Interviews, die Erhebungssituation sich so sehr wiederholt bzw. gleich bleibt, daß aus Gründen der Langeweile, ein Nachlassen der Aufmerksamkeit des Interviewers festzustellen ist. Solche Ermüdungserscheinungen wirken sich als systematische Fehlerquellen aus, weil durch mangelnde Konzentration und Aufmerksamkeit, nachlässige Frageformulierung (damit mangelnde Standardisierung der Reize) und unzuverlässige Registrierung der Antworten eintreten.

Als letzter verzerrender Einfluß des Interviewers sei die bewußte Fälschung des Interviews genannt. Da Interviewer häufig nach der Erfolgsquote bezahlt werden, und deren Qualität nach solchen Erfolgsquoten wie auch danach bemessen wird, ob alle Fragen des Fragebogens beantwortet sind, werden Interviewer dazu verleitet, z.T. ganze Interviews zu fingieren oder Antwortverweigerungen durch Fälschung einzelner Antworten zu kaschieren. Um solche Probleme auszuschließen, sind von den kommerziellen Umfrageinstituten, eine Fülle von subtilen Kontrollmechanismen entwickelt worden, die bewußte Fälschungen sehr schnell aufdecken können.

Begreift man die Interviewsituation – soziologisch gesehen – als sozialen Prozeß, so können die gegenseitigen Beeinflussungen zwischen Befragten und Interviewern in einer schematischen Darstellung nach CANNELL und KAHN (1968) zusammengefaßt werden: Abbildung 20.

Neben den Fehlern, als deren Quelle der Interviewer ausgemacht wurde, sind solche verzerrende Faktoren zu berücksichtigen, die in der *Befragungssituation* selbst liegen können. So gibt es Erkenntnisse darüber, daß der *Befragunsort* das Antwortverhalten und die Antwortverteilung beeinflussen kann. Sollen z.B. bestimmte Einstellungen von Schülern zur Schule erhoben werden, so dürfte es einen erheblichen Unterschied ausmachen, ob das Interview in der Schule, im Elternhaus oder an einem mit dem Schüler gesondert zu vereinbarenden Platz stattfindet. In Schule und Elternhaus werden allein durch die Räumlichkeiten bestimmte institutionelle Zwänge verspürt werden, die das Antwortverhalten verändern können.

Auf ähnlicher Ebene liegt das Argument, daß *Dritte, die beim Interview anwesend sind,* bewußt oder unbewußt Einflüsse auf das Antwortverhalten des zu Befragenden nehmen. So dürfte es plausibel sein, daß Lerneinstellungen – im Beisein des Lehrers oder der Eltern erfragt – nicht als zuverlässige und gültige Informationen angesehen werden können. Die Intervieweranweisungen laufen daher dahingehend, daß Anwesende Dritte beim Interview auszuschließen sind. Tatsächlich wird dieses jedoch nur in den wenigsten Fällen gelingen. (Man stelle sich vor, daß der Haushaltsvorstand in seinem Wohnzimmer im Beisein der Ehefrau befragt wird; es dürfte schwerfallen, die Ehefrau hinauszukomplimentieren). Nun gibt es allerdings

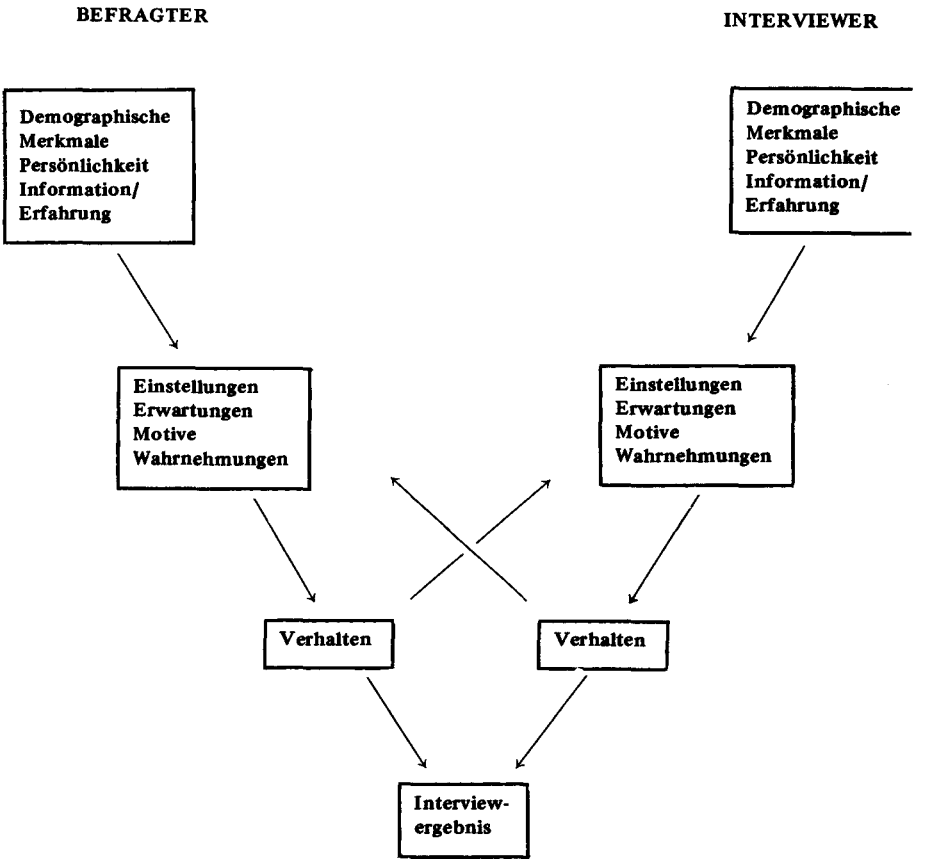


Abbildung 20: Das Interview sozialer Beeinflussungsprozeß (nach CANNELL & KAHN)

auch ein theoretisches Argument, weshalb Dritte nicht notwendigerweise ausgeschlossen werden müssen. Geht es z.B. bei der Befragungsaktion darum, weniger die Einstellung als solche, als vielmehr die Verhaltensrelevanz von Einstellungen zu erheben, so kann es durchaus sinnvoll sein, daß solche Personen beim Interview anwesend sind, die das faktische Verhalten der Befragungsperson auch in realen Situationen, (also ohne den Interviewer) mitbestimmen können. (So ist die unbeeinflußte Antwort auf die Frage, ob sich jemand im Laufe des nächsten Jahres gerne ein Auto kaufen möchte, unter dem Gesichtspunkt der Verhaltensrelevanz weit weniger brauchbar, als wenn in Anwesenheit der Ehefrau, der Befragte deren Auf-

fassung als die seiner eigenen konträr perzipiert und seine Antwort in entsprechender Weise modifiziert hätte).

Nicht unbedeutend für Erfolg oder Mißerfolg der Durchführung eines Interviews ist die Frage, wie der Befragte die Interviewsituation sozial definiert. Eine solche soziale Definition ist abhängig, von dem sozial-kulturellen Milieu des Befragten. So werden Mittelschichtangehörige eher Verständnis für die Notwendigkeit der Durchführung eines Interviews aufbringen, als Unterschicht- und Oberschichtangehörige. Je nach Definition der Rollensituation entwickeln sich auch verschiedene Bereitschaftsgrade an der Interviewsituation mitzuwirken. So berichten Interviewer häufig davon, daß sie abgewiesen wurden, weil die zu Befragenden den Verdacht geäußert hätten, es handele sich bei der Bitte um ein Interview, nur um einen simplen Verkaufstrick. Erfahrungshorizont, kulturelles Milieu des Befragten und Auftreten des Interviewers, können so zu Antwortverweigerungen oder Antwortverzerrungen führen.

Die in diesem Abschnitt angegebenen, wichtigsten Fehlerquellen beim Interview können – und darüber sollte man sich im Klaren sein – bestensfalls kontrolliert, nicht jedoch völlig ausgeschaltet werden. Selbst die Kontrolle dürfte schwierig werden, weil sie voraussetzt, daß alle möglichen verzerrenden Einflüsse vorab bekannt sind. Die unendliche Variabilität der Erhebungssituationen beim Interview schließt jedoch im Grunde eine umfassende Kenntnis der Verzerrungsfaktoren aus. Die bisher in der Literatur und durch Primärerfahrung vorliegenden Erkenntnisse zu den verzerrenden Faktoren im Interview, können allerdings dazu benutzt werden, entsprechende Interviewerschulungen durchzuführen, um somit wenigstens den Interviewern bewußt zu machen, welche unbemerkten und unerkannten Einflüsse sie ausüben können. Damit erhalten sie die Chance, solchen Einflüssen bewußt, gezielt und kontrolliert gegenzusteuern.

4.1.8 Gütekriterien der Befragung

Die Befragungsmethoden werden nach den allgemein entwickelten Gütekriterien, Objektivität, Zuverlässigkeit und Gültigkeit beurteilt. Allerdings geht der Gesichtspunkt der Objektivität beim Interview voll in das der Reliabilität mit ein, weil der Inteviewer selbst einen Teil der Meßsituation, ja des Meßinstruments darstellt. Daher ist normalerweise nur die Zuverlässigkeit und Gültigkeit des Erhebungsinstrumentes zu prüfen.

Die Zuverlässigkeit von Befragungen kann generell durch einen hohen Grad der Standardisierung erreicht werden. Je standardisierter nämlich die Fragen sind, umso weniger können verzerrende Einflüsse durch den Interviewer wirken. Weiterhin gilt allgemein: Je klarer eine Frage formuliert ist, je einheitlicher sie also von dem Befragten verstanden wird, desto größer ist ihre Zuverlässigkeit. Man kann mit Hilfe der Faktorenanalyse zeigen, daß Fragebatterien zu einem Gegenstand zuverlässigere Ergebnisse liefern, als Einzelfragen. Solche allgemeinen Erkenntnisse zur Zuverlässigkeit der Befragungsmethode müssen jedoch immer von Fall zu Fall geprüft

und auf den jeweiligen Untersuchungsgegenstand bezogen werden. Eine solche Prüfung genügt häufig nur Plausibilitätskriterien, weil die im Abschnitt 3.1.2. zu den Testverfahren entwickelten Methoden der Zuverlässigkeitsprüfung bei der Befragung nur sehr bedingt anwendbar sind: Die dort vorgestellten Reliabilitätskoeffizienzen können im Regelfall deshalb nicht berechnet werden, weil die einmalige Durchführung einer Befragung durch den subjekthaften Charakter des Probanden, durch das einmalige Anlegen des Meßinstruments bereits Veränderungen in den zu messenden Variablen provoziert. Test-Retest-Methoden sind aus diesem Grunde praktisch nicht durchführbar. Auch Pralleltestformen lassen sich bei der Befragung nur schwer realisieren, was auch für die Split-Half-Methode gilt.

Aus den Überlegungen zu den möglichen Fehlerquellen der Befragung kann geschlossen werden, daß sie erhebliche Gültigkeitsprobleme aufwirft, die bereits als Reliabilitätsprobleme virulent werden. Wenn man von diesen Problemen abstrahiert, bleiben weitere Gültigkeitsprobleme die FRIEDRICHS (Methoden empirischer Sozialforschung, Reinbek 1973, S. 223f) in drei Fragen formuliert.

1. Verstehen alle Befragten dasselbe unter der Frage, bzw. unter den Antwortvorgaben wie der Forscher selbst.
2. Sagt der Befragte, was er tatsächlich denkt und meint.
3. Handelt der Befragte so, wie er es im Fragebogen angibt.

Die erste Frage wird durch Plausibilität bzw. durch Erfahrungen, die beim Interviewen gesammelt werden, etwa annähernd beantwortet werden können. Die zweite Frage jedoch, ist wie die Erkenntnisse aus der Wahrnehmungs- und der Sozialpsychologie zeigen, nur mit erheblichen Schwierigkeiten, – wenn überhaupt – beantwortbar. So wissen wir, daß häufig nicht die eigene Auffassung zu einem, in Frage stehenden Problem wiedergegeben wird, sondern jene, von der man glaubt, daß sie *sozial erwünscht* sei. Da wir jedoch kein Instrument haben, das uns erlaubt, in die Denkvorgänge eines Individuums hineinzuschauen, wird man die Gültigkeit in Bezug auf die gestellte zweite Frage eher hypostasieren, als nachweisen können. Etwas anders liegt der Sachverhalt bei der dritten Frage: Auch hier wissen wir aus der Sozialpsychologie, daß nicht notwendigerweise Einstellungen unmittelbar zu einem mit ihnen kompatiblen Verhalten führen. Vielmehr können situative Faktoren ein Verhalten provozieren, das von den originären Einstellungen abweicht. Wenn aber eine Kongruenz zwischen Einstellungen und Verhalten nicht notwendigerweise zu erwarten ist, so erübrigt sich die dritte Frage ebenso, wie deren Gültigkeitsüberprügung. Wird jedoch das Verhalten als solches erfragt, dann gelten für die hierauf gegebenen Antworten dieselben Gültigkeitsprobleme wie bei der zweiten Fragestellung.

Die Gültigkeitsüberprüfung bei der Befragung gestaltet sich noch aus zwei weiteren Gründen schwierig: Einmal sind die durch Befragung ermittelten Sachverhalte meist nur durch die Methode der Befragung erhebbar, sodaß eine Vergleichbarkeit der Ergebnisse mit solchen anderer Methoden nicht gewährleistet ist. Zum anderen, fehlen zur Gültigkeitsüberprüfung meist die Außenkriterien, die einen Vergleich mit den über Befragung erhobenen Daten zulassen würden. Man wird deshalb grundsätzlich davon ausgehen müssen, daß insbesondere durch die Tatsache,

daß das Befragungsobjekt zugleich Subjekt ist, alle Ergebnisse, die mit dieser Methode erzielt worden sind, mit großer Vorsicht zu analysieren und zu interpretieren sind. Weil Validitätsüberprüfungen praktisch nicht vorhanden sind, sollte man die Methode der Befragung daher immer nur dann einsetzen, wenn andere Methoden sich aus inhaltlichen – oder Praktibilitätsgründen verbieten.

Die hier theoretisch und abstrakt entwickelten Vorbehalte gegenüber der Befragung, sind extrem und idealtypisch an theoretischen Maximierungskriterien ausgerichtet. Tatsächlich kann die Mehrzahl aller durchgeführten Befragungen auf Erfolge zurückblicken, die nicht zuletzt durch praktische Umsetzbarkeit der Ergebnisse (z.B. in Form von Prognosen) ausgewiesen werden. Man wird deshalb abschließend feststellen müssen, daß der Befragung erhebliche, theoretische Mängel und Probleme anhaften, daß jedoch bei Kenntnis dieser und unter Anwendung von Vermeidungs- und Kontrollstrategien die Befragung sich als durchaus nützliches, praktikables und wichtiges Erhebungsinstrument erwiesen hat.

4.2 Anwendung und Beispiel

In den letzten beiden Jahrzehnten hat sich in der wissenschaftlichen Fachwelt ein Streit über die Genese psychiatrischer Krankheiten entwickelt. Diese Kontroverse ist zurückzuführen auf unterschiedliche theoretische Vorstellungen (Schulen). Nachdem nämlich jahrzehntelang das biogenetische Krankheitsmodell mehr oder weniger unumstritten akzeptiert wurde, das von der Endogenität psychiatrischer Erkrankungen ausging, hat sich insbesondere mit dem Aufsatz von SZASZ (The Myth of Mental Illness) die Kritik an diesem Modell zunehmend darauf gerichtet, daß es soziale Aspekte vollständig ignoriere. Er stellt das alte tradierte Krankheitskonzept als solches in Frage und behauptet, daß der Begriff der Krankheit weder ausschließlich biologisch-medizinisch, noch ausschließlich psychologisch gesehen werden kann, sondern das vielmehr soziale Aspekte miteinbezogen werden müssen.

Anliegen einer wissenschaftlichen Untersuchung könnte es nun sein, den psychiatrischen Krankheitsbegriff zum Gegenstand der Analyse zu machen. Man könnte danach fragen, ob das Krankheitsverständnis der Psychiater in der Bundesrepublik Deutschland dem neuesten Stand der wissenschaftlichen Erkenntnis entspricht, bzw. ob neuere Überlegungen für den Krankheitsbegriff rezipiert wurden. (Man beachte, daß die Fragestellung nicht davon ausgeht, daß die neueren Überlegungen zum Krankheitsbegriff gültig und richtig sind. Vielmehr wird nur danach gefragt, ob sie bei den praktizierenden Psychiatern bekannt sind und ob sie in die berufliche Praxis mit einbezogen werden).

Als erste deskriptive Hypothese könnte formuliert werden: „Das Krankheitsverständnis der Psychiater entspricht nicht dem Stand der wissenschaftlichen Forschung". Damit ist der Untersuchungsgegenstand grob und vorläufig als das „Krankheitsbild der Psychiater" eingegrenzt. Die weiteren Überlegungen haben

sich nun darauf zu beziehen, mit welcher Methode dieser Untersuchungsgegenstand am besten zu erfassen ist. Eine zunächst negative Ausgrenzung führt dazu, daß Experimente, Test- und Beobachtungsverfahren als sachlich und ökonomisch inopportun ausscheiden. Da es sich bei dem Krankheitsverständnis um Vorstellungen und Einstellungen der Psychiater zum Phänomen der Krankheit handelt, bietet sich die Befragung an, die ja in besonderer Weise geeignet ist, Attitüden, Vorstellungen, Meinungen und Einstellungen zu erheben. Nehmen wir zudem an, daß aus finanziellen Gründen eine mündliche Befragung ausgeschlossen ist, so bleibt nur die Möglichkeit, einen Fragebogen für eine postalische Befragung zu erstellen. Bevor dies geschehen kann, muß der Untersuchungsgegenstand gerade auch im Hinblick auf das Untersuchungsziel und in Erweiterung der sehr allgemeinen und deskriptiven Hypothese spezifiziert werden.

4.2.1 Der Untersuchungsgegenstand

Will man das Krankheitsverständnis von Psychiatern ermitteln, so muß man zunächst einmal wissen, welche Krankheitsbilder theoretisch und/oder praktisch relevant sein können. Da eine schriftliche Befragung einen hochstandardisierten Fragebogen mit vornehmlich geschlossenen Fragen enthalten muß, können die Krankheitsbilder nicht erst explorativ durch die Fragen selbst erhoben werden, sondern sie müssen bereits als Vorgabe in die Konstruktion des Fragebogens einfließen. Daher sind einige theoretische Vorüberlegungen zum Krankheitsbild erforderlich:

Im wesentlichen lassen sich drei verschiedene Krankheitsbegriffe unterscheiden: der medizinische, der psychogenetische und der soziogenetische. Das medizinische Modell der psychiatrischen Krankheit geht davon aus, daß in Analogie zu organmedizinischen Pathologievorstellungen psychische Störungen biogenetisch determiniert sind. Dieser Krankheitsbegriff, der am naturwissenschaftlichen Modell orientiert ist, nimmt mehr oder weniger an, daß eine objektive und naturhafte Bestimmung des Patientenzustandes in „krank" oder „gesund" möglich ist.

Die psychogenetisch orientierten Vorstellungen von Krankheit, die insbesondere auf die Psychoanalyse zurückgehen, sehen die Ursachen der Erkrankung in unbewußten Konflikten des Individuums, die aus der Unvereinbarkeit der Triebansprüche mit den Anforderungen der sozialen Umwelt resultieren. Dieses psychogenetische Modell verknüpft Umwelt und Individuum, therapiert jedoch (ähnlich wie das medizinische Modell) das Individuum, weil davon ausgegangen wird, daß das Individuum selbst dafür verantwortlich zu machen ist, daß die potentiellen Konflikte zwischen Umwelt und Triebstruktur nicht bewältigt werden.

Beim soziogenetischen Krankheitsbegriff werden die Gründe für psychische Defekte vornehmlich im Zusammenwirken von Individuum und gesellschaftlichem Kontext vermutet, wobei die Priorität beim Sozialen gesehen wird. Wesentlichste Prämisse ist, daß sich gesunde von kranken Verhaltensweisen in deren Genese nicht voneinander unterscheiden.

Diese grobe (für den Demonstrationszweck aber völlig ausreichende) Differenzie-

rung in drei Krankheitsbegriffe, sollte als erster, theoretisch-begrifflicher Einstieg genügen. Je nach zugrundeliegendem Krankheitsverständnis sollte sich auch zeigen, daß verschiedene Therapieformen angewandt werden; geht man davon aus, daß Therapien nicht nur Symptome kurieren, sondern gerade die Ursachen der Symptome beseitigen wollen, so müssen sich aus einem unterschiedlichen Krankheitsverständnis auch unterschiedliche Therapieformen ergeben (von dem Fall, daß verschiedene Therapieformen für gleiche Ursachen angewandt werden können, sei aus Gründen der Einfachheit einmal abgesehen). Es käme also darauf an, Kategorien von Therapieformen zu finden, die den einzelnen Krankheitsbildern mehr oder weniger eindeutig zuzuordnen wären. So könnte man also über die Erfragung der Therapieformen indirekt auf das zugrundeliegende Krankheitsverständnis der Psychiater schließen (vgl. hierzu auch 4.2.2).

Wissenschaftliche Fragestellungen dienen nicht nur der Absicht beschreibende Aussagen zu machen, sondern es wird immer auch versucht, Erklärungen für bestimmte Sachverhalte zu finden. Solche Erklärungen für bestimmte Sachverhalte zu finden. Solche Erklärungen werden zunächst hypothetisch in relationale Hypothesen gefaßt, die dann mittels der erhobenen Daten überprüft werden sollen. Will man erklären, so fragt man nach potentiellen Ursachen oder unabhängigen Variablen. In unserem Beispielfalle würden wir also nach Ursachen dafür suchen, daß ein bestimmtes Krankheitsverständnis und mithin bestimmte Therapieformen favorisiert, andere jedoch ausgeschlossen werden. Man könnte die Vermutung haben, daß die „Art der Praxis" die Therapieformen determiniert. Differenziert man nur danach, ob jemand freipraktizierend oder in einem Krankenhaus psychiatrisch arbeitet, so könnten sich zwischen beiden Gruppen Unterschiede in der Therapieform ergeben. Als erste Hypothese, die wir überprüfen wollen, könnte man formulieren: „Die Art der Praxis determiniert die angewendeten Therapieformen" (oder in anderer Formulierung: „Die jeweils angewandte Therapieform ist abhängig von der Art der Praxis").

Wenn tatsächlich stimmt, daß sich die jeweilige Therapieform nur auf eine jeweils spezifische Krankheitsursache bezieht, so läßt sich daraus die weitere Hypothese ableiten: „Die Vermutung der Genese psychischer Erkrankungen (Krankheitsverständnis) determiniert die angewandten Therapieformen". Sucht man nach weiteren unabhängigen Variablen und zieht die nicht unbegründete Vermutung heran, daß solche Therapieformen praktiziert werden, die man zu beherrschen glaubt, die man also erlernt hat, so ließe sich daraus die Hypothese formulieren: „Je nach Ausbildung werden unterschiedliche Therapieformen angewandt".

Diese drei kursorisch und knapp entwickelten Hypothesen zum Untersuchungsgegenstand sollten zur Demonstration der Verfahrensweise bei der Befragung ausreichen. Tatsächlich stellen wir damit nur einen winzigen Ausschnitt des Hypothesensets zu dem Objektbereich dar. So müßte weiter danach gefragt werden, ob die hier als unabhängig angegebenen Variablen (Ausbildung der Psychiater, Art der Praxis und Genesenverständnis von psychischen Störungen) nicht selbst irgendwie voneinander abhängen. Die Aufstellungen solcher Hypothesen würde in deren

Überprüfung die Kenntnis multivariater statistischer Verfahren voraussetzen, die bei einer grundlegenden Darstellung, wie sie hier angestrebt wird, natürlich nicht behandelt werden können. Die Reduktion der Hypothesen auf drei bivariate ist also nicht methodologisch sondern ausschließlich pragmatisch begründet. Wir nehmen damit in Kauf, daß bei der statistischen Auswertung mögliche Phänomene, wie Scheinkorrelationen und Interventionen, nicht erkannt werden. (Diese Untersuchung wurde von einer studentischen Arbeitsgruppe unter meiner Leitung am Institut für Soziologie der Universität München durchgeführt; alle Überlegungen gehen (verkürzt) auf sie zurück.)

4.2.2 Die Operationalisierung und Erstellung des Erhebungsinstrumentes

Mit Operationalisierung war die Transformation der theoretischen Begriffe in Forschungsoperationen bezeichnet worden. Es geht also darum, jene, in den Hypothesen enthaltenen Begriffe (Variablen) durch entsprechende Fragen zu erheben. Für die gültige und zuverlässige Operationalisierung, also wie man zu Formulierungen gelangt, von denen man begründet annehmen kann, daß sie den in den theoretischen Begriffen angesprochenen Sachverhalt, auch messen, gibt es natürlich keine allgemeingültigen Regeln. Hier kommt es auf die Intuition und Phantasie des Forschers an.

Die in den Hypothesen aufscheinenden Begriffe waren: Ausbildung der Psychiater, Art der Praxis, Krankheitsverständnis (Vermutung über die Ursachen psychischer Erkrankungen), und Therapieformen. Aufgrund der theoretischen Überlegungen wollten wir die Art der Praxis nach freipraktizierend und im Krankenhaus arbeitend dichotomisieren. Selbstverständlich wären weitere Differenzierungen möglich, doch erschienen nur diese hypothetisch-theoretisch nicht relevant zu sein. Die Art der Praxis könnte demnach erhoben werden über die Frage: „Sind Sie freipraktiziernd oder arbeiten Sie in einem Krankenhaus?"

Die Frage der Ausbildung der Psychiater ist nicht mehr ganz so einfach zu beantworten, weil hier die unterschiedlichsten fachspezifischen und fachübergreifenden Kombinationen von Ausbildungen auftreten können. Um alle diese Möglichkeiten erfassen zu können, könnte man den Begriff der Ausbildung in zwei Fragen fassen; einmal nach der Grundausbildung und zum anderen nach der Zusatzausbildung. Werden diese Fragen offen gestellt, so werden die gegebenen Antworten nachträglich bei der Auswertung kategorisiert und in eine theoretisch sinnvolle und trennscharfe Klassifikation gebracht. So könnte sich z.B. durch die deskriptive Häufigkeitsverteilung herausstellen, daß folgende Klassifikation auch theoretisch sinnvoll erscheint: „Facharzt für Neurologie und Psychiatrie" als erste Kategorie, als zweite „Facharzt und psychotherapeutische Ausbildung" und als dritte „Facharzt und psychoanalytische Ausbildung".

Die Frage, wo die Psychiater die Ursachen für psychische Störungen sehen, kann offen gestellt werden, womit keine bestimmten Antworten suggeriert würden. Andererseits bestünde die Gefahr, relativ unvergleichbare Antworten zu bekommen,

sodaß eine quantitative Auswertung erschwert wäre. Gerade bei schriftlichen Befragungen wird man jedoch häufig auf geschlossene Fragen zurückgreifen, sodaß die Frage nach den Ursachen als Cafeteriafrage gestellt werden kann: einige denkbare Ursachen für psychische Störungen werden angegeben, und eine Restkategorie „Sonstiges", unter die alle anderen Antworten subsumiert werden, stellt die Vollständigkeit des Kategoriensystems her. Eine diesbezügliche Frage könnte z.B. lauten: „Worauf sind nach ihrer Ansicht psychische Störungen zurückzuführen?"

Als vierte, wichtige Variable muß noch die der Therapieformen operationalisiert werden. Da solche Therapieformen sicherlich nicht in Ausschließlichkeit verwandt werden, (ebenso wenig, wie prinzipiell nur eine Ursache für psychische Störungen angegeben werden kann), empfiehlt es sich danach zu fragen, in welcher Häufigkeit welche Therapieformen praktiziert werden. Der entsprechende Fragebogen zur Überprüfung der oben formulierten Hypothesen, könnte demnach das folgende Aussehen haben:

1. **Sind Sie freipraktizierend oder arbeiten Sie in einem Krankenhaus?**

   ☐ freipraktizierend

   ☐ Krankenhaus

2. **Welche Grundausbildung haben Sie?**

   .....

3. **Welche Zusatzausbildung haben Sie?**

   .....

4. **Worauf sind Ihrer Ansicht nach psychische Störungen zurückzuführen?**
   Bitte geben Sie die Ihrer Ansicht nach wichtigsten vier Punkte an, indem Sie in die Kästchen die Ziffern 1-4 eintragen, wobei 1 der wichtigste Punkt, 2... ist.

Psychische Störungen sind zurückzuführen auf:

- ☐ frühkindliche Traumata
- ☐ Kommunikationsstörungen in der Familie
- ☐ biochemische Veränderungen des Organismus
- ☐ intrapersonelle Konflikte
- ☐ soziale Lernprozesse
- ☐ Vererbung
- ☐ Identifikation des Kranken mit der ihm zugeschriebenen Krankenrolle
- ☐ sonstiges: .....

  .....

5. **Wie häufig wenden Sie welche Behandlungsform an?**

	sehr oft	oft	manchmal	selten	nie
Gesprächstherapie					
Milieutherapie					
Pharmakotherapie					
Beschäftigungstherapie					
Gruppentherapie					
Schocktherapie					
Psychoanalyse					
somatisch-chirurgische Therapie					
Verhaltenstherapie					
Familientherapie					

Es versteht sich von selbst, daß bei nur drei zu testenden Hypothesen und einem entsprechend kurzen Fragebogen, die theoretisch erarbeiteten Gesichtspunkte zur Makroplanung eines Fragebogens (Fragenreihenfolge etc.), hier keine Berücksichtigung finden können. Daher wurden die Fragen einfach im Verlaufe der oben entwickelten Operationalisierung angeordnet.

4.2.3 Die Datenerhebung

Vor jeder Datenerhebung steht die Frage, welches die *Untersuchungs-* und welches die *Auswahleinheiten* sind. Da wir unsere Aussagen auf die Psychiater der Bundesrepublik Deutschland beziehen wollen, gelten als Untersuchungseinheiten alle Psychiater in diesem lokal abgegrenzten Raum. Da es jedoch einen erheblichen Aufwand bereiten würde, diese Psychiater alle zu ermitteln und sie alle in die Untersuchung einzubeziehen, muß es darum gehen, eine Stichprobe zu ziehen, die eine Katalogisierung der Untersuchungseinheiten nicht erfordert und durch den entsprechend reduzierten Personenkreis gleichwohl repräsentative Aussagen für alle Psychiater in der Bundesrepublik Deutschland zuläßt. Gehen wir der Einfachheit halber von der Annahme aus, daß alle Münchener Psychiater sich nicht signifikant von den Psychiatern der gesamten Bundesrepublik Deutschland unterscheiden, so könnten wir als Stichprobe definieren: In die Auswahl gelangen alle Münchener Psychiater. Diese sind mittels Telefonbuch festzustellen, wobei auch hier noch Stichprobenverzerrungen zu berücksichtigen wären.

Der ausgewählte Personenkreis wird angeschrieben und darum gebeten, den beiliegenden Fragebogen auszufüllen und innerhalb einer bestimmten Zeitspanne mit Hilfe des beiliegenden Freiumschlages an die angegebene Institution zurückzuschicken. Nun werden nicht alle angeschriebenen Psychiater den Fragebogen beantworten, sodaß nur der Teil, der ausgefüllt und auswertbar zurückgeschickt wird, in die Analyse gelangen kann. Je niedriger die *Rücklaufquote* ist, desto eher muß davon ausgegangen werden, daß *systematische Verzerrungen* bei der statistischen Interpretation im *Repräsentativitätsschluß von der Stichprobe auf die Grundgesamtheit* vorgenommen werden. Wenn wir all diese Detailprobleme im weiteren Fortgang vernachlässigen und davon ausgehen, daß unsere Aussagen, die wir für die eingegangenen Fragebogen machen, auch für die Grundgesamtheit Geltung haben, so verkennen und verharmlosen wir diese Schwierigkeiten nicht, meinen jedoch, daß sie für die Darstellung der Auswertungstechniken von untergeordneter Relevanz sind.

4.2.4 Datenauswertung- und Interpretation

Jede Datenanalyse wird zunächst damit beginnen, daß die auf die Fragen erhaltenen Antworten akkumuliert und als Häufigkeitsverteilung ausgegeben werden. Diese erste deskriptive Analyse liefert über einfache statistische Überlegungen und Manipulationen, Parameter der Verteilungen, die in der Lage sind, die aufgestellten deskriptiven Hypothesen zu überprüfen. So könnte man z.B. als solche Hypothese vermutet haben, daß freipraktizierende Psychiater gegenüber den im Kran-

kenhaus angestellten, in der Minderheit sind. Die Auszählung der Antworten auf die diesbezügliche Frage im Fragebogen ergibt jedoch eine annähernde Gleichverteilung, nämlich: 54% sind freipraktizierend und 46% in Krankenhäusern beschäftigt. Die Differenz zwischen beiden dürfte innerhalb der statistischen Schwankungsbreiten liegen, sodaß in der Tat von einer annähernden Gleichverteilung gesprochen werden kann und die oben formulierte deskriptive Hypothese als widerlegt gilt.

Die Auszählung der Frage 4 und 5 mag ergeben haben, daß es sinnvoll ist, beide zu einer Variablen zusammenzufassen, da praktisch alle Psychiater die gleiche medizinische Grundausbildung in unserer Stichprobe haben. Bildet man die drei Klassen „Facharzt für Neurologie und Psychiatrie", „Facharzt und psychiatrische Ausbildung", „Facharzt und psychoanalytische Ausbildung", so zeigt sich, daß etwa 1/4 der ersten Kategorie, 2/5 der zweiten und 1/3 der dritten Kategorie zuzuordnen sind.

Die deskriptive Analyse ist eine Vorstufe für die Überprüfung der relationalen Hypothese, sodaß die oben angestellten und nur paradigmatisch-kursorisch vorgenommenen deskriptiven Überlegungen, notwendigerweise durch weitere Analyseschritte ergänzt werden müssen. Wir kehren daher zu der zuerst aufgestellten Hypothese zurück. Dort wurde ein Zusammenhang zwischen den angewendeten Therapieformen und der Art der Praxis postuliert. Die Operationalisierung der Therapieformen wurde jedoch sehr differenziert vorgenommen (vgl. Frage 5) und zwar in einer Weise, die zunächst nicht kompatibel mit dem theoretischen Konzept des Krankheitsverständnisses zu sein scheint, weil dieses selbst nur trichotomisiert war. Die zu differenziert erhobenen Informationen müssen so zusammengefaßt werden, daß sie das theoretisch Gemeinte repräsentieren. Daher werden Gesprächstherapie und Verhaltenstherapie zu der Therapieform vereint, die am Patienten ansetzt, Milieutherapie und Familientherapie werden als am sozialen Umfeld sich orientierende Therapie verstanden, während Pharmakotherapie und Schocktherapie als medizinische Vorgehensweise klassifiziert werden. Somit können wir drei unterschiedliche Korrelationstabellen konstruieren, die jeweils die Art der Praxis mit den Therapieformen kombinieren.

Die Tabelle 27 zeigt, wie sich die Häufigkeit einer patientorientierten Therapie auf die Art der Praxis verteilt:

Tabelle 27: Art der Praxis und personenorientierte Therapieform

Praxis / personenorientierte Therapie	freipraktizierend	Krankenhaus	Σ
ja	50% (50)	94% (80)	130
nein	50% (50)	6% (5)	55
Σ	100	85	185

Wenn wir uns die Prozentzahlen ansehen und miteinander vergleichen, so stellen wir fest, daß die Hälfte der freipraktizierenden Psychiater noch nie eine personorientierende Therapieform angewandt haben, während dies bei den im Krankenhaus angestellten Psychiatern etwa nur 6% sind. Auch die anderen Zellenhäufigkeiten zwischen den Spalten unterscheiden sich gravierend. Es erhebt sich nun die Frage, ob diese aufscheinenden Differenzen zufällig entstanden sind, oder ob diese Differenzen so groß sind, daß sie als statistisch signifikant ausgewiesen werden können. Diese Fragestellung wird mit Hilfe des Chi^2-Tests überprüft.

Wir berechnen also die Erwartungswerte für die Tabelle unter der Voraussetzung der Gültigkeit der Nullhypothese:

$$fe_1 = \frac{100 \cdot 130}{185} \approx 70$$

$$fe_2 = \frac{100 \cdot 55}{185} \approx 30$$

$$fe_3 = \frac{85 \cdot 130}{185} \approx 60$$

$$fe_4 = \frac{85 \cdot 55}{185} \approx 25$$

$$\chi^2 = \Sigma \frac{(fo - fe)^2}{fe}$$

$$= \frac{(50 - 70)^2}{70} + \frac{(50 - 30)^2}{30} + \frac{(80 - 60)^2}{60} + \frac{(5 - 25)^2}{25}$$

$$\chi^2 = 41{,}7$$

Die Berechnung der Freiheitsgrade ergibt:

$$df = (\text{Spalten} - 1)(\text{Zeilen} - 1) = (2 - 1)(2 - 1) = 1$$

Als Signifikanzniveau (Irrtumswahrscheinlichkeit) geben wir 0,001 vor. Für dieses Niveau hätten wir bei einem Freiheitsgrad ein χ^2 von mindestens 10,8 (kritischer Tabellenwert) erhalten müssen. Da unser errechneter χ^2= Wert bedeutend größer ist, können wir die Nullhypothese, daß keine Beziehung zwischen den Variablten besteht, mit einer Irrtumgswahrscheinlichkeit von mindestens 0,1% zurückweisen und unsere aufgestellte Alternativhypothese beibehalten.

Wenn man ermittelt hat, daß ein Zusammenhang zwischen den korrelierten Variablen nicht mehr zufällig begründet werden kann, dann sollte die weitere Frage interessieren, wie stark dieser Zusammenhang denn ist. Diese Fragestellung kann mit Hilfe der Maßzahlen zur Korrelation beantwortet werden. Im Falle einer Vielfeldertafel wird φ berechnet.

$$\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{41{,}7}{185}} \approx 0{,}48 \text{ oder}$$

$$\varphi = \frac{ad - bc}{\sqrt{s_1 \cdot s_2 \cdot s_3 \cdot s_4}} = \frac{(50 \cdot 5) - (80 \cdot 50)}{\sqrt{100 \cdot 85 \cdot 130 \cdot 55}} \approx (-)0{,}48$$

Wir stellen also eine mittlere Stärke der Beziehung zwischen den beiden Variablen fest. (Da wir nur nominal gemessen haben, hat das Vorzeichen bei der zweiten Berechnungmethode von φ keine Bedeutung und kann deshalb auch nicht interpretiert werden.)

Mit der Zurückweisung der Nullhypothese und der Berechnung des Korrelationskoeffezienten ist allerdings die eingangs aufgestellte erste Hypothese noch nicht vollständig überprüft, denn wir haben ja, wie die Tabelle zeigte, nur die am Patienten orientierte Therapieform getestet. Es muß sich also nun dieselbe Vorgehensweise für die medizinisch- und die sozialorientierte Therapie anschließen. Hierzu habe man die folgenden Tabellen erhalten:

Tabelle 28: Art der Praxis und medizinische Therapie

Praxis / medizinische Therapie	freipraktizierend	Krankenhaus	Σ
ja	39% (35)	81% (65)	100
nein	61% (55)	19% (15)	70
Σ	90	80	170

$\chi^2 = 28{,}7; \text{df} = 1; \alpha < 0{,}001$

$\varphi = 0{,}41$

Tabelle 29: Art der Praxis und sozialorientierte Therapie

Praxis / orientierte Therapie	freipraktizierend	Krankenhaus	Σ
ja	32% (30)	82% (70)	100
nein	68% (65)	18% (15)	80
Σ	95	85	180

$\chi^2 = 47{,}7; \text{df} = 1; a < 0{,}001$

$\varphi = 0{,}51$

Die Berechnung der Maßzahlen für die Tabelle 28 mit der medizinisch orientierten Therapie ergibt ebenfalls, daß die Nullhypothese zurückgewiesen werden kann und daß ein mittelstarker Zusammenhang zwischen den beiden Variablen besteht. Der gleiche Sachverhalt zeigt sich auch bei der sozial orientierten Therapie. Die aufgestellte Hypothese, daß sich Unterschiede in den Therapieformen zwischen freipraktizierenden Psychiatern und denen, die in Krankenhäusern angestellt sind, ergeben werden, kann durch die Daten nicht widerlegt werden. Da die Überprüfung der anderen beiden in 4.2.1. formulierten Hypothesen in analoger Weise erfolgt, kann auf deren weitere Darstellung hier verzichtet werden.

Ergänzende und vertiefende Literatur:

KÖNIG, R., Das Interview, Formen, Technik, Auswertung, Köln 1972

NOELLE, E., Umfragen in der Massengesellschaft, Reinbek 1976

RICHTER, H.J., Die Strategie schriftlicher Massenbefragungen, Harzburg 1970

SCHEUCH, E.K., Das Interview in der Sozialforschung, in: KÖNIG, R. (Hg.), Handbuch der empirischen Sozialforschung Bd.1, Stuttgart 1967(2)

5 SYSTEMATISCHE VERHALTENS- UND SELBSTBEURTEILUNG

Dieses Kapitel fällt, gemessen an den Vorhergehenden, in denen Methoden sozialwissenschaftlicher Forschung vorgestellt wurden, etwas aus dem Rahmen. Die systematische Verhaltensbeurteilung müßte unter dem methodischen Kapitel der Beobachtung abgehandelt werden. Eine ausführliche Darstellung der Methode der Beobachtung wird hier jedoch nicht vorgenommen, weil der Gegenstandskatalog der medizinischen Vorprüfung nicht explizit darauf abstellt. In ihm werden nur einige Detailphänomene der Beobachtung unter dem Gliederungspunkt 1.6. „Methoden der systematischen Verhaltensbeurteilung und Selbstbeurteilung" eher eklektizistisch herausgegriffen. Eine ausführliche Behandlung der Methode der Beobachtung verbietet sich daher aus pragmatischen Gründen.

Die Beobachtung kann nicht nur als eine spezielle Methode der Datenerhebung sondern sie kann auch begriffen werden als jenes grundlegende Prinzip, das allen empirischen Wissenschaften zugrundeliegt, nämlich durch den Einsatz der Beobachtung (wie immer diese dann im Einzelfalle aussehen mag), Feststellungen über die Realität zu treffen, damit diese mit den theoretischen Aussagen konfrontiert werden kann. Insoweit ist die Beobachtung methodenübergreifend. Tatsächlich sind auch die im Gegenstandskatalog genannten Unterpunkte hierzu methodisch-methodologische Probleme, wie sie bei allen Methoden auftreten können, soweit sie auf Beobachtungsverfahren rückführbar sind. (So können z.B. Rating-Skalen, sowohl im Fragebogen, wie auch im Test, ja sogar im Experiment auftreten; die mit ihnen verbundene Probleme sind also methodenübergreifend, was auch für die systematischen Tendenzen bei der Verhaltensbeurteilung und andere Fehlerquellen gilt). Deswegen sollte dieses Kapitel 5. als ein methodologisch orientiertes betrachtet werden, weshalb auch auf ein paradigmatisch herausgegriffenes Beispiel, an dem die Einzelphänomene demonstriert werden, verzichtet wird. In seinem Aufbau entspricht daher Kap. 5 etwa dem Kap. 1.

Systematische Verhaltensbeurteilung oder Selbstbeurteilung setzt voraus, daß auf irgendeine Art und Weise das zu beurteilende Phänomen an der Realität beobachtet wurde. Vor der Beurteilung steht also die Beobachtung. Während nun die *Fremdbeobachtung als Verhaltensbeobachtung* zu den kontrollierten Beobachtungsverfahren gehört und sich den methodologischen Kriterien von Objektivität, Reliabilität und Validität unterwirft, also eine intersubjektive Überprüfbarkeit der Beobachtungsresultate akzeptiert und ermöglicht, gilt für die *Selbstbeobachtung* (Introspektion), daß das Anlegen der wissenschaftlichen Gütekriterien an sie zu ungünstigeren Resultaten führt. Neben diesem methodologischen Unterschied, der unten noch spezieller zu behandeln sein wird, gilt auch, daß sich die Verhaltensbeobachtung als *Fremdbeobachtung, auf äußere, overte Verhaltensweisen bezieht, während die Selbstbeobachtung sich eher auf inneres, psychisches Erleben, Fühlen und Empfinden richtet.* Es versteht sich von selbst, daß die Beobachtung innerpsychischer Vorgänge – insbesondere durch den von diesen Vorgängen unmittelbar Betroffenen – problematisch erscheinen muß. Zwar kann die Introspektion nicht

völlig ausgeschlossen werden, weil eben bei innerpsychischen Vorgängen und deren wissenschaftlicher Erkundung man auf die Selbstbeobachtung angewiesen ist, doch muß für sie gelten, daß ihre Erkenntnisse mit relativem Vorbehalt zu sehen sind.

Verhaltensbeurteilung und Selbstbeurteilung stellen den Schritt dar, der sich an die unmittelbare Beobachtung als Interpretation oder Wiedergabe dieser Beobachtung anschließt. Der Begriff der Beurteilung ist dabei etwas unglücklich gewählt, weil er ein evaluatives, normative Element in sich trägt. Es ist jedoch mit Beurteilung nicht gemeint, daß z.B. eine bestimmte Verhaltensweise als gut oder schlecht, als vernünftig oder unvernünftig etc. klassifiziert wird. Vielmehr wird damit ausgedrückt, daß aufgrund rationaler, intersubjektiv nachvollziehbarer Kriterien auf der Basis der durch Beobachtung gewonnenen Informationen Gesamturteile, den zu analysierenden Phänomenkomplex betreffend, abgegeben werden. Daß in solchen Beurteilungen durchaus auch wertende und unbewußt verzerrende Auffassungen einfließen, soll im Abschnitt 5.1. gezeigt werden. Solche Einflüsse sind jedoch streng von dem zu trennen, was hier als bewertende Beurteilung ausgeschlossen werden sollte.

In der Gegenüberstellung von Verhaltens- und Selbstbeurteilung treffen sich auch zwei unterschiedliche Wissenschaftsauffassungen. Während die eine ihren Gegenstand ausschließlich in den objektiv beobachtbaren Verhaltensweisen sieht, richtet die andere ihr Interesse insbesondere auf psychische Phänomene, die nicht unmittelbar beobachtbar erscheinen. In der ersten Position wird versucht, innerpsychisches Erleben durch äußere Verhaltensweisen zu erschließen und zu beurteilen. Daß solche Schlußfolgerungen problematisch sind, ergibt sich schon daraus, daß inneres Erleben und äußere Verhaltensweisen nicht notwendigerweise aufeinander abgestellt sein müssen. Die Schlußfolgerung jedoch, daß auf die Erfassung von nicht unmittelbar beobachtbaren psychischen Phänomenen wegen der Gefährdung der Objektivität und Zuverlässigkeit verzichtet werden solle, wäre jedoch falsch, weil damit der Erkenntnishorizont auf overtes Verhalten eingeschränkt und Erkenntnisrestriktionen hingenommen würden. Selbstbeobachtung ist daher wie die Fremdbeobachtung als Möglichkeit wissenschaftlicher Erkenntnis anzuerkennen, wenngleich die Selbstbeobachtung im Hinblick auf ihren Erkenntniswert kritischer betrachtet werden muß.

5.1 Beurteilung nach verschiedenen Bezugspunkten

Aus der sozialpsychologischen Wahrnehmungsforschung wissen wir, daß die zu beobachtenden Sachverhalte nicht als solche sondern immer nur durch ein vorstrukturiertes Raster perzipiert werden können (vgl. hierzu 5.4.) Dies hat zur Folge, daß die Beurteilung der beobachteten Phänomene je nach Wahrnehmungsraster unterschiedlich ausfällt. Generell wird man sagen können, daß die Wahrnehmung, die Beobachtung und die Beurteilung derselben durch soziale Determinanten mitbestimmt werden.

Aus den Forschungen der Ethnologie und Kulturanthropologie kennt man das Phänomen des *Ethnozentrismus*, das unmittelbar auf Beobachtung und Beurteilung übertragbar erscheint. Zwei Möglichkeiten können dabei unterschieden werden: Ethnozentrismus im engeren Sinne meint das Befangensein in der eigenen Kultur, sodaß die Beobachtung anderer Kulturen und deren Beurteilung und Interpretation nur im Bezugsystem der eigenen Kultur vorgenommen werden kann, was jedoch zu Fehldeutungen und Fehlbeurteilungen führt. Andererseits gilt auch, daß bestimmte kulturelle Selbstverständlichkeiten durch den Angehörigen dieser Kultur nicht mehr erkannt werden können, weil sie so alltäglich und selbstverständlich geworden sind, daß sich das Augenmerk nicht mehr auf diese richtet. Allgemein bezeichnet Ethnozentrismus also das Befangensein in einer Kultur und mithin die Gefahr von Beobachtungs- und Beurteilungsverzerrungen.

Die Soziologie hat gezeigt, daß *Beobachtung und Beurteilung immer nur in einem sozialen Kontext* vor sich gehen, d.h., daß allgemeinste Ordnungsprinzipien, wie Werte und Normen die Wahrnehmung und Beobachtung strukturieren. Wir sehen z.B. bestimmte Verhaltensweisen durch die Brille sozialer Normen und Werte und beurteilen sie entsprechend. Die Wert- und Normabhängigkeit selegiert, strukturiert und evaluiert Beobachtung und Beurteilung.

Die Persönlichkeitspsychologie lehrt, daß bestimmte, *isolierte Eigenschaften konsistent mit einem Persönlichkeitssystem gesehen und beurteilt werden,* d.h. schon die Feststellung von Eigenschaften erfolgt im Kontext einer Persönlichkeitstheorie; unstimmige Eigenschaften werden nicht wahrgenommen, bagatellisiert etc. Die *impliziten Persönlichkeitstheorien*, die insbesondere im alltäglichen Umgang Anwendung finden, sind ein beredtes Beispiel für diesen Sachverhalt.

Versucht man die oben genannten einzelwissenschaftlichen Erkenntnisse, die mehr oder weniger gut empirisch abgesichert sind, unter einen gemeinsamen, theoretischen Hut zu bringen, so wird man sagen können, daß *Beobachtung und Beurteilung nicht unabhängig von dem jeweils ihnen zugrundeliegenden Bezugssystem erfolgen können.* Dieses Bezugssystem ist immer ein *soziales,* indem es Werte, Normen, Vorstellungen, Informationen, Wissen und Erkenntnisse, je nach Typ des Bezugssystems miteinbezieht. Es ist aber auch ein *personales,* das selbst nicht unabhängig vom sozialen zu sehen ist, in dem persönlichkeitsspezifische Elemente wirksam werden. Ein solches Bezugssystem ist aber auch nicht unabhängig von dem jeweiligen *situativen* Kontext, weil die Situationen, in denen beobachtet und beurteilt werden soll, durchaus verhaltensrelevant, persönlichkeitsspezifische und soziale Faktoren modifizieren können. Diese dreifach determinierte Orientierung in der Beobachtung und Beurteilung als Bezugssystem derselben, kann Beobachtung und Beurteilung in entscheidender Weise bezugssystemspezifisch verändern und sich damit verhaltensrelevant auswirken. Dieses soll an einigen Beispielen für den Bereich der Medizin verdeutlicht werden.

Die *Selbstbeurteilung eines Patienten* kann in ihrer Zuverlässigkeit und Gültigkeit durch die verschiedensten Faktoren beeinflußt werden. So kann eine schlechte

Stimmung oder eine schlaflos verbrachte Nacht durchaus dazu führen, daß der Patient seinen Gesundheitszustand schlechter beurteilt, als er ihn beurteilen würde, wenn diese situativen Ereignisse nicht eingetreten wären. Dies bedeutet, daß in der intraindividuellen Selbstbeurteilung durchaus Brüche auftreten können (unterschiedliche Beurteilungen), die nicht auf den eigentlichen Befindlichkeiten beruhen. Oder man stellt sich z.B. einen Hypochonder vor, der Tag und Nacht in sich hineinhorcht, bei der kleinsten Unregelmäßigkeit zum Arzt läuft und eine Selbstbeurteilung seines Zustandes abgibt, die weit von der realen Situation entfernt ist. Man denke auch daran, daß der Patient seinen Zustand günstiger schildert, als er ihn faktisch verspürt, weil er möglicherweise Angst vor einer bestimmten Behandlungsmethode hat, oder weil er vielleicht bestimmte gesellschaftliche Normen so sehr internalisiert hat, daß er glaubt, man können die Schilderung seines Zustandes als Wehleidigkeit oder Überempfindlichkeit interpretieren. All diese Beispiele zeigen, daß eine Selbstbeurteilung offenkundig problembehaftet ist und durch andere Beurteilungen validiert werden müßte.

So kann man der Selbstbeurteilung des Patienten das Arzturteil gegenüberstellen. Zweifelsfrei werden in dem Vergleich beider Beurteilungen gültigere Resultate erzielt, als wenn man sich jeweils auf die einzelne Beurteilung verlassen würde, denn auch das Arzturteil ist nicht frei von äußeren Einflüssen, die oft gar nicht bewußt sind. Zunächst wird man zweifelsohne davon ausgehen können, daß das Arzturteil im wesentlichen auf jenen Erfahrungen oder Kenntnissen beruht, die er beruflich gesammelt hat; insoweit sollte das Urteil korrekt sein. Sieht man einmal von Fehlurteilen ab, die natürlich nie auszuschließen sind, so kann sich gleichwohl ein fehlerbehaftetes Urteil einschleichen. Nehmen wir den Patienten, den der Arzt als Hypochonder erkannt und durchschaut hat, so wird in dem Falle einer tatsächlich begründeten Selbstbeurteilung des Zustandes der Arzt geneigt sein, die Selbstbeurteilung als überzeichnet abzutun. Die individuellen Erfahrungen des Arztes mit diesem Patienten haben ihn zu einem scheinbar konsistenten, jedoch falschen Urteil verleitet. (Als etwas ketzerisches Beispiel für die durch Bezugssysteme verursachte Determiniertheit des Arzturteils, kann auch darauf verwiesen werden, daß bestimmte Untersuchungen, Analysen zum Zwecke der Diagnose oder bestimmte Therapien durchgeführt werden, die aufgrund des Krankheitsbildes nicht unbedingt erforderlich wären, deren Anwendung als z.B. stärker ökonomisch, als medizinisch motiviert ist.)

Auch die Rolle, die die Angehörigen eines Patienten oder andere *soziale Bezugspersonen* spielen, kann in erheblichem Ausmaße die Beurteilung determinieren. Man denke an jene Eltern, die ihre Tochter seit Monaten einer erfolglosen Intensivbehandlung unterzogen sehen, die aber gleichwohl das Bewußtsein nicht wieder erlangt und die den Arzt aus Pietät darum bitten, die Beatmungsgeräte abzustellen, während dieses der Arzt aus ethischen, moralischen, vielleicht auch aus juristischen Gründen ablehnt. Dieses Beispiel zeigt deutlich, wie Werte und Normen zueinander in Widerstreit geraten können und je nach Bezugspunkt unterschiedliche Vorstellungen favorisiert werden. Oder der noch alltäglichere Fall, wo sich ein Kind eine mittelschwere Verletzung beim Spielen zugezogen hat, die eigentlich

nach Arzturteil stationär versorgt werden sollte, die Mutter jedoch aus Mitgefühl dem Drängen des Kindes nachkommt und dafür plädiert, ambulant zu behandeln; sie werde sich zu Hause um das Kind in entsprechender Weise nach den Instruktionen des Arztes kümmern.

Die *Peer-Croup* ist die Gruppe der Gleichaltrigen, bzw. Gleichrangigen (der Begriff wurde ursprünglich nur für Jugendliche entwickelt, wird jedoch heute auch allgemein verwandt) Peer-Groups zeichnen sich normalerweise dadurch aus, daß innerhalb der Gruppen Konsens über bestimmte Normen und Wertvorstellungen herrscht, die aber durchaus von den allgemeingesellschaftlich geteilten Normen abweichen können. So kann der Ärzteschaft als Peer-Group durchaus unterstellt werden, daß sie einen gemeinsamen Vorrat an Werten und Normen hat und daher in der Beurteilung bestimmter Sachverhalte relativ übereinstimmen wird. Ähnliches gilt auch für die Peer-Group der Patienten, die durch das gemeinsame Schicksal sich zusammengehörig fühlen und ähnliche Normvorstellungen entwickeln. Im Regelfall kann davon ausgegangen werden, daß Urteile von Personen aus der Peer-Group relativ gleichlautend abgegeben werden.

5.2 Beurteilungsskalen

Die verschiedenen Positionen von Beurteilungen, die im Abschnitt 5.1. vorgestellt wurden, machen deutlich, daß für eine wissenschaftlich abgesicherte Form der Beurteilung Verfahren entwickelt werden müssen, die die jeweils unterschiedlichen Bezugssysteme der Beurteilung ausschalten mindestens jedoch ihre spezifischen Verzerrungen erkennen und reduzieren können. Hierzu sind Beurteilungsskalen entwickelt worden, die es ermöglichen, systematischere, intersubjektiv nachprüfbare und kontrollierbare Beurteilungen zu liefern.

Bei den Beurteilungsskalen unterscheidet man zwei verschiedene Techniken: *Ratingverfahren und Skalierungsverfahren*. Beides sind Techniken, die dazu dienen sollen, die Beurteilungen zu entsubjektivieren. Die Skalierungsverfahren (wie z.B. Guttmann-Skalierung, Likert-Skalierung und Thurstone-Skalierung) kommen eher in Soziologie, Sozialpsychologie und Psychologie zum Einsatz, weil sie helfen, Einstellungen zu messen. Die Ratingverfahren werden dazu verwandt, Leistungen oder andere Persönlichkeitsmerkmale zu erfassen. Wir beschränken uns in der weiteren Darstellung auf die Ratingverfahren. Sie sind einfacher als die Skalierungsverfahren zu verstehen und zu handhaben und werden im Bereich der medizinischen Psychologie und Soziologie häufig benötigt.

Bevor wir auf die Ratingmethoden eingehen, sollen noch einige methodologische Grundlagen genannt werden, die sowohl für die Ratingskalen, wie auch für die Skalierungsverfahren Gültigkeit besitzen. Beide können personenorientiert, indikatororientiert oder reaktionsorientiert messen. Eine *Personenorientierung* ist dann gegeben, wenn das Verfahren dazu dient, Aussagen über die befragten Personen zu machen, d.h. Meßwertunterschiede werden als Unterschiede zwischen den

Personen interpretiert. (Ein Arzt wird zu seiner Einstellung zur Sterbehilfe standardisiert befragt. Der erhaltene Meßwert gibt Auskunft über die Einstellung (die relative Position des Arztes auf der Skala)). Bei dem *indikatororientierten* Messen werden Merkmale von Objekten durch Personen beurteilt und skaliert. (So kann man einige Personen einen Politiker X im Hinblick auf seine Beliebtheit beurteilen lassen.) Die Meßwerte geben Auskunft (= Indikatoren) über Eigenschaften des gemessenen Objektes. *Reaktionsorientiertes* Messen liegt dann vor, wenn die Versuchspersonen und die Indikatoren zugleich auf der Basis der Skalierung gemessen und beurteilt werden. (Die Skalogrammanylyse von GUTTMAN wäre eine reaktionsorientierte Messung. Sowohl die Personen, wie auch die Indikatoren werden dabei auf der eindimensionalen Skala geordnet.)

5.2.1 Relative Beurteilungsskalen

Bei den Beurteilungsskalen differenziert man zwischen relativen und absoluten Beurteilungsskalen. *Relative Beurteilungsskalen* zeichnen sich dadurch aus, daß die gewonnenen Meßwerte nicht durch das Anlegen eines „absoluten" Maßstabes entstanden sind, sondern daß es sich um relationale Meßwerte derart handelt, daß vergleichende Beurteilungen vorgenommen wurden. (Beispiel: Die Krankenschwester X hat sich mehr um den Patienten bemüht, als die Krankenschwester Y. Dies ist eine relationale oder relative Aussage, weil ihr Aussagewert nur durch das Inbeziehungsetzen der beiden, zu beurteilenden Objekte X und Y sinnvoll interpretierbar ist.) Relative Beurteilungsskalen haben von daher zur Voraussetzung, daß mehrere Objekte in die Beurteilung einbezogen werden müssen.

Die einfachste Form einer relativen Beurteilungsskala ist der *Rangreihenvergleich oder die Rangordnung*. Bei der Rangordnung wird versucht, die relative Position der zu beurteilenden Objekte zu den anderen Objekten anzugeben. Nehmen wir an, der Patient A solle 4 Krankenschwestern im Hinblick auf deren Bemühen um die Patienten in eine Rangordnung bringen. Er wird dann z.B. sagen, die Krankenschwester X habe sich am stärksten um die Patienten bemüht, die Krankenschwester Z am zweitstärksten, Y am drittstärksten und V am wenigsten. Damit hat man eine Ordinalskala gewonnen; die Personen sind hinsichtlich eines bestimmten Merkmals (Bemühen um den Patienten) durch den Rangreihenvergleich in eine Rangordnung gebracht worden. Eine andere Vorgehensweise zur Ermittlung einer Rangordnung wäre gewesen, die Patienten aufzufordern, die zu beurteilenden Objekte im Hinblick auf die vorgegebene Dimension mit Noten zu versehen. Die Krankenschwester X hätte die Note 1, Z die Note 2, Y die Note 3 und V die Note 4 erhalten. So simpel dieses Verfahren ist, es liefert Meßwerte auf ordinalem Niveau und kann eigentlich immer eingesetzt werden. Eine Restriktion muß allerdings genannt werden: wenn zu viele Objekte beurteilt werden sollen, kann der Proband überfordert sein, weil es ihm nicht gelingt, gleichzeitig alle Objekte im Auge zu haben. Das Aufstellen der Rangordnung wird dadurch zeitlich aufwendig, intellektuell anstrengend und die Reihe höchstwahrscheinlich inkonsistent.

Insbesondere um die Inkonsistenz (die ja nicht als faktische Inkonsistenz in der

Beurteilung zu sehen, sondern durch Überforderung des Probanden entstanden war) zu vermeiden, kann man eine andere Methode der relativen Beurteilungsskalen anwenden: den *Paarvergleich.* Der Paarvergleich besteht darin, daß aus allen zu beurteilenden Objekten, alle möglichen Objektpaare gebildet und zur relativen Beurteilung dem Probanden vorgelegt werden. Um in dem Beispiel mit den Krankenschwestern zu bleiben: man könnte dem Probanden also 6 Paare vorlegen und er müßte jeweils entscheiden, welche Krankenschwester sich mehr um die Patienten bemüht hat. Allgemein ergibt sich die Zahl der zu bildenden Paare aus:

$$P = \frac{N(N-1)}{2}$$

Nehmen wir an, der Proband hätte wie im obigen Falle beurteilt, dann hätten wir folgende Relationen erhalten:

$$\begin{array}{l} X > Z \\ X > Y \\ X > V \\ Z > Y \\ Z > V \\ Y > V \\ \hline X > Z > Y > V \end{array}$$

Diese Eindeutigkeit der Relationen als Konsistenz der Beurteilungen, ist nicht immer zu erzielen. Für die Inkonsistenz von Beurteilungen können verschiedene Gründe verantwortlich gemacht werden: allein schon der, daß der Proband tatsächlich inkonsistente Beurteilungen abgibt (Inkonsistenz also nicht als Fehler zu sehen ist). Es wäre auch denkbar, daß die Zahl der Objekte so groß geworden ist, daß sie für den Probanden unüberschaubar wäre, weshalb inkonsistente Urteile zu erwarten sind. Aus der obigen Formel ist zu entnehmen, daß bei nur 10 zu beurteilenden Objekten, bereits 45 Paare gebildet werden müssen. Gerade wenn es sich um viele Objekte, damit um noch mehr Paare handelt, besteht die Gefahr, daß das Kriterium, innerhalb dessen beurteilt werden soll (die Dimension), in Vergessenheit gerät und unbewußt andere Beurteilungsdimensionen Platz greifen, sodaß Inkonsistenzen verursacht werden. Eine weitere Ursache für Inkonsistenz kann darin liegen, daß die Objekte in der Beurteilung sehr nahe beianander liegen, der Proband jedoch gezwungen ist, immer eine Präferenzentscheidung zu treffen, (Gleichheitsrelationen sind ausgeschlossen). Es werden somit künstliche Differenzen erzeugt, die zu einer Inkonsistez führen können.

Bisher waren wir beim Paarvergleich davon ausgegangen, daß ein Proband verschiedene Objekte beurteilt. Tatsächlich jedoch werden in der Regel mehrere Probanden die Objekte ordnen, d.h. es muß versucht werden, eine über alle Probanden hinweg geltende Rangreihe zu finden. Die einfachste Möglichkeit besteht nun darin, daß ausgezählt wird, wie häufig jedes Objekt in dem vorgegebenen Paar preferiert wurde. Summiert man über die einzelnen Probanden diese Werte, so erhält man

eine allgemeine und ordinale Maßzahl für die Beurteilung des jeweiligen Objektes. Hierbei ist natürlich nicht ausgeschlossen, daß zwei Objekte eine gleiche Maßzahl erhalten, sodaß die absolute Rangfolge zwar gestört, das Ergebnis jedoch immer verwertbar bleibt, weil interpretiert werden kann, daß die beiden Objekte gleichrangig beurteilt werden.

Zu den relativen Beurteilungsskalen zählt man auch die *soziometrischen Wahlverfahren*. Die Soziometrie ist eine Methode zur Erfassung und Darstellung sozialer Beziehungen in Gruppen. Sie geht auf Jakob L. MORENO zurück (1934) und wurde von ihm (da er Arzt war), auch zu therapeutischen Zwecken benutzt; auf diese soll jedoch hier nicht eingegangen werden.

Die Methode der Soziometrie ist das soziometrische Wahlverfahren oder der *soziometrische Test*. Mit seiner Hilfe ermittelt man Sympathie, Antipathie und Gleichgültigkeitsbeziehungen innerhalb einer Gruppe auf vorzugebenden Dimensionen. Die Gruppenmitglieder werden aufgefordert, im Hinblick auf die angegebene Dimension andere Gruppenmitglieder zu wählen, die ihnen sympathisch, unsympathisch oder gleichgültig sind. Diese Wahlverfahren können auf den unterschiedlichsten Techniken beruhen:

1. Es kann eine *bestimmte Anzahl von Wahlen vorgegeben* werden, oder aber *die Zahl der Wahlen bleibt unbegrenzt*. Werden Wahlbegrenzungen vorgenommen, so kann man folgende Möglichkeiten wählen:
   a. Das *Maximumlimit* (ceiling limit); es fordert die Angabe einer bestimmten Anzahl von Wahlen oder weniger, auf keinen Fall jedoch mehr.
   b. Das *Minimumlimit* (Floorlimit); es verlangt eine bestimmte Anzahl von Wahlen oder mehr, auf keinen Fall jedoch weniger.
   c. Das *Maximum-Minimumlimit*, hier wird eine bestimmte Zahl vorgegeben, die weder über- noch unterschritten werden darf.
   d. Das *Cirkalimit*, hier wird eine bestimmte Anzahl von Wahlen vorgeschrieben, Variationen nach oben und unten sind jedoch möglich.

Bei der Benutzung dieser Wahlmethoden muß allerdings theoretisch bedacht werden, daß jede Methode unterschiedliche Erkenntnismöglichkeiten eröffnet und verschiedene Erkenntnisgrenzen hinnimmt. So wäre z.B. denkbar, daß in einer Gruppe von 20 Personen bei einer begrenzten Wahlzahl von 2 Personen, eine Person X als von der Gruppe ausgeschlossen, als isoliert erscheint; läßt man jedoch eine dritte Wahl zu, kann dieselbe Person plötzlich voll in die Gruppe integriert sein.

2. Neben der freien Wahl und der einfachen Begrenzung der Wahlmöglichkeiten, kann man auch nach einer *Rangordnung* fragen. („Wen möchtest du am liebsten, am zweitliebsten...") Es wäre aber auch möglich, statt einer Rangordnung die Möglichkeit zu eröffnen, im Hinblick auf die erfragte Dimension bestimmte *Punktwerte* vergeben zu können, z.B. von −10 bis +10. Auch hiermit sind theoretische Implikationen verknüpft, die vor der Anwendung dieser Verfahren durchdacht werden müssen.

3. Es ist zu unterscheiden zwischen *positiven und negativen Wahlen* (Sympathie und Antipathie). Während die soziometrischen Wahlverfahren zunächst auf Sympathie ausgerichtet waren, hat man mehr und mehr auch Antipathiewahlen vorgenommen. In der Literatur wird gelegentlich vor negativen Wahlen gewarnt, weil damit das Gruppenklima tangiert werden kann. Andererseits bedeutet jedoch der Verzicht auf negative Wahlen eventuell eine unvollständige und unrealistische Erhebung und damit mangelhafte Interpretation der Gruppensituation und Gruppenstruktur.

4. Die beim soziometrischen Test vorzugebenden Wahlfragen, können *indikativ oder konjunktiv* formuliert werden. Sind sie indikativ formuliert, dann handelt es sich um eine Bestandsaufnahme der realen Gruppensituation. Sind sie konjunktiv formuliert, so stellen sie eher auf Intentionen, Wünsche und Vorstellungen als auf die Realität selbst ab. Sind gewis-

se reale Strukturen bekannt, so kann mit Hilfe der hypothetisch-konjunktiven Form aus einer auftretenden Diskrepanz auf Probleme innerhalb dieser Gruppe geschlossen werden.

5. Um die Beurteilungsfähigkeit der einzelnen Gruppenmitglieder überprüfen zu können, hat man den soziometrischen Test auf *vermutete Wahlen* angewandt. Man fragt dabei danach, von wem man glaubt, gewählt worden zu sein oder wer einen vermutlich abgelehnt hat. Aus der Übereinstimmung oder Nichtübereinstimmung mit den tatsächlichen Wahlen, kann dann auf eine gültige Selbst- und Gruppenbeurteilung geschlossen werden.

Die mit Hilfe des soziometrischen Wahlverfahrens ermittelten Sympathie-, Antipathie- und Gleichgültigkeitsstrukturen innerhalb von Gruppen, können auf die unterschiedlichste Weise ausgewertet werden. So kann man die Ergebnisse in Form einer *Soziomatrix* (Kreuztabelle) anordnen und mit Hilfe der schon bekannten statistischen Maßzahlen einige Aussagen und Interpretationen vornehmen. Aus einer solchen Tabelle, wie auch aus den Rohwerten, können *Sozioindizes* gebildet werden. Hat man z.B. die Wahlmöglichkeiten in der Zahl nicht begrenzt, so können die abgegebenen positven Wahlen als Index für die *aktive Soziabilität* des Gruppenmitgliedes herangezogen werden. Auch kann die Summe der erhaltenen Wahlen auf den Grad der Akzeptierung eines bestimmten Gruppenmitgliedes durch die Gruppe hinweisen.

Tabelle 30: **Soziomatrix:** ***Interpersonelle Attitüden innerhalb einer Kleingruppe***

(Die Vpn hatten 5 Kategorien zu vergeben: von −1 über −0,5; +0,5 zu +1; Beurteilungsdimension: Grad der Akzeptierung)

gegeben Wahlen / Wahlen erhalten	A	B	C	D	E	F	G	Σ erhalten
A	–	1	0	0	0	0,5	1	2,5
B	1	–	1	1	0,5	0,5	0,5	4,5
C	0	0	–	0	0	0,5	0	0,5
D	0,5	0	0,5	–	0,5	1	0,5	3,0
E	0,5	−1	0	0,5	–	1	0	1,0
F	0	0,5	−1	−1	0	–	0	−1,5
G	0	0	−0,5	0	0,5	0,5	–	0,5
Σ erhalten	2.0	0,5	0	0,5	1,5	4,0	2,0	

Für die obige Tabelle seien mit Hilfe einfachster statistischer Maßzahlen (Standardabweichung und arithmetisches Mittel) einige Sozioindizes berechnet:

Die passive Soziabilität (Akzeptierung durch andere)

$$\bar{X}_{\text{erh. Noten}} = \frac{\Sigma xi}{N}$$

A = 0,41 E = 0,17
B = 0,75 F = −0,25
C = 0,08 G = 0,08
D = 0,50

Grad der Übereinstimmung der Beurteilung

$$\sigma = \sqrt{\frac{\Sigma(xi - x)^2}{N}}$$

A = 0,42 E = 0,50
B = 0,25 F = 0,50
C = 0,14 G = 0,25
D = 0,17

Die aktive Sozialibilität

$$\bar{X}_{\text{gegeb. Noten}} = \frac{\Sigma xi}{N}$$

A = 0,33 E = 0,25
B = 0,08 F = 0,67
C = 0,00 G = 0,33
D = 0,08

Grad der Diskriminierung

$$\sigma = \sqrt{\frac{\Sigma(xi - \bar{x})^2}{N}}$$

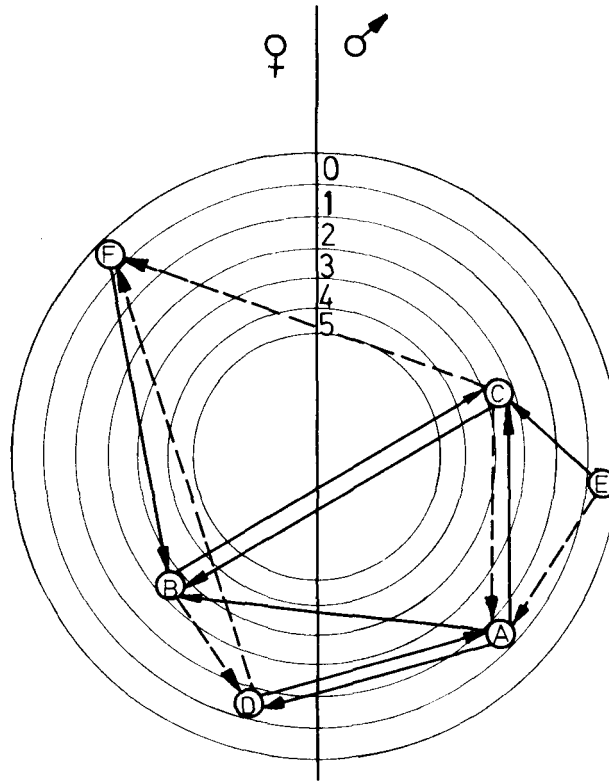
A = 0,33 E = 0,25
B = 0,43 F = 0,22
C = 0,50 G = 0,33
D = 0,43

Am deutlichsten wird die über soziometrische Wahlverfahren ermittelte Gruppenstruktur dann, wenn die Konfiguration der Gruppe als *Soziogramm* gezeichnet wird. Aus der Gruppenkonstellation im Soziogramm kann man mögliche Konflikte diagnostizieren und Therapiemöglichkeiten entwickeln. *Das Soziogramm ist die grafische Darstellung der Ergebnisse des soziometrischen Tests.* In der Analyse von Soziogrammen haben sich bestimmte Konfigurationen als häufig und relevant herausgestellt, sodaß diesen bestimmte Bezeichnungen zugewiesen wurden, wie z.B. Paar, Kette, Stern, Star, Clique, Isolierter, graue Eminenz usw.

In den soziometrischen Wahlverfahren können auch der Rangreihenvergleich oder der Paarvergleich angewandt werden. Daraus kann abgeleitet werden, daß die soziometrischen Wahlverfahren einerseits ein Spezialfall der relativen Beurteilungsskalen darstellen, weil allgemein entwickelte Beurteilungsskalen in ihnen angewandt werden, andererseits sind sie jedoch allgemeiner, weil sie sich eher theoretisch als methodisch orientieren.

5.2.2 Absolute Beurteilungsskalen

Absolute Beurteilungsskalen sind solche, wo bestimmte Objekte anhand eines vorgegebenen „Maßstabes" – ohne daß vergleichende Objekte herangezogen würden – zu beurteilen sind. Die Versuchsperson schätzt das Objekt, die Eigenschaft des Objektes oder dessen Merkmalsausprägungen unter Zuhilfenahme von Skalen, die als Maßstab fungieren. Solche absolute Beurteilungsmethoden bezeichnet man daher als *Schätzskalen* oder *Ratingskalen.* So kann der Patient danach gefragt werden, ob er die Krankenschwester X als „sehr freundlich", „freundlich" oder „weniger freundlich" beurteilt. Maßstab sind dabei die vorgegebenen Antwortkatego-



Legende:

– – –	Ablehnung
———	Wahlen
0 – 5	Anzahl der erhaltenen Wahlen
♀	weibliche Gruppenmitglieder
♂	männliche Gruppenmitglieder

Beim Target-Soziogramm (= Schießscheibensoziogramm) wird versucht, eine Struktur in das Soziogramm zu bringen, indem je nach Häufigkeit der Wahlen die Gruppenmitglieder auf den Ringen der Scheibe plaziert werden.

Die Gruppe besteht aus drei weiblichen (F, B, D) und drei männlichen (C, A, E) Gruppenmitgliedern.

D, A, B, C bilden eine Kette.

F erhält nur Ablehnungen, während E weder gewählt noch abgelehnt wird. Beide sind in der Gruppe isoliert.

D und A wählen sich gegenseitig, erhalten aber duch andere Gruppenmitglieder keine Wahlen, sondern nur Ablehnungen.

Man kann A und D als Paar bezeichnen.

B und C erhalten die meisten Wahlen (je drei) und keine Ablehnungen.
Sie wählen sich gegenseitig, aber keine anderen Gruppenmitglieder.
Die weiblichen Gruppenmitglieder sind untereinander ebenso wenig verbunden wie die männlichen.

Die Gruppe wird durch B und C zusammengehalten.

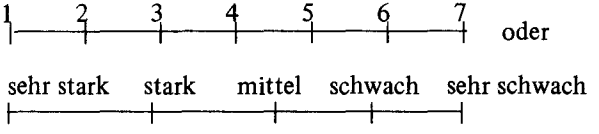
(RATHGEBER Walter, (Hrsg.), Medizinische Psychologie, München 1977, 2. Aufl.)

Abbildung 21: Soziogramm, hier Target- oder Schießscheibensoziogramm

rien. Dieser ist absolut, weil keine Vergleichsobjekte zur Verfügung stehen, die seine Richtigkeit validieren könnten. So ist durchaus denkbar, daß zwei verschiedene Versuchspersonen die Antwortkategorie „sehr freundlich" qualitativ und quantitativ unterschiedlich interpretieren, sodaß sie zu unterschiedlichen Aussagen über die Krankenschwester X kommen, obwohl sie im Hinblick auf deren Verhalten und dessen Beurteilung einer Meinung sind.

Je nachdem, wie ein solcher Maßstab für Schätzskalen konstruiert wird, lassen sich unterschiedliche Ratingskalen unterscheiden (vgl. hierzu auch GUILFORD, J.P., Psychometric Methods, New York, Toronto, London 1954, 2. Aufl. S.236–301). Die wohl bekannteste Form der Schätzskala ist die *numerische Ratingskala*. Bei der numerischen Schätzskala wird jedem Objekt oder jedem Merkmal durch den Probanden ein bestimmter numerischer Wert zugewiesen, so z.B. der minimale Wert 1, wenn eine Merkmalsausprägung nur äußerst schwach ist und der Maximalwert 7, wenn ein Merkmal stark ausgeprägt ist. Die Schätzskala besteht also aus den Werten 1–7. (Selbstverständlich kann auch jeder andere Maßstab, z.B. Werte 1–9 oder 0–4 gewählt werden. Zu berücksichtigen bleibt dabei nur, daß eine solche Schätzskala symmetrisch aufgebaut werden sollte, damit nicht schon durch die Vorgabe des Maßstabes positive oder negative Verzerrungen eintreten).

Eine numerische Schätzskala kann man auch grafisch anordnen, indem man eine Gerade analog dem entwickelten Maßstab in einzelne Abschnitte unterteilt und diese Abschnitte mit den Zahlen (exakter: Ziffern) der numerischen Schätzskala oder mit qualitativen verbalen Erläuterungen kennzeichnet. Man spricht dann von *grafischen Schätzskalen.*



Bei der *Standardschätzskala* wird nun, nicht wie bei der numerischen und grafischen Ratingskala nur ein Maßstab (eine Gerade) vorgegeben, sondern es wird versucht ein Objekt im Hinblick darauf beurteilen zu lassen, inwieweit es mit mehreren vorgegebenen Standardbeispielen übereinstimmt. Der Proband erhält also eine bestimmte Anzahl von Beispielen (Statements oder Items) und der Proband sucht sich dasjenige aus, das nach seiner Auffassung das zu beurteilende Objekt am besten charakterisiert. Sind diese Standardbeispiele auf einer Liste aufgeführt, so spricht man von einer *Case-List*, aus der das für das Untersuchungsobjekt passende Beispiel ausgewählt werden soll.

Läßt man aus diesen vorgegebenen Standardbeispielen nicht nur eine Wahlmöglichkeit zu, sondern verlangt von dem Probanden, daß er die Beispiele in Bezug auf das Untersuchungsobjekt in eine Rangfolge bringt, so handelt es sich um ein *Sortierverfahren.*

Hat man mehrere Schätzskalen, die alle bezüglich einer Dimension für einen Ob-

jektbereich entwickelt worden sind, und wählt man zur Charakterisierung dieses Objektes nicht nur eine Schätzskala aus, sondern zieht zur Auswertung alle Schätzskalen heran, so spricht man von *kumulierten Punktskalen* oder *Punktsummenskalen.* Hierbei werden also die einzelnen Punktwerte der Skalen aufsummiert und die Summe der Punktwerte ergibt die Maßzahl für das zu beurteilende Objekt. Voraussetzung für die Anwendung der Punktsummenskala ist deren Eindimensionalität, d.h.: jede einzelne Skala muß auf der Dimension des zu messenden Merkmals oder Objekts liegen, (weil man ja schließlich nicht „Äpfel und Birnen" addieren kann). Die Richtung der einzelnen Skalen muß auch jeweils identisch sein (niedrige Werte müssen immer gleichgerichtet bei jeder Skala niedrigen oder hohen Ziffern entsprechen), weil sonst die Summierung der Skalenwerte sinnlos wird und den Meßwert eher verschleiert, als erhellt.

5.3 Systematische Beurteilungseffekte

Bei der Verhaltensbeurteilung hat man durch eine Fülle von Untersuchungen spezifische Fehlerquellen, die in systematischer Weise auf die Untersuchungsergebnisse einwirken, erkannt. Wie schon in Abschnitt 5.1. einleitend dargestellt wurde, spielen für die Verhaltensbeurteilung soziologische, psychologische und kulturspezifische Variablen eine entscheidende Rolle. Neben diesen allgemeinen Erkenntnissen sind eine Fülle von spezifischen, systematisch verzerrenden Faktoren entdeckt worden, über deren Vorhandensein und Wirksamkeit der Forscher Bescheid wissen muß, ist ihm daran gelegen, gültige Ergebnisse zu erzielen. Das Wissen um solche systematischen Beurteilungseffekte, stellt einen ersten Schritt in Richtung auf eine Objektivierung der Beobachtungs- und Beurteilungsergebnisse im Sinne einer Reduzierung oder gar Vermeidung solcher Fehlerquellen dar.

Der *Haloeffekt* ist eine bedeutsame Fehlerquelle der Verhaltensbeurteilung. *Er meint die Beurteilung einzelner Eigenschaften, oder Merkmalsausprägungen im Lichte schon bekannter anderer Eigenschaften oder eines Gesamtbildes.* Dieses Effekt tritt bei den Ratingmethoden relativ häufig auf, weil die Beurteilung nachfolgender Items durch die vorausgegangenen mehr oder weniger stark beeinflußt wird. Allgemein wirkt der Haloeffekt im Sinne einer Tendenz zur Vereinfachung, zur Konsistenz und zur Stabilität. Da die einzelnen Items nicht unabhängig voneinander gesehen werden, tritt durch den Haloeffekt eine Pseudokonsistenz und Pseudostabilität der Ergebnisse auf, die sich auch in hohen Korrelationskoeffizienten niederschlagen.

Gerade in der Persönlichkeitsbeurteilung ist der Haloeffekt nur schwer auszuschließen. Aufgrund des Vorhandenseins eines Merkmals, wird auf das Vorliegen weiterer Merkmale und einer bestimmten Persönlichkeit mittels impliziter Persönlichkeitstheorien, die sich aufgrund eines dominanten Merkmals ergeben, geschlossen. Der Haloeffekt beruht also auf einer suggestiven Wirkung von Merkmalen, die als zusammengehörig assoziiert werden, was jedoch nicht inhaltlich gerechtfertigt sein muß. Er gilt für Probanden, die in einem Fragebogen sich selbst zu beurteilen ha-

ben, genauso, wie für „neutrale" Beobachter, die bestimmte Verhaltensweisen oder Eigenschaften an anderen Personen feststellen sollen. Gerade wenn dem Beobachter schon Informationen über den zu Beobachtenden vorliegen, bildet sich aufgrund dieser Informationen eine bestimmte Erwartungshaltung heraus, die solche Verhaltensweisen und Eigenschaften des zu Beurteilenden im Sinne der Konsistenz stärker in den Vordergrund treten läßt und die vernachlässigt, die nicht mit dem Erwartungshorizont vereinbar sind. Wichtig ist beim Haloeffekt festzuhalten:

1. Seine Wirkung besteht darin, daß eine scheinbare Stabilität und Konsistenz durch ein einheitliches Bild geschaffen wird, das aber eher durch systematische Fehler, als durch die Realität selbst entstanden ist;
2. und daß solche Wirkungen durch die unbewußte „Abstimmung" von Verhaltensweisen und Eigenschaften auf der Basis eines vorgefaßten Gesamtbildes oder vorhergegangener, einzelner Merkmale und Eigenschaften eintreten.

Unter die *Einstellungsfehler* werden die *Kontrast- bzw. Ähnlichkeitsfehler* subsumiert. Bei diesen wird je nach Situation die *Beobachtung bzw. Beurteilung in bewußter oder unbewußter Abhängigkeit von der eigenen Person vorgenommen.* Der Ähnlichkeitsfehler besteht nun darin, daß der zu beurteilenden Person ähnliche Eigenschaften zugeschrieben werden, wie man sie bei sich selbst vorzufinden glaubt, während beim Kontrastfehler der Beurteilende davon ausgeht, daß die zu beurteilende Person ganz anders ist, als er selbst. Dieser Kontrasteffekt ist nicht nur im wissenschaftlichen sondern auch im Alltagsbereich häufig vorzufinden. So konnte man z.B. nachweisen, daß solche Beamte, die einen Bewährungsaufstieg (außerhalb der „normalen" Laufbahn) geschafft haben, besonders rigide und „scharf" sind und anderen den potentiellen Bewährungsaufstieg nicht gönnen. Auf Persönlichkeitseigenschaften bezogen, läßt sich auch sehr oft feststellen, daß ein äußerst zuverlässiger, pünktlicher und genauer Beurteiler den zu Beurteilenden an seinen eigenen Maßstäben mißt und ihm attestiert, er wäre unordentlich, schlampig und ungenau. Da der Ähnlichkeitseffekt und der Kontrasteffekt beide auf der gleichen Basis beruhen, nämlich Einstellungseffekte sind, muß im konkreten Einzelfalle jeweils überprüft werden, in welcher Richtung der Einstellungseffekt gewirkt hat.

Der *leniency error* oder *generosity error* ist der *Mildefehler*. In ihm kommt die Tendenz zum Ausdruck, nicht die Sachverhalte als solche wahrzunehmen und zu beurteilen, sondern sie immer in ein günstigeres Licht zu rücken, Milde walten zu lassen. Die Ursachen für den Mildefehler können vielfältiger Natur sein. Ein äußeres Erscheinungsbild eines Probanden, das als sympathisch empfunden wird, kann eine generell positive Grundstimmung in dem Beurteilenden erzeugen, sodaß er nicht abgeneigt ist, Milde walten zu lassen. So wird z.B. auch der psychiatrische Gerichtsgutachter im Zweifelsfalle bei seinem Gutachten mit in Betracht ziehen, welche Konsequenzen sein Urteil für den Probanden hat. Glaubt er, daß eine psychiatrische Behandlung wirksamer und weniger persönlichkeitszerstörend ist, als eine langjährige Inhaftierung, so wird sein Gutachten zwar nicht unabhängig von den Tatsachen gefällt werden, doch immerhin mit einer Tendenz zur Milde (= psychiatrische Behandlung) versehen sein.

Projektion bei der Verhaltensbeurteilung meint die Übertragung eigener Wünsche

oder auch eigener Fehler auf andere, sodaß unerkannte Projektion als erhebliche Fehlerquelle auftreten kann. Der Begriff der Projektion stammt aus der Psychoanalyse und bezeichnet das Wirksamwerden eines Abwehrmechanismus, der dafür sorgt, daß eigene Gefühle anderen Personen zugeschrieben werden. Verspürt ein Beurteiler beispielsweise einen starken Sexualtrieb, den er aber aufgrund des ausgeprägten Über-Ichs nicht bei sich akzeptieren kann, weil die gesellschaftlichen Normen dem Ausleben dieses Triebes entgegenstehen, so wird dieser interne Konflikt allzu leicht auf andere Personen, die zu beurteilen sind, übertragen; das kleinste Indiz genügt als Auslöser und man schreibt dem zu Beurteilenden die eigenen Sexualtriebe zu.

Von diesem Begriff der Projektion, der in der Psychoanalyse als theoretischer Begriff durchaus erklärungskräftig ist, der bei der Verhaltensbeurteilung aber als Fehlerquelle wirkt, ist der Projektionsbegriff zu trennen, wie er in dem Kapitel zu den Testverfahren entwickelt wurde. Dort war darunter verstanden worden: „Methoden, welche die Persönlichkeit dadurch untersuchen, daß sie die Versuchsperson einer Situation gegenüberstellen, auf welche die Versuchsperson entsprechend der Bedeutung reagiert, die diese Situation für sie besitzt. Das Wesen eines projektiven Verfahrens liegt darin, daß es etwas hervorruft, was – auf verschiedene Art – Ausdruck der Eigenwelt des Persönlichkeitsprozesses der Versuchsperson ist". (nach FRANK. L.K., Projective Methods, Springfield 1948, S.46f.). Der klassische Projektionsbegriff, aus dem die Projektion als Fehlerquelle abgeleitet ist, meint hingegen, die Übertragung (Zuschreibung) eigener unangenehmer Eigenschaften auf Objekte der Außenwelt (im Regelfalle Personen).

Der Fehler der *zentralen Tendenz,* tritt insbesondere dort auf, wo bestimmte Eigenschaften skaliert werden sollen, (sei es über Rating- oder Skalierungsverfahren). Er meint, daß die von den Probanden geforderten Selbsturteile über Eigenschaften oder Sachverhalte, aber auch die vorzunehmenden Fremdbeurteilungen die *Tendenz aufweisen, gehäuft auf die Skalenmitte zu fallen, extreme Ausprägungen also zu vermeiden.* Diese Tendenz entspringt einerseits einem Mangel an Differenziertheit des Urteils, andererseits aber auch der Unsicherheit über das eigene Urteil. Drittens kann man anführen, daß gerade dann, wenn die soziale Wünschbarkeit bestimmter Eigenschaften nicht eindeutig ist, gehäuft mittlere Positionen für die Beurteilung herangezogen werden. Die Scheu vor Extremurteilen führt dazu, grau in grau zu malen (Man denke z.B. an den Deutschlehrer, der unsicher darüber ist, welche Note er einem Schüler auf einen Aufsatz geben soll. Er wird den geringsmöglichen Fehler begehen, wenn er eine mittlere Note, vielleicht „befriedigend" wählt). Während in dem Beispiel der mittlere Wert durchaus eindeutig ausgelegt werden kann, ist das bei den Skalen nicht immer möglich. Da der mittlere Wert dort als neutral, indifferent, ambivalent usw. interpretiert werden kann, ist der Fehler der zentralen Tendenz besonders gravierend, weil eine eindeutige Entscheidbarkeit nicht gewährleistet ist.

Als letzte Fehlerquelle soll der sog. Primacy-Effekt angesprochen werden. Dieser *Fehler des ersten Eindrucks* wirkt sich so aus, daß alle weiteren, zu beobachtenden Verhaltensweisen im Lichte eines ersten Eindrucks beurteilt und interpretiert wer-

den. Man kann daher davon ausgehen, daß das Vorurteil oder Vorwegurteil des ersten Eindrucks als self-fulfilling prophecy derart wirkt, daß dieser erste Eindruck durch weitere Verhaltensweisen bestätigt zu werden scheint. Wenn aber zeitlich vorausgehende Beobachtungen und Beurteilungen alle nachfolgenden in gleichem Sinne, jedoch ungültigerweise bestimmen, dann wirkt sich der primacy error wieder in einer scheinbaren Stabilität und Konsistenz aus, deren Erkennen schwierig wird. Während nämlich inkonsistente Ergebnisse den Forscher dazu veranlassen, darüber nachzudenken, wie solche Inkonsistenzen inhaltlich oder methodisch entstanden sein könnten, suggeriert Konsistenz die Richtigkeit der Resultate, wodurch das Entdecken von Fehlerquellen erschwert wird.

5.4 Fehlerquellen der Selbstbeurteilung

Die im Abschnitt 5.3. genannten Beurteilungseffekte, können sowohl für die Verhaltensbeurteilung, wie für die Selbstbeurteilung relevant werden. Bei der Selbstbeurteilung gilt jedoch darüber hinaus, daß weitere Probleme mit ihr verbunden sind, die gelegentlich dazu geführt haben, daß jeglicher Verzicht auf Selbstbeurteilung aus methodisch-methodologischen Gründen gefordert wurde. Wie jedoch schon eingangs erwähnt, würde ein solcher Verzicht einen erheblichen Erkenntnisverlust darstellen. Auch wenn die Erkenntnisse, die durch Introspektion gewonnen werden, nicht jenes Maß an Gültigkeit und Zuverlässigkeit besitzen, wie die Fremdbeobachtung und -beurteilung kann auf sie nicht verzichtet werden. Bestimmte Sachverhalte sind eben der wissenschaftlichen Beobachtung als Fremdbeobachtung völlig entzogen und müssen notwendigerweise durch Introspektion erhoben werden. Dies gilt für alle „inneren Verhaltensweisen", die weder direkt noch indirekt empirisch erfaßbar sind. So ist die Medizin noch nicht so weit, Schmerz objektiv (und im Extremfall skalierbar) zu ermitteln. Vielmehr ist man auf die Mitteilung der subjektiven Empfindungen der Patienten angewiesen. Hierbei versteht sich von selbst, daß das Fehlen eines absoluten Maßstabes notwendigerweise dazu führen muß, daß gleiche Schmerzqualitäten von einzelnen Patienten unterschiedlich wahrgenommen und mitgeteilt werden. Bestimmte innerpsychische Zustände, Affekte, Gefühle etc. sind auch nicht unmittelbar wahrnehmbar und bedürfen daher der notwendig subjektiven Verbalisierung. (So ist es einem Menschen nicht notwendigerweise anzusehen, ob er z.B. einen anderen liebt. Zwar gibt es durchaus bestimmte Verhaltensweisen, mit deren Hilfe auf solche Gefühle geschlossen werden kann, doch im Regelfalle müssen sie in Form der Verbalisierung der Introspektion mitgeteilt werden. Daß solche Mitteilungen nicht in jedem Falle den Gefühlen entsprechen müssen, liegt auf der Hand.)

Die in der Sozialpsychologie ermittelten Phänomene der sozialen Wahrnehmung (vgl. RATHGEBER, W. (Hrsg.), Medizinische Psychologie, München 1977, 2. Aufl. S. 261ff), wie Wahrnehmungsabwehr, Wahrnehmungsakzentuierung usw. sind gute Belege dafür, daß scheinbar objektive Wahrnehmungen subjektive Veränderung erfahren, die sozial und personal bedingt sind. Deutlich wird auch die Notwendig-

keit, auf durch Introspektion gewonnene Informationen zurückzugreifen, wenn es sich um nur vorgestellte, z.B. krankhaft bedingte Wahrnehmungen handelt. Ein sehr ängstlicher Mensch, der einen Abendspaziergang im Park unternimmt, wird möglicherweise eine bestimmte Konfiguration von Ästen und Blättern als menschliche Gestalt wahrnehmen und keinen Zweifel an der Richtigkeit seiner Wahrnehmung haben. Solche Trugwahrnehmungen müssen nicht notwendigerweise auf krankhaften Phobien beruhen. Ursache dafür können auch stark ausgeprägte Vorstellungskraft, lebhafte Phantasie oder die äußerst zielgerichtete selektive Wahrnehmung sein. Wahnvorstellungen sind der typische Fall dafür, daß subjektive Wahrnehmungserlebnisse vorliegen, die jeglicher objektiver Grundlagen entbehren. So kann ein unter Wahnvorstellungen Leidender in einem völlig leeren Raum sich befinden und gleichwohl das subjektive Gefühl haben, daß in diesem Raum Menschen sind, die ihn verfolgen. Gerade an den Wahnvorstellungen läßt sich ablesen, daß Introspektion ein notwendiges Hilfsmittel einer objektiven Beobachtung sein kann. Seit wir nämlich durch das *Thomastheorem* wissen, daß für menschliches Verhalten nicht die Dinge an sich so sehr entscheidend sind, als vielmehr die als real vorgestellten und geglaubten Dinge („If people think, something is real, then it is real in it's consequences"), haben wir die Legitimation dafür erhalten, sich in der wissenschaftlichen Arbeit nicht allein auf das objektiv Festell- und Meßbare zu beschränken, sondern gerade auch die subjektiven von den objektiven Tatsachen abweichenden Elemente in die wissenschaftlichen Überlegungen einzubeziehen. Es zeigt sich somit, daß ein „Entweder-Oder" zwischen Fremdbeobachtung und Introspektion durch ein „Sowohl-als-auch" ersetzt werden muß.

In den genannten Beispielen für Introspektionen als Selbstbeobachtung und Selbstbeurteilung, konnte man (ähnlich, wie bei der Gegenüberstellung von theoretischen - und Beobachtungsbegriffen) ein Kontinuum angeben zwischen solchen Phänomenen, die ausschließlich durch Introspektion erfaßbar erscheinen und solchen, die sowohl introspektiv, wie auch annähernd durch Fremdbeobachtung festgestellt werden können. Daher sollte die Ausschließlichkeit zugunsten einer Gemeinsamkeit aufgehoben werden, auch wenn die Introspektion einige zusätzliche Fehlerquellen beinhaltet.

Jede Selbstbeobachtung und Selbstbeurteilung ist, wenn sie wissenschaftlich nutzbar gemacht werden soll, darauf angewiesen, daß eine *Mitteilungsbereitschaft* bei dem Probanden existiert, oder aber geweckt wird. Mitteilungsbereitschaft meint hier zweierlei: einmal, die Motivation, sich anderen mitzuteilen und andererseits, in der Konkretion dieser abstrakten Mitteilungsbereitschaft, tatsächlich das mitzuteilen, was im inneren Erleben sich relevant abgespielt hat. (So setzt z.B. der Gang zum Psychoanalytiker einerseits die Entscheidung voraus, sich mitteilen zu wollen, ein entsprechendes Bedürfnis entwickelt zu haben. Dieses allein hilft allerdings nicht weiter, wenn der Patient nicht willens ist, alle relevanten Sachverhalte und alle richtig (ohne Beschönigung, ohne Übertreibung etc.) mitzuteilen.)

Setzen wir abstrakte und konkrete Mitteilungsbereitschaft voraus, so bleibt das für die Introspektion weit schwerwiegendere Problem der *Mitteilungsfähigkeit* der

Probanden in seiner absoluten Schärfe bestehen. Man kann eindeutig interindividuelle und intraindividuelle Unterschiede der Mitteilungsfähigkeit feststellen. Die Person X vermag etwa das Schmerzempfinden sehr spezifisch zu lokalisieren und im Ausprägungsgrad adäquat wiederzugeben, während bei gleicher Symptomatik die Person Y nur relativ diffuse Angaben darüber machen kann. Die Person Z ist durchaus in der Lage, sehr exakt ihr Schmerzempfinden wiederzugeben, sieht sich jedoch außerstande, positive Gefühle zu beschreiben.

Neben anderen Fehlerquellen in der Mitteilungsfähigkeit, wie z.B. Scham, Angst, Imponiergehabe, etc. spielt gerade die Variable des *Verbalisierungsvermögens* innerer Erlebnisse und Zustände bei der Introspektion eine entscheidende Rolle. Sprachlich Begabte, den sog. „elaborated code" sprechende Angehörige der Oberschicht, werden eher in der Lage sein, differenzierte Stellungnahmen aufgrund von Introspektion zu liefern, als den „restricted code" sprechende Angehörige der Unterschicht. Mitteilungsbereitschaft und Mitteilungsfähigkeit sind also jene theoretischen Konzepte, die bei dem Verfahren der Introspektion in besonderer Weise zusätzlich berücksichtigt werden müssen, sollen Fehlinterpretationen und falsche Schlußfolgerungen vermieden werden. Diese zusätzlichen Fehlerquellen der Introspektion können jedoch nicht dazu herangezogen werden, die Introspektion als nicht erklärungskräftig und wissenschaftlich unbrauchbar abzulehnen. Vielmehr ergänzt die Selbstbeurteilung die Fremdbeurteilung und Fremdbeobachtung. Wie die Geschichte der Psychologie gezeigt hat, wurde die Introspektion häufig fruchtbar eingesetzt; so wäre manche Persönlichkeitstheorie ohne die Introspektion sicher nicht möglich gewesen.

LITERATURVERZEICHNIS

Das Literaturverzeichnis ist nicht nach Sachgebieten sondern alphabetisch geordnet. Aus den jeweiligen Titeln ist jedoch unschwer zu entnehmen, welchem Sachgebiet diese zuzurechnen sind. Die Liste der Titel ist auch ausführlicher, als es die Literaturangaben im Text sind, um dem Leser ein breiteres Spektrum anzubieten. Gleichwohl handelt es sich immer um Grundlagenliteratur, die über die hier gegebene Einführung hinausgeht.

ALEMANN, H.v. Der Forschungsprozeß, Stuttgart 1977

BARTON, A.H., Asking the Embarrassing Question, in: Public Opinion Quarterly, 22, 1958

BASTIN, G., Die soziometrischen Methoden, Bern und Stuttgart 1967

BELSER, H., Testentwicklung, Weinheim 1972

BJERSTEDT, A., The Methodology of Preferential Sociometry (Sociometry Monographs, No. 37), Lund 1956

BENNINGHAUS. H., Deskriptive Statistik, Stuttgart 1974

BREDENKAMP, J., Experiment und Feldexperiment, in GRAUMANN, C.F. (Hrsg.), Handbuch der Psychologie Bd. 7, 1. Halbband, Sozialpsychologie, Göttingen 1969

CAMPBELL, D.T., u. STANLEY, J.C., Experimental and Quasi-Experimental Designs for Research, Chicago 1963

CANNEL, Ch. und KAHN, R. The Collection of Data by Interviewing, in: Research Methodes in the Behavioral Sciences (hg. von L. FESTINGER und D. KATZ), New York 1953

CLAUSS, G. u. EBNER, H., Grundlagen der Statistik, Thun und Frankfurt 1977

DOLLASE, R., Soziometrische Techniken, Weinheim und Basel 1973

EDWARDS, A.L., Techniques of Attitude Scale Construction, N.Y. 1957

ERBSLÖH, E., Techniken der Datensammlung 1, Das Interview, Stuttgart 1972

FISCHER, G.H., Einführung in die Theorie psychologischen Testens, Bern 1974

FRANK, L.K., Projective Methods, Springfield 1948

FRIEDRICHS, J. u. LÜDTKE, H., Teilnehmende Beobachtung, Weinheim-Berlin-Basel 1971

FRIEDRICHS, J., Methoden der empirischen Sozialforschung, Reinbek 1973

GRAUMANN, C.F., (Hrsg.), Handbuch der Psychologie Bd. 7, 1. Halbband, Sozialpsychologie, Göttingen 1969

GRÜMER, K.-W. Techniken der Datensammlung 2: Beobachtung, Stuttgart 1973

GUILFORD, J.P., Psychometric Methods, New York, Toronto, London 1954 (2)

HARTMANN, H., Empirische Sozialforschung, München 1970

HEISS, R., (Hrsg.), Handbuch der Psychologie Bd. 6, Psychologische Diagnostik, Göttingen 1964

HEMPEL, C.G. u. OPPENHEIM, P., Logic of Explanation, in: FEIGL, H. u. BRODBECK, M. (Hrsg.), Redings in the Philosophy of Science, N.Y. 1953

HILTMANN, H., Kompendium der psychodiagnostischen Tests, Bern, Stuttgart, Wien 1977

HOLM, K., Theorie der Frage bzw. Theorie der Fragebatterie, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Heft 1 und 2, 1974

HOLM, K., Die Befragung, 6 Bände, München 1975 – 77

KAHN, R.L. und CANNEL, C., The Dynamics of Interviewing, N.Y. 1957

KELLERER, J., Theorie und Technik des Stichprobenverfahrens, München 1963 (3)

KISH, A., Some Statistical Problems in Research Design, in: American Sociological Review Nr. 24, 1959

KÖNIG, R., (Hrsg.) Praktische Sozialforschung Bd. I, Das Interview, Formen, Technik, Auswertung, Köln 1972

KÖNIG, R., (Hrsg.) Praktische Sozialforschung Bd. II, Beobachtung und Experiment, Köln 1966 (3)

KÖNIG, R., (Hrsg.) Handbuch der empirischen Sozialforschung Bd. 1, Stuttgart 1967 (2), in Taschenbuchausgabe 1973

KOOLWIJK van, J., Die Befragungsmethode, in: KOOLWIJK van, J. u. WIEKEN-MAYSER, M., (Hrsg.), Techniken der empirischen Sozialforschung Bd. 4, Erhebungsmethoden: Die Befragung, München 1974

KOOLWIJK van, J., u. WIECKEN-MAYSER, M., (Hrsg.) Techniken der empirischen Sozialforschung Bd. 1 – 6, München (unterschiedliche Jahrgänge)

KRAPP, A. u. PRELL, S., Studienhefte zur Erziehungswissenschaft, Heft 5, Empirische Forschungsmethoden, München 1975

KRIZ, J. Grundlagen der Statistik, Reinbek 1973

LAKATOS, J. u. MUSGRAVE, H., Criticism and the Growth of Knowledge, Cambridge 1970

LAMNEK, S. u. KRUTWA, A., Grundriß der elektronischen Datenverarbeitung, München 1975

LIENERT, G.A., Testaufbau und Testanalyse, Weinheim 1969 (3)

MAYNTZ, R. u.a., Einführung in die Methoden der empirischen Soziologie, Köln und Opladen 1972

MICHEL, L., Allgemeine Grundlagen psychometrischer Tests, in: HEISS, R., (Hrsg.), Handbuch der Psychologie, Bd. 6, Psychologische Diagnostik, Göttingen 1964

MORENO, J.L., Die Grundlagen der Soziometrie, Köln und Opladen 1954

MYRDAL, G., Objektivität in der Sozialforschung, Frankfurt 1971

NOELLE, E., Umfragen in der Massengesellschaft, Reinbek 1976

OPP, K.D., Methodologie der Sozialwissenschaften, Reinbek 1971

OPPEL, U.G., u. RÜGER, B., Biomathematik, medizinische Statistik und Dokumentation, München 1975

POPPER, K.R., Logik der Forschung, Tübingen 1966 (2)

POPPER, K.R., Objektive Erkenntnis, Hamburg 1973

PRIM, R. u. TILMANN, H., Grundlagen einer kritisch-rationalen Sozialwissenschaft, Heidelberg 1973

RATHGEBER, W., (Hrsg.), Examensfragen zur medizinischen Psychologie, München 1975

RATHGEBER, W., (Hrsg.), Examensfragen zur medizinischen Soziologie, München 1975

RATHGEBER, W., (Hrsg.), Medizinische Psychologie, München 1977 (2)

RICHTER, H.J., Die Strategie schriftlicher Massenbefragungen, Harzburg 1970

ROEDE, H., Befrager und Befragte, Probleme der Durchführung des soziologischen Interviews, Berlin 1965

SAHNER, H., Schließende Statistik, Stuttgart 1972

SCHEUCH, E.K., Das Interview in der Sozialforschung, in: KÖNIG, R., (Hrsg.), Handbuch der empirischen Sozialforschung Bd. 1, Stuttgart 1967 (2) als Taschenbuchausgabe 1973

SCHEUCH, E.K., Skalierungsverfahren in der Sozialforschung, in: R. KÖNIG (Hg.), Handbuch der empirischen Sozialforschung, 1. Bd., Stuttgart 1962

SCHEUCH, E.K., Auswahlverfahren in der Sozialforschung, in: R. KÖNIG (Hg.), Handbuch der empirischen Sozialforschung, Bd. I., Stuttgart 1967

SIEGEL, S., Nonparametric Statistics for the Behavioral Sciences, N.Y. 1956, deutsch: Nichtparametrische statistische Methoden, Frankfurt 1976

SIXTL, F., Meßmethoden der Psychologie, Weinheim 1967

STELZL, I., Experimentelle Versuchsanordnungen, in KOOLWIJK van, J. u. WIECKEN-MAYSER, M., (Hrsg.) Techniken der empirischen Sozialforschung Bd. 6, Statistische Forschungsstrategien, München 1874

STROSCHEIN, F.-R., Die Befragungstaktik in der Marktforschung, Wiesbaden 1965

ZEISEL, H., Die Sprache der Zahlen, Köln-Berlin 1970

ZETTERBERG, H., On Theory and Verification in Sociology, Totawa, N.J. 1968

SACHREGISTER