Published as:
Greve, W. & Wentura, D. (1997). *Wissenschaftliche Beobachtung. Eine Einführung*. 2. Aufl. Weinheim : Beltz , 1997. - 182 S. : Ill., graph. Darst. - ISBN 3-621-27360-3

Veröffentlichungsversion:
Aus rechtlichen Gründen sind Cover und Rückseite, Titelblatt und Verlagsprogramm entfernt.

Werner Greve/Dirk Wentura

Wissenschaftliche Beobachtung

Anschriften der Autoren:
Dr. Werner Greve
Kriminologisches Forschungsinstitut Niedersachsen
Lützerodestraße 9
30161 Hannover

Dr. Dirk Wentura
Universität Münster
Psychologisches Institut IV
Fliednerstraße 21
48149 Münster

Umschlaggestaltung: Dieter Vollendorf, München
Druck und Bindung: Druckhaus „Thomas Müntzer", Bad Langensalza
Printed in Germany
Gedruckt auf säurefreiem Papier

ISBN 3-621-27360-3

Vorwort zur ersten Auflage

Systematische Beobachtung ist in aller Regel die erste Datenerhebungsmethode, die Studierende der Psychologie kennenlernen. Nicht selten gehört das Thema zu den ersten Dingen, die man überhaupt im Psychologiestudium lernt (oft genug hört man danach freilich nie wieder etwas davon – bedauerlicherweise). Dieses Buch ist aus unserer Erfahrung mit diesen Lehrveranstaltungen entstanden. Studierende können hier nach unserem Eindruck kaum auf systematische und einführende Darstellungen zurückgreifen. Viele Aufsätze sind abschreckend speziell, mitunter auch schwierig, und behandeln Details, die jedenfalls in einer einführenden Veranstaltung, insbesondere wenn sie auch praktische Übungen enthalten soll, nur kurz benannt werden müssen. Andere legen eine systematische Ordnung zugrunde, die ihre sachlichen Gründe haben mag, aber dem Einsteiger den Überblick – und das heißt: die Einsicht und das Behalten – nicht unbedingt erleichtern. Wieder andere sind für den Anfänger unlesbar, sei es, weil sie zu technisch geschrieben sind, sei es weil sie zu differenziert für den Einstieg sind, kurzum: weil an „den Anfänger" als Leser nicht gedacht wurde. Damit ist die Absicht dieses Bändchens deutlich: Wir wollen kein umfassendes Lehrbuch über Beobachtungsmethoden (in) der Psychologie vorlegen, sondern wir möchten Einsteiger an das Thema heranführen, ihr Interesse dafür wecken, mit wichtigen Problemen vertraut machen und Lösungsansätze in Umrissen vermitteln. Wir wollten kein hochaktuelles Nachschlagewerk für den Forscher, sondern eine semesterbegleitende Lektüre für Studierende schreiben. Da es uns dabei auch um eine Heranführung an die Psychologie allgemein geht, haben wir gerne auch ältere, „klassische" Arbeiten aufgegriffen, wenn sie das, was wir zeigen wollen, gut illustrieren.

Wir haben dabei die Wege des Üblichen nicht nur nicht verlassen, sondern uns möglichst mitten auf ihnen bewegt. Der Leser wird hier wenig finden, was nicht schon irgendwo von irgendwem bemerkt und diskutiert worden ist (manches davon sicher häufiger). Wir hoffen jedoch, daß der Einsteiger es hier müheloser und verständlicher findet – und überdies nicht an verschiedenen Orten suchen muß. Allerdings haben wir dabei nicht alle Verweise auf historische Vollständigkeit abgeklopft; mancher wird den einen oder anderen Klassiker vermissen. Dieses Einführungsbuch kann und soll das detaillierte Literaturstudium nicht ersetzen, vielleicht aber anregen und ein wenig erleichtern.

Wir schulden mehreren Personen Dank für Ihre Unterstützung bei diesem Projekt. Insbesondere der Beitrag von Dr. Horst Gräser und Dipl.-Psych. Ulrich Schmitz zu unterschiedlichen Phasen und Abschnitten der Arbeit war schließlich so bedeutend, daß eine besondere Nennung angemessener schien als eine Danksagung im Vorwort. Herr Prof. Dr. Jochen Brandtstädter hat das Manuskript in einer früheren Fassung aufmerksam gelesen; ihm sowie Herrn Prof. Dr.

Günter Krampen und Herrn Dr. Friedrich E. Heil verdanken wir außerdem verschiedene Hinweise und Ratschläge, nicht nur in Hinblick auf diesen speziellen Text. Frau Dipl.-Psych Ute Wahner hat bei der Literaturrecherche wichtige Hilfe geleistet und uns den Text von Binet (1897) zugänglich gemacht (d.h. übersetzt). Herr Dr. Jürgen Kagelmann vom Quintessenz-Verlag hat durch seine unbürokratische Zusammenarbeit das zügige Erscheinen des Buches ermöglicht. Auch ihm gilt unser Dank. Ein Wort in eigener Sache zum Schluß: Die Reihenfolge der Autoren ist nur durch das Alphabet diktiert. Wir haben den gesamten Text gemeinsam entworfen, diskutiert und überarbeitet.

Trier, im Mai 1991

Werner Greve *Dirk Wentura*

Vorwort zur zweiten Auflage

Daß eine zweite Auflage unseres Buches schon jetzt, nur fünf Jahre nach der ersten Veröffentlichung, möglich wurde, freut uns natürlich sehr. Ein äußerer Anlaß für sie ist auch der Wechsel des Quintessenz-Programmes zur Psychologie Verlags Union. Dies gibt uns aber auch Gelegenheit, ein paar kleinere Fehler in der ersten Auflage zu korrigieren.

Hannover und Münster, im September 1996

Werner Greve *Dirk Wentura*

Inhalt

Kapitel 1

„Seh' ich was, was Du nicht siehst?"
Bestimmung und Einführung in wissenschaftliche Beobachtung

Was ist eigentlich das Besondere an wissenschaftlicher Beobachtung? Ist Beobachtung nicht etwas, was jeder täglich macht, was jeder - mehr oder weniger - kann und können muß, um im Alltag zu bestehen? Was ist Beobachtung überhaupt mehr als das bloße Wahrnehmen, das Hinsehen oder Bemerken? Warum beobachten wir als Wissenschaftler? Was tun wir, wenn wir wissenschaftlich beobachten, anderes oder anders als beim alltäglichen Hinschauen?

Wissen wir nicht auf der anderen Seite gerade aus der alltäglichen Erfahrung mit uns selbst und anderen, wie unzuverlässig menschliche Beobachtung ist? Sind nicht beispielsweise die Gerichte ständig damit beschäftigt, das Durcheinander und die Widersprüchlichkeiten der verschiedenen Zeugenaussagen wenigstens annähernd zu klären? Bemerken wir nicht selbst immer wieder, daß wir uns auf unser Wahrnehmungs- und Urteilsvermögen auch bei scheinbar ganz einfachen Beobachtungsaufgaben nur unzureichend verlassen können? Laufen wir nicht oft blind durch die Welt, ohne viel von dem zu bemerken, was um uns herum vorgeht? Wie könnte dieses offenbar mangelhafte und nicht immer zuverlässige Instrument „Beobachtung" für Wissenschaft nutzbar gemacht werden? Welche konkreten Probleme tauchen bei Beobachtung in wissenschaftlicher Absicht auf, und vor allem: Was kann man tun, um sie zu vermeiden, zu beheben oder zu verringern?

Wir wollen in den folgenden Kapiteln die Vorteile und Stärken menschlicher Beobachtung nachzeichnen, ihre wichtigsten Schwächen diskutieren und Lösungsmöglichkeiten für sie skizzieren. Wir werden uns dabei auf die psychologische Beobachtung beziehen. Hinter dieser Bezeichnung verbirgt sich jedoch nichts Geheimnisvolles. Es geht um systematische Beobachtung mit verschiedenen Methoden in wissenschaftlicher - und das heißt hier: psychologischer - Absicht, die bestimmten Kriterien genügen muß. Die Grundlage dieser wissenschaftlichen Beobachtung und die Hauptursache der meisten Schwierigkeiten ist aber natürlich die allgemeine menschliche Fähigkeit zu beobachten. Insofern wird das, was wir in diesem Buch diskutieren, nicht nur für wissenschaftliche Beobachtung, sondern allgemein für menschliche Beobachtung in den verschiedensten Zusammenhängen gelten – mit der einen oder anderen Einschränkung vielleicht.

Es ist dabei vermutlich angebracht, zwei Einschränkungen vorab ausdrücklich hervorzuheben. Zum einen ist das Konzept der wissenschaftlich-psychologischen Beobachtung, um das es in diesem Buch gehen wird, insofern eine Beschränkung, als vieles, was man unter „psy-

chologischer" Beobachtung verstehen könnte, in diesem Rahmen nicht zur Sprache kommen wird. So wird etwa die Selbstbeobachtung, die in vielen praktischen Kontexten eine Rolle spielt, nur in der historischen Einleitung kurz berührt, oder eine therapeutische Beobachtungsfähigkeit (im Sinne einer besonderen *Sensibilität*), überhaupt die praktische Relevanz von Beobachtung in vielen Anwendungskontexten kaum angesprochen. Damit soll deren Wichtigkeit nicht bestritten werden; in akademischen Kontexten werden diese Aspekte von Beobachtung freilich seltener diskutiert.

Zum anderen ist es – wir haben das im Vorwort bereits betont – nicht unsere Absicht, gängige Verfahren repräsentativ oder umfassend aufzulisten, vorzustellen oder ausführlich zu diskutieren. Derartige Sammlungen finden sich für einzelne Forschungsbereiche etwa bei Manns, Schultze, Herrmann und Westmeyer (1987; diese Sammlung enthält 26 nach einheitlichen Kriterien ausgewählte Verfahren zu sechs verschiedenen Bereichen), Rosenshine und Furst (1973; vor allem Studien und Instrumente zu Untersuchungen im Unterricht) oder in der umfassenden Sammlung von Simon und Boyer (1974; insgesamt 99 Verfahren). Alle Beispiele, die im weiteren genannt oder beschrieben werden, dienen vielmehr zur Illustration.

1.1 Vorab ein Beispiel

Bevor wir uns den im vorherigen Abschnitt angesprochenen Fragen zuwenden, wollen wir zunächst mit einem einfachen Beispiel aus der Forschungspraxis illustrieren, wie Beobachtung eingesetzt wird. Das Beispiel ist - wie in den weiteren Kapiteln deutlich werden wird - sicherlich nicht in jeder Hinsicht repräsentativ für den Einsatz der Methode; es werden jedoch einige wichtige Punkte deutlich.

Lernen Kinder neue Verhaltensweisen schon dadurch, daß sie dieses Verhalten bei einer anderen Person „abschauen"? Oder werden solche Nachahmungen nur dann vom Kind gelernt, wenn die Modellperson für ihr Verhalten belohnt, nicht aber, wenn sie bestraft wird? Albert Bandura widmete dieser Frage ein interessantes Experiment (1966). Er zeigte 66 Kindern im Alter von 3 bis knapp 6 Jahren eine Filmszene, in der eine Person („Rocky") eine Puppe in Erwachsenengröße recht aggressiv traktierte und beschimpfte. Insbesondere zeigte Rocky vier für die Kinder neue Verhaltensweisen; sie sind in Tabelle 1 wiedergegeben.

Bei einem Drittel der Kinder endete der Film damit, daß Rocky für sein aggressives Verhalten belohnt wurde; ein weiteres Drittel sah eine Bestrafung des Modells, und in dem Film, den die restlichen Kinder sahen, blieb das Modellverhalten ohne Konsequenzen. Nach der Filmszene wurden die Kinder in einen „Spielraum" geführt, in dem neben vielen anderen Spielsachen auch die Requisiten des Films (Puppe, Bälle, Hammer) bereit lagen. Jedes Kind durfte dort 10 Minuten für sich allein spielen.

Tabelle 1: Das Modellverhalten in Banduras Experiment (Bandura, 1966)

	Rocky ...
1	... legt Puppe auf Boden, ... setzt sich auf Puppe, ... haut ihr auf die Nase, ... sagt „Pow, right in the nose, boom, boom".
2	... richtet Puppe auf, ... schlägt ihr mit einem Holzhammer auf den Kopf, ... sagt „Sockero ... stay down".
3	tritt die Puppe durch den Raum, ... sagt „Fly away".
4	wirft Gummibälle gegen die Puppe, ... sagt „Bang".

Der Spielraum war über eine sogenannte „Einwegscheibe" mit einem Beobachtungsraum verbunden: Die Kinder konnten so unbemerkt während ihres Spiels beobachtet werden. Die Beobachter sollten von dort aus jede Nachahmung des Modells registrieren. Ganz konkret hatten sie die Aufgabe, für jedes 5-Sekunden-Intervall zu notieren, ob eine (und wenn ja, welche) der vier Verhaltensweisen vom Kind gezeigt wurden. In Abbildung 1 ist ein (fiktiver) Protokollbogen wiedergegeben, wie er so oder ähnlich von Banduras Beobachtern benutzt wurde.

Versuchsperson: Peter

Beobachter: Klaus

Verhalten	Zeit												
	5	10	15	20	25	30	35	40	45	50	55	60	65
1	I	I	I										I
2					I	I							
3													
4										I			

Abbildung 1: Ein fiktiver Protokollbogen der Untersuchung von Bandura

Nach dieser Phase wurden die Kinder an den Film erinnert und aufgefordert, wenn möglich die Modellverhaltensweisen nachzuahmen („Zeige mir, was Rocky getan hat!"). Dafür wurden ihnen kleine Belohnungen versprochen. Auch in dieser zweiten Phase wurde wieder beobachtet, welches Modellverhalten die Kinder zeigten. Das Ergebnis dieses Experiments war, daß das spontane Imitieren des Modells in der ersten Phase deutlich davon abhing, ob das Modell im Film bestraft worden war oder nicht. Dieser Unterschied verschwand jedoch in der

zweiten Phase: *Gelernt* hatten die Kinder die Verhaltensweisen also unabhängig davon, welche Konsequenzen das Modell zu ertragen hatte. Ob das Gelernte *spontan in Verhalten umgesetzt* wurde, hing jedoch deutlich von diesen Konsequenzen ab.

> Vielleicht kann schon hier ein Hinweis auf die Probleme der *Interpretation* derartiger Befunde nicht schaden. Eine wichtige Frage, auf die Brandtstädter (1981, S. 96f.) hingewiesen hat, besteht nämlich darin, ob die Kinder von dem „Modell" Rocky tatsächlich *aggressives* Verhalten gelernt haben. Wir werden im zweiten Kapitel noch genauer sehen, daß menschliche Handlungen (z.B. Aggressionshandlungen) nur unter Bezug auf subjektive Überzeugungen und Intentionen identifiziert werden können. Anders gesagt: Eine Handlung ist nur dann aggressiv, wenn sie aggressiv *gemeint* ist; und genau das kann man bei den Kindern, die Rockys Verhalten äußerlich zeigen, bezweifeln. Vielleicht *spielen* sie nur.

1.2 Kennzeichnung: Was ist wissenschaftliche Beobachtung?

Wir wollen nun etwas genauer klären, was wissenschaftliche Beobachtung eigentlich ist. Was kennzeichnet sie, wie läßt sie sich definieren, was ist das Besondere an ihr? Die endgültige Klärung wird dabei natürlich nicht mit ein paar Sätzen zu leisten sein. Im Grunde ist es das Ziel des gesamten Buches, ein möglichst zutreffendes und hinreichend differenziertes Verständnis dessen zu vermitteln, was psychologische Beobachtung bedeutet. So gesehen ist die Begriffsklärung am Anfang eher ein Einstieg ins Thema als die endgültige Bedeutungsfestschreibung.

Zunächst: Was meinen wir mit unserem alltäglichen Begriff der Beobachtung? Mit Graumann (1966) kann man ihn so umschreiben:

> „Die absichtliche, aufmerksam-selektive Art des Wahrnehmens, die ganz bestimmte Aspekte auf Kosten der Bestimmtheit von anderen beachtet, nennen wir Beobachtung. Gegenüber dem üblichen Wahrnehmen ist das beobachtende Verhalten planvoller, selektiver, von einer Suchhaltung bestimmt und von vorneherein auf die Möglichkeit der Auswertung des Beobachteten im Sinne der übergreifenden Absicht gerichtet. Im alltäglichen Verhalten gehen Wahrnehmen und Beobachten oft unmerklich ineinander über." (S. 86)

Festzuhalten sind also vor allem folgende Punkte:

- *Absicht*: Wir sprechen von Beobachtung im Sinne eines absichtlichen, geplanten Unternehmens. Beobachtung setzt einen Zweck, ein Ziel voraus. Im Experiment, das in Abschnitt 1.1 dargestellt wurde, ist das die Frage: Imitieren Banduras Kinder das Modellverhalten in Abhängigkeit davon, ob das Modell bestraft oder belohnt wurde?
- *Selektion*: Das bedeutet, daß wir bestimmte Aspekte unseres Wahrnehmungsfeldes genauer betrachten, andere aber vernachlässigen. Banduras Beobachter beschrieben nicht die Kleidung oder das Aussehen der Kinder, sondern konzentrierten sich auf das Imitationsverhalten.
- *Auswertung*: Beobachtung ist immer auf eine Auswertbarkeit der Ergebnisse ausgerichtet. Wahrgenommenes muß dementsprechend auf ein System von Zeichen, die vereinbarte Be-

deutung tragen, „abgebildet" werden. Die Beobachter in unserem Beispiel kodierten auf ihrem Protokollbogen durch einfaches Anstreichen „Versuchsperson Peter zeigte im zehnten 5-Sekunden-Intervall das Modellverhalten Nr. 4". Aber auch das einfache Beschreiben mit der Alltagssprache ist in diesem Sinne ein „Abbilden auf ein System von Zeichen".

Was macht nun eine Beobachtung zu einer wissenschaftlichen Beobachtung? Feger (1983) ergänzt die Formulierung von Graumann:

> „Wenn die übergreifende Absicht ist, eine wissenschaftliche Annahme zu prüfen, und wenn sie in Planung und Bewertung bestimmten Kriterien genügt, geht die vorwissenschaftliche in die wissenschaftliche Beobachtung über." (S.3)

Zwei Gegebenheiten müssen also nach Feger erfüllt sein, damit wir von wissenschaftlicher Beobachtung sprechen können:

- Ziel der Beobachtung ist es, eine Theorie bzw. Hypothese zu prüfen (dazu gleich mehr; s. Abschnitt 1.4.2, Exkurs). Diese Hypothesen können zunächst noch vage sein; in diesen Fällen würden wir eher von Vermutungen bzw. einer „explorativen" Untersuchung reden. Bandura hatte dagegen präzise Hypothesen über die Auswirkung von Belohnung und Bestrafung auf das Lernen und das spontane Imitationsverhalten der Kinder.
- „Bestimmte Kriterien" müssen erfüllt sein. Um welche Kriterien es sich dabei handelt, wird Thema späterer Abschnitte sein (vgl. z.B. die Abschnitte 3.2.1, 4.4); hier nur soviel: Die Ergebnisse einer Beobachtungsstudie sollten natürlich wiederholbar sein (Stichwort: *Replizierbarkeit*), und verschiedene Beobachter sollten bei Beobachtung desselben Sachverhaltes zu demselben Ergebnis kommen (Stichwort: *Objektivität*). In der Veröffentlichung von 1966 gibt Bandura alle wesentlichen Informationen, so daß andere Forscher seine Studie wiederholen können. Außerdem hat er zumindest bei einem Teil der Kinder zwei Beobachter unabhängig voneinander beobachten lassen. So konnte er aufzeigen, daß die Verhaltensweisen so eindeutig definiert waren, daß verschiedene Beobachter praktisch immer zum selben Zeitpunkt die gleiche Kodierung vorgenommen hatten.

Die vier wesentlichen Punkte in der Unterscheidung von wissenschaftlicher Beobachtung und einfacher Wahrnehmung sind – stichwortartig nochmals zusammengefaßt – also:

1. die Absicht, Annahmen zu überprüfen;
2. die systematische Selektion bestimmter Aspekte;

(Diese beiden Punkte unterscheiden *alltägliche* Beobachtung von einfacher Wahrnehmung.)

3. die beabsichtigte Auswertung der erhobenen Daten und
4. die Kriterien der Replizierbarkeit und Objektivität.

(Diese beiden Punkte kennzeichnen *wissenschaftliche* Beobachtung.)

Wir wollen in den folgenden Abschnitten dieses Kapitels diese wissenschaftliche Beobachtung in Stellenwert und Bedeutung auf dreierlei Weise weiter einordnen, eingrenzen und bestimmen.

Zunächst werden wir sie grob in die historische Entwicklung der akademischen Psychologie einordnen (Abschnitt 1.3), dann gegen die Begriffe der Messung und des Experimentes abgrenzen (Abschnitt 1.4) und schließlich verschiedene Klassifikationen und Vorgehensweisen psychologischer Beobachtung vorstellen (Abschnitt 1.5).

1.3 Historische Skizze

Hätte man etwa um die Jahrhundertwende Psychologie studiert, ständen wohl nicht „Beobachtungsmethoden" im Ausbildungsplan, sondern ein Fach wie „Selbstbeobachtung" bzw. „Introspektion". Damit war das gezielte „Beobachten" der eigenen Bewußtseinsvorgänge gemeint, eine Art „Vivisektion" der subjektiven Empfindungen, Gefühle und Denkprozesse.

Diese Methodik ergab sich direkt aus einem Verständnis von Psychologie als der Wissenschaft, die die bewußte Erfahrung untersucht. Einer solchermaßen verstandenen Wissenschaft von den Bewußtseinsprozessen muß natürlich eine Beobachtung *objektiver* Gegebenheiten eher fremd sein. Auf den ersten Blick liegt es für Psychologen ja auch nahe, Menschen sich selbst aus *subjektiver* Perspektive „beobachten" zu lassen. Betrachtet man sich dieses Verfahren der Introspektion jedoch genauer, wird es schnell etwas suspekt: Was soll es bedeuten, möglichst genau in sich hineinzuschauen, um die Abfolge von Empfindungen zu beschreiben? Die „Introspektionisten" hofften aber, Versuchspersonen[1] soweit zu schulen, daß sie auf dem Weg einer „Innenschau" etwas über psychische Prozesse und Vorgänge erfahren könnten (ein geschichtlicher Rückblick hierzu findet sich etwa bei Danziger, 1980). Festzuhalten sind vor allem zwei wesentliche Kritikpunkte am Introspektionismus:

1. Der „Introspektionismus" verlangte zuviel von seinen Versuchspersonen, insbesondere Auskunft über Prozesse, zu denen diese keinen Zugang hatten.
2. Die Ergebnisse müssen deswegen beliebig sein. Ein Wissenschaftsideal ist dabei gravierend verletzt: Man möchte immer wissen, unter welchen objektiven Bedingungen sich ein Phänomen einstellt, um es jederzeit wiederholen zu können.

   Bei introspektionistischen Untersuchungen ist genau dies jedoch häufig gescheitert. So beschreibt Young (1927) die Wiederholung eines Experimentes zur Erfassung von Gefühlsqualitäten. Nafe (1924) hatte aus Introspektionsberichten geschlossen, daß sich „angenehm" und „unangenehm" als Erscheinungsweisen von Druck erwiesen: das Angenehme sei ein heller Druck (bright pressure), das Unangenehme ein dumpfer Druck (dull pressure) (Nafe, 1924, S. 508). Bei Youngs Replikationsversuch fand sich jedoch, daß seine Versuchspersonen (zwei Wissenschaftler und eine Studentin) so über diese Empfindungen berichteten, wie ihre „Schule" bzw. ihr Kenntnisstand es erwarten ließ.

Die „Wachablösung" kam dann auch bald: Bereits 1913 veröffentlichte J.B. Watson seinen berühmten Aufsatz „Psychology as a behaviorist views it" und leitete damit für die Psychologie

[1] „Versuchspersonen" waren dabei nicht etwa psychologische Laien, sondern die Forscher und ihre Schüler!

einen neuen Abschnitt ein. Karl Bühler beschreibt in seinem 1927 erschienenen Buch „Die Krise der Psychologie" diese „behavioristische Wende" wie folgt:

> „Während unsere [„introspektionistische"; G/W] Weisheit in dem Satz gipfelte: Erkenne Dich selbst, kamen die Behavioristen und lehrten: Betrachte das Benehmen (behavior) der Menschen und Tiere von außen. Du wirst, wenn Du es nur systematisch genug vollbringst, auch damit zu wichtigen psychologischen Erkenntnissen gelangen. Damit haben sie die Tierpsychologie reformiert und in der Tat einen neuen, unentbehrlichen Grundaspekt vom Gegenstand der Seelenlehre gewonnen." (S. 19)

Bis in die späten fünfziger Jahre dominierte dann das Forschungsprogramm des „Behaviorismus", jedenfalls die amerikanische und - nach dem Zweiten Weltkrieg - auch die deutsche (akademische) Psychologie. Daneben hielten sich, gewissermaßen am Rande des Fahrwassers, natürlich auch andere Strömungen nicht nur am Leben, sondern gewannen in anderen, z.B. therapeutischen Kontexten nicht unbeträchtlichen Einfluß. Prominentestes Beispiel hierfür ist vermutlich die Psychoanalyse, ein anderes die Gestaltpsychologie. In der akademischen Psychologie freilich setzten sich zwei Hauptforderungen, mit denen sich der Behaviorismus vor allem von den „Introspektionisten" (zugleich aber auch wenigstens teilweise von der Psychoanalyse) abgegrenzt hatte, deutlich durch, eine methodische und eine thematische:

1. Die Psychologie soll nur objektive Methoden anwenden; Ergebnisse sollten wiederholbar sein.
2. Ausschließlich beobachtbares Verhalten soll zum Gegenstand der Psychologie erhoben werden. Die Rede in psychischen Kategorien, sogenannten „mentalen" Begriffen, wie Überzeugung, Absicht, Gefühl oder gar Bewußtsein, muß für die wissenschaftliche Psychologie verworfen werden.

Dementsprechend war objektive Beobachtung im Grunde die Methode der Wahl des Behaviorismus, allerdings zunächst im Sinne einer einfachen Registrierung von Reiz-Reaktions-Abfolgen: Drückt die Ratte auf den Knopf, woraufhin sie eine Futterpille bekommt, oder nicht?

Mittlerweile ist freilich klargeworden, daß ein Forschungsprogramm, das *nur* die Untersuchung von Reiz-Reaktions-Sequenzen ohne Rückgriff auf innere Zustände zum Programm erhebt, scheitern muß. In den späten fünfziger Jahren mehrte sich dann auch die Kritik am Behaviorismus. Berühmt geworden ist z.B. die Kritik von Noam Chomsky (1959) an dem Versuch von B.F. Skinner (1957), der beherrschenden Figur des Behaviorismus, auch Sprechen als reines (Sprech-)*Verhalten* behavioristisch zu rekonstruieren. Zum einen ist Sprechen mehr als nur die Produktion von Schallereignissen oder eine Folge von Muskelbewegungen, es hat *Bedeutung*. Das Lernen von Sprache ist zum anderen nicht nur als reaktiver Vorgang denkbar. Der Lernende lernt aktiv, selektiv, unterschiedlich leicht und gut, wendet Regeln aktiv an etc. Ohne die Annahme von Voraussetzungen auf seiten der Person ist die Vielfalt der Phänomene schlechterdings unerklärlich. Die Forschung zu komplexen Vorgängen wie eben menschlichem Sprechen oder dem Spielen eines Musikinstrumentes etc. konnte mit einem theoretischen Modell, das jedes Verhalten als Reaktion auf äußere Reize verstand, nicht weiterkommen: Man

kann auf die theoretische Annahme von internen Zuständen (z.B. Kognitionen) nicht verzichten (eine Geschichte der „Kognitionsforschung" findet sich bei Gardner, 1985/1989).

Die behavioristische Psychologie machte Forschungsprogrammen Platz, die es wieder für angemessen hielten, von inneren Zuständen und dementsprechend in „mentalen Begriffen" (Absicht, Überzeugung, Gefühl) zu reden. Man arbeitet seither an und mit Theorien, die diese Begriffe zueinander in Beziehung setzen. Auf der *thematischen* Ebene ist man somit *auch* wieder zu den Inhalten der alten Bewußtseinspsychologie zurückgekehrt (vgl. Brandtstädter, 1991). Wie versucht man jedoch heute, diese Theorien zu testen? In *methodischer* Hinsicht wurde die erste Forderung der Behavioristen nie aufgegeben: Nach wie vor gilt das Postulat, wiederholbare Ergebnisse zu erbringen. Nur: Die Theorien legen nahe, welche Methode angemessen ist. Auf der Methodenebene entspricht daher der Theorienvielfalt eine Methodenvielfalt. So werden auch heute wieder Selbstauskünfte der Versuchspersonen benutzt.

> Allerdings ist das insgesamt keine einfache Sache. Menschen sind sich in aller Regel durchaus nicht aller Aspekte ihrer psychischen Situation bewußt und können schon deswegen über vieles davon nicht sprachlich Auskunft geben. Die Psychoanalyse Sigmund Freuds (1938) hat diesen Punkt besonders betont und neben das Bewußtsein, das von Freud als „höchst flüchtiger Zustand" charakterisiert wurde, die psychischen Qualitäten des Vorbewußten und des Unbewußten gestellt, die nur unter bestimmten Umständen bewußt werden können oder sogar nur aus indirekten Zeichen (im Verhalten oder Erleben) erschlossen werden können. Mit Selbstauskünften sind hier insofern nicht nur sprachliche Mitteilungen gemeint, sondern alle Informationen, mit denen jemand Auskunft über seine psychische Situation gibt. Die Interpretationsleistungen, durch die etwa ein Therapeut von derartigen Anzeichen auf psychische Vorgänge schließt, sind allerdings mit Irrtumsrisiken behaftet. In bezug auf einige Wahrnehmungen („Dies sieht rot aus") oder Empfindungen ("Das tut mir weh") ist dieses Irrtumsrisiko für die eigene Person ausgeschlossen (vgl. auch Greve, 1996).

In einigen Fällen haben wir zu Selbstauskünften auch keine Alternative: Wie sollten wir z.B. sonst etwas über die Meinungen oder Empfindungen von Personen erfahren? Wir bitten freilich unsere Versuchspersonen in der Regel nicht mehr, Unmögliches zu tun.

Welche Rolle dabei die Beobachtung spielen kann, wollen wir an der Reaktion auf eine klassische Untersuchung zeigen: der Forschung zum sogenannten „Pygmalion-Effekt"[2]. Die zugrundeliegende Studie von Rosenthal und Jacobsen (1968/1971) basierte zwar nicht auf Beobachtung, wir müssen eine kurze Schilderung aber zum Verständnis des Folgenden voranschicken.

> Es ging in dieser Studie um die Rolle von Lehrererwartungen bei der geistigen Entwicklung von Schulkindern. Die Autoren testeten 255 Kinder der unteren Schulklassen mit einem Test zur Erfassung allgemeiner intellektueller Fähigkeiten. Den Lehrern war jedoch mitgeteilt worden, daß mit dem Test „intellektuelle Spätentwickler" erkannt werden können, die in der nächsten Zeit einen besonders starken Leistungszuwachs zeigen würden. Tatsächlich wurden den Lehrern nach dem Test einige Kinder (etwa 20%) als „Auf-

[2] Benannt nach einem griechischen Künstler, der eine Statue der Aphrodite schuf, die ihm so gut gelang, daß er sich unsterblich in sie verliebte. Die Göttin, durch seine Verzweiflung gerührt, erweckte dieses Produkt aus seinen Wünschen und Phantasien schließlich zum Leben.

> blüher" benannt, die per Zufall ausgewählt worden waren. Nach einem Jahr wurde erneut getestet: Es zeigte sich, daß sich in den unteren beiden Klassen die vermeintlichen „Aufblüher" tatsächlich besser als der Rest der Klasse entwickelt hatten.
>
> Das Experiment von Rosenthal und Jacobson wurde sehr heftig diskutiert, in sehr vielen Aspekten der Untersuchungsmethodik zu Recht scharf kritisiert (Elashoff & Snow, 1971/1972) und häufig wiederholt: Dabei erwies sich der „Pygmalioneffekt" als eher flüchtiges Phänomen, welches sich nicht einfach replizieren ließ (zum Überblick: Brophy & Good, 1974/1976, Kap. 3). Rheinberg und Minsel (1986) stellen fest: „Rückblickend erscheint der Pygmalion-Effekt wie ein trojanisches Pferd, mit dem (Lehrer-)Kognitionen in die vormals behaviorale Erforschung der Lehrer-Schüler-Interaktion eindrangen und dort bis heute bemerkenswert Raum gewannen." (S. 333)

Die nachfolgende Forschung zeigte, daß die Vermittlung von Lehrererwartungen auf Schülerleistungen offenbar ein komplexer Prozeß ist. In diesem Prozeß spielen sehr viele nicht direkt beobachtbare Größen eine Rolle. Die Lehrer*erwartung* über die *Fähigkeit* des Schülers hängt damit zusammen, was der Lehrer über eine schlechte Leistung des Schülers *denkt*. *Glaubt* der Lehrer z.B., der Schüler sei *begabt*, wird er den Fehler auf änderbare Faktoren *zurückführen* („Er hat heute einen schlechten Tag", „Er hat zu wenig gelernt" etc.) und dementsprechend reagieren. Das *Selbstbild* des Schülers bestimmt die *Motivation*, die er einer Aufgabe entgegenbringt: *Glaubt* er beispielsweise, bestimmte Fähigkeiten nicht zu besitzen, wird er vielleicht etwas schneller aufgeben.

In der Pädagogischen Psychologie versucht man all diese - hier nur angedeuteten - Zusammenhänge in Theorien auszudifferenzieren und testbar zu machen. Interviews, Fragebögen und Reaktionen auf raffiniert konstruierte Experimentalsituationen werden hierzu eingesetzt. Aber auch Beobachtungsmethoden erfüllen eine wichtige Funktion, denn letztlich vermittelt sich die Auswirkung der Erwartung auf die Leistung über das *beobachtbare Verhalten* von Lehrer und Schüler, über ihre Interaktion. Dabei wird dem Lehrer sicherlich nur zum geringsten Teil bewußt sein, welch unterschiedliches Verhalten er gegenüber Schülern zeigt; teilweise wird er sich vehement gegen die Vermutung wehren, die Kinder unterschiedlich zu behandeln. Also muß man beobachten, welche Verhaltensmuster sich zeigen, Verhaltensmuster, die man zu den theoretischen Begriffen in Beziehung setzen kann. Wie setzt der Lehrer Lob und Tadel ein? Wie sehr suchen einzelne Kinder die Bestätigung durch den Lehrer? Brophy und Good (1974/1976) fanden etwa heraus, daß Kinder, von denen der Lehrer relativ wenig erwartete, tatsächlich weit häufiger getadelt als gelobt wurden. Außerdem zeigte sich, „daß gegen Ende des Schuljahres die hohen Gruppen wesentlich aktiver waren, indem sie Reaktionsgelegenheiten suchten und zum Lehrer gingen, um ihre selbständig angefertigte Arbeit zu erörtern" (S. 139). Sicherlich sind solche Ergebnisse noch recht grobe Bausteine im Verständnis dieser Prozesse; gleichwohl erweist sich hier die Beobachtung als unverzichtbar im Kanon der psychologischen Forschungsmethoden - sie wird dabei nicht mit dem fragwürdigen Anspruch der Introspektion oder der Ausschließlichkeit des Behaviorismus eingesetzt, aber als dennoch wichtige Methode.

1.4 Abgrenzung: Beobachtung und Messung, Beobachtung und Experiment

Diese bislang noch recht grobe Kennzeichnung von „wissenschaftlicher Beobachtung" soll nun durch Abgrenzung („De-finition") gegen zwei andere Begriffe ergänzt werden: „Messung" und „Experiment". In beiden Fällen wird sich herausstellen, daß eine strikte Abgrenzung unangebracht ist. Beobachtung in dem Sinne, in dem wir den Begriff hier verwenden werden, *ist* eine Form der Messung und – potentieller – Teil eines Experimentes. Durch die beiden folgenden Abschnitte werden außerdem eine Reihe weiterer Konzepte eingeführt, die zum Verständnis des Folgenden nützlich oder sogar nötig sind. Wir werden etwas darüber lernen, was wissenschaftliches Vorgehen überhaupt ausmacht, und von daher dann auch genauer sehen, was wissenschaftliche Beobachtung auszeichnet.

1.4.1 Beobachtung und Messung

Beginnen wir mit der Frage, ob (und wenn ja: inwieweit) Beobachtung und Messung dasselbe, vergleichbar oder grundsätzlich verschieden sind. Wir wollen diesen Punkt hier zunächst nur kurz andeuten (im Kapitel 5 wird deutlicher werden, wo die Parallelen liegen und wo sie gegebenenfalls enden). So schreibt Beck (1987, S. 14, Fußnote 3) für das verallgemeinerungsfähige Beispiel der Unterrichtsforschung: „Das Ziel, durch Beobachtung zu theorierelevanter Beschreibung zu gelangen, ist das gleiche geblieben, und es wird ... im Prinzip noch immer auf die gleiche Weise zu erreichen versucht, nämlich durch die Einschaltung des Beobachters als 'Meßinstrument', als 'Gerät', das den input 'Unterrichtsereignisse' intern verarbeitet und umwandelt zum output 'Unterrichtsbeschreibung'."

Was verstehen wir sinnvollerweise unter „Messen"? Betrachten wir einen einfachen Fall: das Messen von Temperatur (Abb. 2).

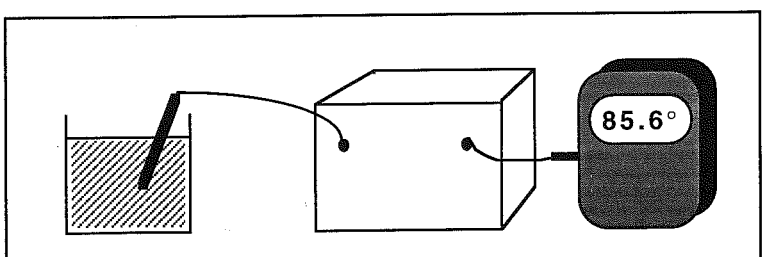


Abbildung 2: Temperaturmessung

Charakteristisch ist zunächst wiederum der Selektionsaspekt: Bestimmte Gegebenheiten des eingesetzten „Reizempfängers" - im Beispiel ein elektronisches Bauelement - verändern sich analog zur Temperatur des Wassers - und eben nicht analog zu seiner Härte, Konsistenz oder Farbe. Es werden dann den jeweiligen Temperaturen (in diesem Fall wiederum durch ein mechanisches oder elektronisches Bauteil) Zahlenwerte zugeordnet. Man sagt, die Eigenschaft „Temperatur" wird auf den Bereich der Zahlen *abgebildet*. Wie diese Abbildung vorzunehmen ist, wird durch die Konstruktion des Thermometers festgelegt. Anders formuliert: Durch die spezifische Konstruktion des Apparates werden Regeln der Zuordnung von Zahlen zur Eigenschaft „Temperatur" aufgestellt, bzw. bei seiner Konstruktion wird diesen Regeln gefolgt. Man kann an diesem Beispiel sehr schön sehen, daß diese Regeln explizit vereinbart wurden: Es ist bekanntlich eine Konvention (oder Normierung), daß im Falle der Celsius-Skala die Null dem Gefrierpunkt und 100 dem Siedepunkt des Wassers zugeordnet ist. Diese Abbildungsvorschriften für die Zuordnung von Symbolen (bei der Messung in der Regel Zahlen) zu empirischen Gegebenheiten macht einen Teil unseres Begriffs von Messung aus. Der andere wichtige Aspekt besteht darin, daß wir bei einer Messung immer ausdrücklich vereinbaren, welche mathematischen Operationen wir mit den so gewonnenen Zahlen anstellen dürfen, so daß deren Ergebnisse wieder als Eigenschaft des empirischen Bereichs interpretiert werden dürfen.

> Was heißt das etwas konkreter ausgedrückt? Angenommen, wir messen zu fünf verschiedenen Zeitpunkten die Temperatur des Wassers, so besagt die Vereinbarung für eine solchermaßen vorgenommene Temperaturmessung, daß ich einen Mittelwert errechnen und mit den Worten „das Wasser hatte eine mittlere Temperatur von x Grad Celsius" interpretieren kann. Demgegenüber stelle man sich die Zuordnung von Zahlen zu Farben des Wassers vor: Z.B. vergeben wir eine Eins, wenn das Wasser blau ist, eine Zwei, wenn es rot ist usw. Dürfen wir hier über verschiedene erhobene Werte einen Mittelwert bilden? Natürlich - aber wir dürfen in diesem Fall nicht den erhaltenen Wert in dem Sinne „das Wasser hatte eine mittlere Farbe von blau" interpretieren. In dem Fall dieser zweiten Anordnung müßte festgelegt sein, daß die Ergebnisse solcher Operationen nicht interpretiert werden dürfen.

Kurz gesagt: Wir wollen von Messung dann sprechen, wenn nicht nur die Abbildungsvorschrift gegeben ist, sondern auch etwas über die Interpretierbarkeit der Ergebnisse von Operationen auf den zugeordneten Symbolen gesagt wird.

Beobachtung ist in vielen Fällen auch Messung - aber nicht immer werden wir diesen strengen Begriff in Anspruch nehmen, da nicht immer Aussagen über den Bereich der Operationen gemacht werden. Das gilt insbesondere dann, wenn wir Wahrgenommenes in unserer Alltagssprache beschreiben und dabei von einer formal präzise beschreibbaren Meßvorschrift sehr weit entfernt sind (vgl. dazu das folgende Kapitel 2). Dagegen wird das erste Kriterium der Messung - das Angeben einer Abbildungsvorschrift - in jedem Fall wissenschaftlicher Beobachtung erfüllt. Sehen wir uns dazu Abbildung 3 an.

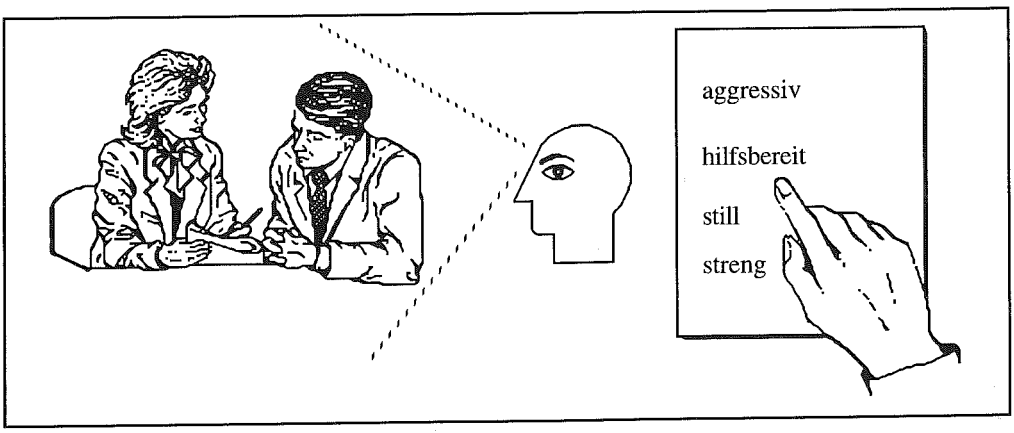


Abbildung 3: Beobachtung

Auch hier wird ausgewählt: Beobachtet wird nur das Verhalten der linken Person, nicht etwa das der rechten oder die Kleidung beider oder ein anderer Aspekt der Situation. Der Beobachter nimmt das Gesehene auf und ordnet entsprechend einer Abbildungsvorschrift ein Symbol einer Symbolmenge zu, in diesem Fall das Wort „hilfsbereit". Formal lassen sich also Temperaturmessung und Beobachtung bis hierher in gleicher Weise als Abbildungsvorgang verstehen. Auf die gleichwohl erheblichen Unterschiede werden wir zurückkommen (Kap. 5).

1.4.2 Beobachtung und Experiment: Gegensatz, Ergänzung, Kategorienverwechslung[3]?

Die gelegentlich heftige Diskussion, ob Beobachtung und Experiment in der Psychologie - aber auch in der Wissenschaft allgemein - ein Gegensatz seien, ob sich beides ergänze oder ob der Vergleich überhaupt auf einer Kategorienverwechslung beruhe, lebt von einem Mißverständnis. Der Begriff „Beobachtung" bezeichnet nämlich in diesem Zusammenhang mindestens zweierlei.

(1) Heuristische Beobachtung: Beobachtung als Haltung

Zum einen ist mit „Beobachtung" so etwas wie eine *Haltung* gegenüber dem Gegenstand der Forschung (im Falle der Psychologie also dem Menschen bzw. dem menschlichen Verhalten)

---

[3] Der Begriff der Kategorienverwechslung geht auf Gilbert Ryle zurück („Der Begriff des Geistes", 1949/1969). Was ist damit gemeint? Betrachten wir folgendes Beispiel: Ein fremder Besucher bittet mich, ihm die Universität Trier zu zeigen. Ich führe ihn also durch das Fachgebäude „Psychologie", zeige ihm kurz das Rechenzentrum, gehe mit ihm durch das Juristen- und Verwaltungsgebäude, esse mit ihm in der Mensa, schleiche durch den Lesesaal der Bibliothek und so weiter. Nach einem langen Nachmittag fragt er mich: „Das war ja alles ganz beeindruckend, aber ich hatte ja eigentlich gehofft, Sie würden mir die *Universität* zeigen. Was davon war denn nun die Universität?" Dieser Besucher begeht eine Kategorienverwechslung: „Universität" gehört nicht in dieselbe Kategorie wie „Fachgebäude Psychologie", denn letzteres kann man sehen, erstere ist ein abstrakter Begriff (der ggf. auch ein solches Gebäude umfassen kann).

gemeint. Diese Haltung läßt sich vielleicht folgendermaßen skizzieren: Der Forscher beteiligt sich nicht aktiv, er mischt sich nicht ins Geschehen ein, er kontrolliert und manipuliert nicht, er „läßt sich auf seinen Gegenstand ein", ohne bereits konkrete explizite Vermutungen oder sogar Theorien zu haben, d.h. ohne von vorneherein auf bestimmte Aspekte besonders oder ausschließlich zu achten und dabei andere zu ignorieren oder zu vernachlässigen. Er betrachtet seinen Gegenstand vielmehr so unvoreingenommen und so vollständig wie irgend möglich. Wir werden diese Bedeutung des Begriffs im folgenden als *heuristische* Beobachtung bezeichnen. Der heuristische Beobachter hofft, daß sich das, was er da betrachtet, zu einer Struktur, zu einer Systematik ordnet, daß er das Muster, das sich hinter dem individuellen Durcheinander verbirgt, nach und nach erkennt, daß er interessante Zusammenhänge *entdeckt*. Darin liegt die besondere Stärke der heuristischen Beobachtung: Mit einer sozusagen „gleichschwebenden" Aufmerksamkeit nach und nach die Kennerschaft erwerben, aus der die guten Ideen erwachsen, weil einem Dinge auffallen, die durch ihre Ungewöhnlichkeit aus den unmerklich ausgebildeten Strukturen herausragen oder so prägnante Muster bilden, daß sie zu einer expliziten Struktur werden. Diese Beobachtung ist nicht (explizit) theoriegeleitet, kann aber insofern systematisch sein, als sie um Erfassung und *Ordnung* bestimmter (weiter) Ausschnitte des Geschehens systematisch bemüht ist. Unter der Überschrift „Zurück zu den Sachen" bringt Norbert Bischof (1989) im Kontext eines Plädoyers für eine „biologische Wende" der Psychologie diese Auffassung sehr schön auf den Punkt:

> „Biologie ist vor allem eine *Haltung*. Der Kern dieser Haltung ist umschreibbar durch die vielleicht etwas anspruchsvolle Formel vom Sich-Einlassen aufs Leben. Den guten Biologen kennzeichnet die Besessenheit, mit der er einen Wurm oder eine Alge liebevoll bis ins kleinste Detail nachzeichnet, die Intensität, mit der eine Forscherin wie Jane Goodall Jahrzehnte im Urwald verbringt, um jeden Tag erneut mit disziplinierter Einfühlungskraft den Ereignisstrom in ihrer Schimpansenkolonie zu registrieren.
>
> Konrad Lorenz ... hat es als 'Modetorheit' der Psychologie gegeißelt, zu glauben, man könne auf Beschreibung verzichten, bevor man Theorien baut. 'Beschreibung' ist dabei nicht das, als was sie gern denunziert wird: Kuhäugige Anfertigung von Strichlisten über belanglose Kleindetails. Schon eher hat sie etwas mit der ebenfalls vielgeschmähten anekdotischen Betrachtung zu tun, sofern diese einer Gestaltwahrnehmung entspringt, die dem Leben kongenial ist und daher fähig, seine Prägnanzschwerpunkte aufzunehmen und zu Kristallisationskernen der Theorienbildung werden zu lassen.
>
> Was eigentlich gefordert wird, ist die Überwindung der Berührungsscheu vor dem Gegenstand, ... die Bereitschaft, Kennerschaft zu erwerben, bevor man mit Wissenschaft beginnt." (S. 203).

Es hat in der Diskussion um psychologische Beobachtung immer wieder Stimmen gegeben, die auf die Notwendigkeit hingewiesen haben, in diesem Sinne (wieder) mehr zu beobachten. So weisen etwa auch Mees und Selg (1977) darauf hin, daß Psychologen ihren Gegenstand zu wenig kennen. Sie zitieren in diesem Zusammenhang Barker und Wright (1971), die schreiben, dies unterscheide die Psychologie von vielen anderen Wissenschaften. Man wisse, wie häufig Kalium vorkomme, wieviel Prozent diesen oder jenen Gesteins es ausmache und so weiter, aber man wisse fast nichts über Konflikte, Lachen, Reden etc. Hier liege eine monumentale und

unbewältigte Aufgabe für die Psychologie. Auch Bischof ist überzeugt, daß wir Psychologen „... nur dann, wenn wir wieder zu den hartnäckigen und unbeirrbaren Beobachtern geworden sind, wie es die Begründer unserer Wissenschaft einmal waren" (1989, S. 204), auf eine ertragreiche Wissenschaft, auf substantielle, realitätsnahe und nicht banale Theorien und Ergebnisse hoffen dürfen.

Diese Haltung, die heuristische Beobachtung, steht in der Tat in einem gewissen Gegensatz zum Experiment. Der Experimentator kontrolliert die Untersuchungsbedingungen, er manipuliert sie absichtlich und systematisch (denn das ist die effizienteste Form, sie zu kontrollieren), er mischt sich ein, er hat ein spezielles Interesse, eine spezielle Fragestellung, er geht von einer expliziten Hypothese aus, die er anhand konkreter Daten überprüfen will, er registriert nur einen speziellen (kleinen) Ausschnitt der Wirklichkeit (weil er sich nur dafür interessiert). Man kann diese beiden Haltungen auch in einer etwas anderen Beschreibung einander gegenüberstellen: Der heuristische Beobachter beobachtet, um Hypothesen, Zusammenhänge, vielleicht sogar Theorien zu *generieren*, der Experimentator experimentiert, um Zusammenhänge, Hypothesen oder Theorien zu *überprüfen*. Noch anders ausgedrückt: Der heuristische Beobachter geht *induktiv* vor, d.h. er schließt aus dem Speziellen, das er beobachtet, auf das Allgemeine: die Theorie. Der Experimentator arbeitet dagegen *deduktiv*: Er leitet aus der Theorie eine konkrete Hypothese ab und konfrontiert sie mit speziell zu diesem Zweck erhobenen Daten (zu diesem Vorgehen gleich mehr; s.u.: Exkurs).

(2) Deduktive Beobachtung: Beobachtung als Methode

Von dieser ersten Bedeutung des Begriffs „Beobachtung" (Beobachtung als Haltung; heuristische Beobachtung) ist eine zweite Bedeutung zu unterscheiden; wir wollen sie im folgenden die deduktive Beobachtung nennen. Mit ihr ist eine *Methode der Datenerhebung* angesprochen. Vor allem von dieser Bedeutung des Begriffs ist in diesem Buch die Rede (Ausnahmen sind erkennbar), und hauptsächlich sie ist durch die in Abschnitt 1.3 skizzierte Kennzeichnung von wissenschaftlicher Beobachtung zutreffend charakterisiert. Die deduktive Beobachtung steht im Prinzip gleichrangig neben anderen Datenerhebungsmethoden wie etwa Gesprächsmethoden (z.B. in Interviews), Fragebogen (z.B. zur Persönlichkeit), standardisierten Tests (z.B. Intelligenztests) und apparativen Verfahren (z.B. bei EEG-Messung oder Reaktionszeitmessung mit dem Computer). Diese durch systematische Beobachtung gewonnenen Daten dienen der Untersuchung von Hypothesen oder Vermutungen. Sie sind aus einem speziellen Interesse, aus einer konkreten Fragestellung heraus entstanden, sie repräsentieren einen absichtlich und systematisch ausgewählten Aspekt der Wirklichkeit. Der deduktive Beobachter ist ausdrücklich gehalten, *nicht* auf andere Aspekte des beobachteten Verhaltens, der beobachteten Personen zu achten. Diese deduktive Beobachtung steht natürlich nicht in irgendeinem Gegensatz zur experimentellen Haltung. Sie ist im Gegenteil bei bestimmten Fragestellungen bzw. Forschungsfragen ein Bestandteil eines Experimentes. Das Experiment besteht, wie oben bereits angeklungen ist, aus der systematischen Kontrolle (ggf. Variation) von Bedingungen und Konstellationen,

unter deren verschiedenen Ausprägungen verschiedene Ergebnisse erwartet werden. Etwas technischer gesprochen: Man kontrolliert[4] die Randbedingungen und variiert die (theoretisch) relevanten *unabhängigen Variablen*, um seitens der *abhängigen Variablen* einen (vorhergesagten) Effekt zu erhalten (vgl. dazu ausführlich z.B. Bortz & Döring, 1995; Lüer, 1987). Diese abhängige Variable muß also gemessen werden, und eine der hierbei zur Verfügung stehenden Methoden ist eben die Beobachtung: Wir können das (unter den verschiedenen experimentell kontrollierten oder manipulierten) Bedingungen auftretende Verhalten auch *beobachten*, nicht nur mit Instrumenten, Fragebögen oder auf andere Weise messen. Mitunter ist es nötig oder sinnvoll, daß wir die Bedingungen planen, aber nicht experimentell herstellen oder manipulieren, also in einer bestimmten Umgebung unter bestimmten Bedingungen beobachten. Häufig werden wir aber einfach gezwungen sein, unter den natürlichen Bedingungen zu beobachten und dabei diese Bedingungen möglichst genau zu registrieren, wie sie - sozusagen natürlich variiert - auftreten. Indem wir auf die Wiederholung derselben Bedingungen warten, können wir dann (etwas langwieriger und in gewisser Weise unter der Bedingung, daß die Umstände „mitspielen") schließlich doch zu einer „systematischen Variation" der Umstände kommen und damit zu einer Einschätzung darüber, was unter welchen Bedingungen wie auftritt. Diese Bedeutung des Begriffs „Beobachtung" mit „Experiment" zu vergleichen oder beide gegenüberzustellen wäre natürlich eine Kategorienverwechslung der oben erläuterten Art: Der deduktive Beobachter *ist* ein Experimentator (vgl. dazu auch die Beispiele in Abschnitt 1.1 und 1.2).

Selbstverständlich ist diese deduktive Beobachtung auch für die praktische Tätigkeit eines Psychologen von Bedeutung, etwa im Falle eines Therapeuten, der sich Klarheit über ein verhaltensgestörtes Kind verschaffen will und dementsprechend zunächst das Spiel- und Interaktionsverhalten des Kindes beobachten wird. Auch hier kann er dieses Verhalten durchaus nicht nur im Sinne von heuristischer Beobachtung freischwebend, sondern - unter Bezugnahme auf sein in einer langen Ausbildung zweifellos erworbenes Wissen - im Sinne von deduktiver Beobachtung systematisch hypothesentestend beobachten. Wenn im folgenden vorrangig von Forschungskontexten die Rede ist, muß die Übertragbarkeit dieser Methoden auf praktische Zusammenhänge etwa der Verhaltensdiagnostik immer mitgedacht werden (zur ersten Einführung vgl. etwa Bayer, 1974). Und um es abschließend explizit klarzustellen: Es geht hier nicht darum, eine der beiden Haltungen (heuristische versus deduktive Beobachtung) zu kritisieren bzw. zu bevorzugen. Sicher haben beide ihre Vorzüge.

Exkurs: Wozu überhaupt Daten? Ein kurzer Blick in die Wissenschaftstheorie
Wozu brauchen wir aber überhaupt Daten (z.B. Beobachtungsdaten) in der Wissenschaft? Was steckt eigentlich hinter der Idee einer „deduktiven" Prüfung? Zunächst: Wir brauchen Daten nicht unbedingt, um Theorien oder Hypothesen zu *generieren*. Allerdings *können* wir sie dazu gebrauchen: Beobachtungen bringen uns manchmal auf Ideen (Newtons berühmter

[4] Mit „Kontrolle" ist in diesem Zusammenhang nicht nur die reine Kenntnisnahme eines Zustandes gemeint (etwa in dem Sinne, in dem man bei der Inventur die Lagerbestände „kontrolliert"), sondern auch die Herstellung eines gewünschten Zustandes durch entsprechende Maßnahmen (etwa in dem Sinne, in dem ein Heizungsthermostat die Raumtemperatur „kontrolliert").

Apfel), aber sicher auch nur dann, wenn unser Kopf entsprechend „vorbereitet" und „aufnahmebereit" ist (denn das Fallen eines Apfels hat - vor Newton - niemanden zur Entwicklung einer Gravitationstheorie angeregt). An diesem Beispiel sieht man nebenbei, daß die systematische Beobachtung zur Generierung von Hypothesen zunächst um nichts besser geeignet ist als die „anekdotische Betrachtung" (allerdings mag es sein, daß man durch sie auf ergiebigere Gedanken kommt).

Daten taugen auch nicht dazu, allgemeine Vermutungen (Theorien oder Hypothesen) zu *beweisen* (zu verifizieren).[5] Wenn ich eine Vermutung aus einer bestimmten Menge beobachteter Tatsachen ziehe („Ich habe Hunderte von Schwänen in meinem Leben gesehen, und alle waren weiß. Vielleicht sind ja tatsächlich *alle* Schwäne weiß."), das heißt, wenn ich eine abstrakte Beschreibung finde, die diese verschiedenen, sozusagen „individuellen" Tatsachen in wichtiger Hinsicht verbindet, dann stimmt diese Beschreibung natürlich mit diesen Beobachtungen überein. Das beweist aber für sich genommen leider noch nichts. Man kann das auch etwas hochgestochener ausdrücken: Der Schluß vom faktisch Gegebenen auf das Mögliche und auf das Zukünftige ist unzulässig.

Aber selbst das Eintreffen eines weiteren Ereignisses beweist die allgemeine Theorie nicht („Siehe da: schon wieder ein Schwan! Und siehe da: auch er ist weiß!"), nicht einmal dann, wenn es aus der Theorie vorhergesagt wurde („Wenn wirklich *alle* Schwäne weiß sind, dann *muß* ja auch der, der im Teich des Stadtparks von Wien schwimmt - den ich noch nicht gesehen habe - weiß sein. Tatsächlich!"). Warum nicht? Ganz einfach: Schon morgen kann uns der kritische, der widerlegende Fall begegnen: der erste schwarze Schwan, den wir in unserem Leben sehen (die gibt's übrigens tatsächlich: cygnus atratus). Es ist einfach nicht auszuschließen, daß dieser Fall irgendwann irgendwo auftritt. Damit sind wir an einem wichtigen Punkt: Beobachtungen nützen, um eine Vermutung zu widerlegen. Eine einzige Beobachtung genügt, um eine allgemeine Theorie zu *falsifizieren* („Tatsächlich, ein schwarzer Schwan! Also sind doch *nicht alle* Schwäne weiß!").

Das klingt bis hierher nicht sehr schwierig. Ist das alles, was es mit der Wissenschaft auf sich hat? Nicht ganz! Natürlich ist das Unternehmen „Wissenschaft" so einfach nun doch nicht. Es gibt zahlreiche Probleme.

Ein wichtiges: Man muß darauf achten, daß nicht unversehens das vorhergesagte Merkmal („weiß") zum *Kriterium* für die Klassifikation wird („Dieses Tier da - was immer es ist - ist schwarz, *also* kann es kein Schwan sein, denn alle Schwäne sind weiß"), oder mit anderen Worten: Man muß aufpassen daß die Vermutung nicht zur *Definition* wird („Alle Schwäne sind weiß, sonst sind es eben keine Schwäne").

Ein anderes Problem: Was macht man mit einer „falsifizierten" Theorie? Komplett „wegwerfen"? Sicher nicht ohne weiteres. Zuerst muß man natürlich die Beobachtung genau prüfen („War das Tier da wirklich ein Schwan?" Denken Sie daran: Die Tatsache, daß es schwarz war, darf dabei keine Rolle spielen!). Können wir die Beobachtung wiederholen (*replizieren*)? Falls ja: Können wir die Theorie - durch leichte Änderungen - so verbessern, daß sie weiterhin nützt, aber mit den Tatsachen nicht mehr in Widerspruch steht (z.B.: „Alle europäischen Schwäne sind weiß, nur australische Schwäne sind schwarz")? Diese verbesserte Theorie muß sich nun erstens in neuen Prognosen bewähren, sie darf zweitens nicht ihrerseits den Fehler begehen, Hypothese und Definition zu verwechseln (z.B.: „Australische Schwäne *erkennt* man an ihrer schwarzen Farbe"), und sie darf drittens durch ihre Anpassung nicht völlig „immun" gegen zukünftige Widerlegungen werden. Dieser letzte Punkt ist vermutlich der wichtigste. Was ist damit gemeint? Nun, die angepaßte Theorie in

---

5 Die Betonung liegt hierbei auf *allgemeine* Vermutung (wie „*Alle* Gase dehnen sich bei Erwärmung aus"). Spezielle Vermutungen („Mein Schal liegt vielleicht in dieser Schublade") lassen sich durch die Beobachtung des Schals in eben dieser Schublade natürlich schon beweisen. Das grundsätzliche Problem, das sich allerdings auch in diesem einfachen Beispiel stellt - die Theorieabhängigkeit *jeder* Beobachtung - ist Gegenstand eines eigenen Abschnitts (3.1).

unserem Beispiel ist nicht immun, denn sie könnte durch einen europäischen schwarzen Schwan widerlegt werden (wieder Achtung: Es darf *nicht* von seiner Farbe abhängen, ob ein Schwan ein europäischer Schwan ist!). Wenn wir unsere Hypothese dagegen folgendermaßen „verbesserten": „Alle Schwäne sind weiß – außer, wenn sie schwarz sind", kann kaum noch ein widerlegender Fall auftreten. Allerdings nur „kaum": Es könnte ja immerhin noch grüne, rote oder blaue Schwäne geben. Wenn man also ganz sicher gehen will: „Alle Schwäne sind weiß - außer sie sind nicht weiß". Dies aber, das kann man sofort sehen, ist keine Vermutung, sondern eine Gewißheit. Anders formuliert: Diese Behauptung hat *keinen empirischen Gehalt* mehr. Es handelt sich vielmehr um eine sogenannte Tautologie, also einen Satz, der *aus logischen Gründen immer* wahr ist, ganz unabhängig davon, was mit den Schwänen dieser Welt tatsächlich los ist.

(3) Menschliche Beobachtung als Grundlage aller Forschung

Der Vollständigkeit halber sollte noch ein weiteres Verständnis des Begriffs „Beobachtung" kurz angesprochen werden. Gelegentlich findet sich der Hinweis, Beobachtung läge eigentlich *jeder* Untersuchung zugrunde; der Gegensatz sei schon von daher ganz unzutreffend. „In jedem psychologischen Verfahren zur Datensammlung kommt irgendeine Form der Beobachtung zur Anwendung. Ob Experiment, Test- bzw. Fragebogenuntersuchung oder Interview – immer wird - zumindest im weiten Sinne - beobachtet" (Mees & Selg, 1977, S. 15). Ähnlich von Cranach und Frenz (1969): „Beobachtung ist eine grundlegende Methode der Psychologie und in fast allen Verfahren der Datenerhebung enthalten" (S. 269).

Der Punkt, um den es in diesem Zusammenhang geht, ist der folgende. Natürlich forschen immer Menschen. Ihre Interessen und ihre Vorgehensweise (Methoden) sind insofern immer auch „menschlich", als sie letztlich auf den Menschen als „Endadressat" zugeschnitten sein müssen: Ein Bericht im Ultraschallbereich nützt niemandem. Selbst wenn wir ein Meßinstrument benutzen, so könnte man argumentieren, muß es immer eines sein, von dem wir letztlich ablesen können. In diesem Sinne „beobachten" wir immer. Dies ist freilich ein sehr weites Verständnis des Begriffs Beobachtung; es schließt das „Beobachten" des Voltmeters einer physikalischen Versuchsanordnung ebenso ein wie die tagelange minutiöse Beobachtung eines Kindes. Damit wird freilich der Begriff vermutlich überdehnt.

Mit dieser Antwort lösen wir allerdings ein grundsätzliches Problem noch nicht. Selbst wenn wir das Meßinstrument nicht bei seinem konkreten Einsatz „beobachten", muß es doch irgendwann vor seinem Einsatz überprüft worden sein: Wir müssen sichergestellt haben, daß es das (zuverlässig) mißt, was es messen soll. Wie aber können wir das sicherstellen? Anhand von Kriterien, die uns im Verhältnis zu jenem Instrument nicht oder jedenfalls deutlich weniger zweifelhaft sind. Und dies wird sehr häufig - vielleicht über mehrere Zwischenschritte - die menschliche unmittelbare „Anschauung" sein (wir kommen auf diesen Punkt wieder zurück; s. Abschnitte 3.2.1 und 4.4.2).

1.5 Einordnung: Vorgehensweisen, Unterschiede, Klassifikationsmöglichkeiten

Verschiedene übliche, in der Literatur vorfindliche Begriffsdifferenzierungen bzw. Klassifikationen von „Beobachtung" lassen sich nun kurz und recht zwanglos einführen und erläutern.[6] Mit „unsystematischer" oder auch „freier" Beobachtung ist in aller Regel so etwas wie heuristische Beobachtung gemeint, d.h. ein ungesteuertes und nicht „vorsortiertes" Betrachten ohne besondere Fragestellung. „Wissenschaftliche" Beobachtung, insbesondere im Sinne von deduktiver Beobachtung, ist immer systematische Beobachtung. Nur wer eine konkrete Frage hat, eine Hypothese prüft, einer Zusammenhangsvermutung nachgeht, betreibt Wissenschaft. Ganz analog gilt dies auch für die Unterscheidung zwischen strukturierter und unstrukturierter Beobachtung (vgl. etwa Grümer, 1974, S. 36-44). Die explizite Überprüfung einer Hypothese impliziert, wie wir gesehen haben, im Normalfall zugleich die Kontrolle der Bedingungen. „Kontrollierte" Beobachtung (im Gegensatz zu „unkontrollierter" Beobachtung) hebt sozusagen einen – freilich wesentlichen – Aspekt der wissenschaftlichen bzw. systematischen Beobachtung hervor, nämlich eben die Kontrolle der (Rand-) Bedingungen. Anders ausgedrückt: Der Begriff der systematischen bzw. wissenschaftlichen Beobachtung ist umfassender als der einer kontrollierten Beobachtung: Man kann auch systematisch (d.h. mit einer bestimmten Fragestellung unter einer bestimmten Perspektive einen ganz bestimmten Realitätsausschnitt) beobachten, ohne dabei die Randbedingungen (aktiv) zu kontrollieren.

Wir wollen in den folgenden Abschnitten (1) bis (4) einige weitere wichtige Klassifikationen, anhand derer sich auch verschiedene methodische Vorgehensweisen psychologischer Beobachtung erläutern lassen, kurz skizzieren, um abschließend (Abschnitt 5) diejenige Klassifikation vorzustellen, die der weiteren Darlegung in den folgenden Kapiteln des Buches als Gliederungsgedanke zugrunde liegt.

(1) Technisch vermittelte versus unvermittelte Beobachtung

Glück (1971) nennt diese Klassifikationsdimension die der Hilfsmittel. Es ist unmittelbar evident, was mit dieser Kennzeichnung gemeint ist: die Frage, ob zwischen beobachtetem Sachverhalt und Beobachter ein Hilfsmittel als Informationsüberträger „zwischengeschaltet" ist. Allerdings ist das nicht ganz so leicht zu entscheiden, wie es auf den ersten Blick aussieht: Ist mit technischer Vermittlung erst eine komplizierte Videoanlage, eine optische Unterstützung (Fernglas mit Nachtlichtverstärkung) oder schon eine normale Sehhilfe (Brille) angesprochen? Sinnvoll erscheint es (etwa mit Graumann, 1966), die *Speicherbarkeit* der Beobachtung (Film, Ton) als Kriterium der Unterscheidung heranzuziehen.

---

[6] Möglicherweise entspricht das im folgenden eingeführte Verständnis der entsprechenden Begriffe nicht immer dem Gebrauch in allen Quellen oder Diskussionen. In diesen Fällen ist unser Vorgehen als Vorschlag zur Vereinheitlichung aufzufassen, nicht als Ignoranz.

Wichtig ist in diesem Zusammenhang die Frage, ob das verwendete Instrumentarium einen (verzerrenden) Einfluß auf das Ergebnis hat. Kent und Foster (1977) diskutieren Studien zur Frage, ob das Beobachtungsmedium einen Einfluß auf die Zuverlässigkeit der Beobachtung hat. Sie kommen zu dem beruhigenden Ergebnis, daß dem im allgemeinen nicht so ist (außer bei „Sprache" als zu beobachtendem Phänomen): Zwischen verschiedenen Medien (z.B. Spiegel oder Video) gab es nicht einmal dann bedeutsame Unterschiede, wenn die Beobachter in diesen Medien unterschiedlich große Erfahrung hatten. Manns et al. (1987, S.25) diskutieren neuere Literatur, die nach ihrer Einschätzung zu uneinheitlichen Ergebnissen kommt. Sie empfehlen den Einsatz etwa von Videogeräten vor allem wegen der Wiederholbarkeit der Beobachtungen. Neben dieser ist vor allem die Möglichkeit der verlangsamten Wiedergabe (Zeitlupe) ein weiterer wichtiger Vorteil vermittelter Beobachtung. Dem stehen allerdings auch Nachteile gegenüber: So wird etwa die systematische Ausblendung irrelevanter Information erschwert, die Menschen in natürlicher Umgebung offenbar mühelos beherrschen („Cocktailparty-Phänomen": Man kann einem Gespräch folgen, obwohl in der näheren Umgebung zahlreiche andere Gespräche in gleicher Lautstärke stattfinden; vgl. etwa Hussy, 1986, S. 173ff.). Manns et al. (1987) weisen außerdem darauf hin, daß Videoaufnahmen im Falle der Aufnahme vieler Personen bei häufigen Ortswechseln der beobachteten Personen ungenau werden können. So schränkt beispielsweise ein zu enger oder starrer Blickwinkel die Aufnahme ein, oder die Kamera kann nicht allen Bewegungen schnell genug folgen. Mittenecker (1987) weist in seiner ausführlichen Darstellung u.a. darauf hin, daß Kameraschwenks und Zoomingeffekte auch aufmerksamkeitssteuernd und -selegierend wirken können.

(2) Labor- versus Feldbeobachtung, naturalistische Beobachtung

Der Unterschied zwischen Feld- und Laborbeobachtung (zur Einführung siehe z.B. Glück, 1971) bezieht sich auf den Ort bzw. die Umgebung, in dem/der beobachtet wird. Sofern darüber hinausgehende Unterschiede (beispielsweise unterschiedliche *Kontrolle* der Randbedingungen; s.o.) nicht impliziert sind, ist die bloße Kennzeichnung des Beobachtungsortes solange wenig aussagekräftig, wie nicht aufgrund konkreter empirischer oder theoretischer Argumente Unterschiede im beobachteten Verhalten oder in den Beobachtungsleistungen erwartet werden müssen. Die wichtigste Form der Feldbeobachtung ist die „naturalistische" Beobachtung, d.h. die Beobachtung eines Verhaltens unter den natürlichen Bedingungen, in denen es normalerweise auftritt (zum Überblick vgl. z.B. Longabaugh, 1980).

Naturalistische Beobachtung wurde bereits früh, aber insgesamt eher selten verwendet. Sie ist als Methode nach der Einschätzung von Longabaugh (1980) ebenso schwierig wie notwendig, insbesondere für interkulturelle Forschungsvorhaben oder in der Ethologie (vgl. etwa Blurton Jones, 1972; Eibl-Eibesfeld, 1984). Das bekannteste Beispiel einer naturalistischen Beobachtung in der Psychologie sind vermutlich die Arbeiten von Barker und Wright (1971; wir gehen auf ihren Ansatz ausführlicher in Abschnitt 2.1 ein).

(3) Offene versus verdeckte Beobachtung

Im Grunde ist auch hier unmittelbar klar, welches spezielle Merkmal einer Beobachtung damit angesprochen ist. Es geht darum, ob die Tatsache des Beobachtens vor dem Beobachteten verborgen wird („verdeckt"; ein Beispiel dafür ist etwa die Studie von Barash, 1972, die wir in Abschnitt 4.5 darstellen) oder ihm deutlich gezeigt oder mitgeteilt wird („offen"). Ein wichtiges Problem der verdeckten Beobachtung besteht sicher darin, daß sie häufig ethisch nicht unbedenklich ist. Davon abgesehen ist sie vor allem deshalb oft die Methode der Wahl, weil sie sicherstellt, daß Effekte zu Lasten der Beobachtung weitestgehend ausgeschlossen sind, d.h. daß die Tatsache, daß beobachtet wird, auf den beobachteten Sachverhalt keinen Einfluß hat (Kapitel 3.2.3 diskutiert diese Effekte ausführlich).

(4) Teilnehmende versus nicht-teilnehmende Beobachtung

Häufig kann es nützlich oder sogar notwendig sein, einen Beobachter einzusetzen, der sozusagen selbst Teil des Geschehens ist bzw. wird (zur Einführung vgl. etwa Grümer, 1974, S. 45-53; Jahoda, Deutsch & Cook, 1968). Bei einer teilnehmenden Beobachtung hat der Beobachter weitere (aktive, gewollte) Rollen im Geschehen. Häufige Beispiele hierfür sind Beobachtungen in Partnerschaften (durch einen oder beide Partner; z.B. Jarrett & Nelson, 1984; Jacobson & Moore, 1981) oder Unterrichtssituationen (beispielsweise durch den Lehrer; z.B. Hay, Nelson & Hay, 1980). Diese Klassifikation ist (im Gegensatz zur Einschätzung etwa von Glück, 1971) unabhängig von der vorigen (verdeckte vs. offene Beobachtung). Man kann auch verdeckt als teilnehmender Beobachter arbeiten; das wird faktisch sogar nicht selten der Fall sein (vgl. die Beispiele bei Friedrichs & Lüdtke, 1973). Natürlich gibt es erst recht verdeckte oder offene nicht-teilnehmende Beobachter. Eine ausführlichere Darstellung bietet etwa Jorgensen (1989); dort findet man u.a. auch Hinweise auf Zeitschriften, in denen entsprechende Studien publiziert werden. Jorgensen empfiehlt teilnehmende Beobachtung insbesondere dann, wenn

(a) wichtige Unterschiede zwischen der Perspektive von „Insidern" und „Outsidern" zu erwarten und zu beachten sind (Jorgensen nennt als Beispiele Okkultisten, Pokerspieler, Anhänger von FKK, sowie bestimmte Beschäftigungen: Politik, Wissenschaft etc.),

(b) über das zu untersuchende Phänomen wenig bekannt ist (ein besonderer Effekt der teilnehmenden Beobachtung ist das „becoming the phenomenon" [Jorgensen, 1989, S. 62], die Tatsache, daß man ein Phänomen nicht einfach von außen beobachtet, sondern ein Beispiel für das wird, was man beobachten will[7]),

(c) die zu beobachtenden Phänomene normalerweise verborgen bzw. schwer zugänglich sind; Jorgensen nennt als Beispiel Vorgänge in der Familie oder in religiösen Gemeinschaften,

(d) die zu beobachtenden Phänomene systematisch versteckt werden bzw. bestimmte Gruppen anders nicht zugänglich sind (Beispiele sind hier vor allem kriminelle Aktivitäten).

---

[7] Dies ist freilich eher ein Vorgehen im Sinne von heuristischer Beobachtung.

Insbesondere die Argumente (c) bzw. (d) sind überzeugend (vgl. hierzu auch Mees & Selg, 1977): Für manche Fälle bzw. Bereiche wird nicht-teilnehmende Beobachtung (durch fremde Beobachter) praktisch aussichtslos sein, es sei denn, man kann verdeckt beobachten. Besonders deutlich wird das in Bereichen delinquenten oder devianten Verhaltens, beispielsweise Straftäter („Wie kommunizieren Mafiamitglieder?") oder Neonazis. Beispiele für teilnehmende Beobachtungen finden sich in der Edition von Friedrichs und Lütke (1973), bei Bain (1968) oder Kluckhohn (1968), die wegen ihrer gewissermaßen impressionistischen Anschaulichkeit lesenswert sind.

Ein wichtiges Argument für „teilnehmende Beobachtung" ist häufig die Annahme, daß sogenannte „Reaktivitätseffekte" (d.h. Reaktionen der Beobachteten auf die Tatsache, daß sie beobachtet werden) in diesen Fällen weniger auftreten, bzw. - anders formuliert - daß die Reaktivität der Beobachteten bei nicht-teilnehmender offener Beobachtung höher sei. Es gibt jedoch Hinweise darauf, daß das so nicht stimmt. Die vorliegenden Studien hierzu zeigen, daß auch im Fall teilnehmender Beobachtung Reaktivität auftritt (z.B. Hay, Nelson & Hay, 1980). Je nach Situation kann das sogar stärker sein, z.B. deswegen, weil das gewohnte Verhalten der als Beobachter eingesetzten Personen sich ändert. Wir kommen auf diesen wichtigen Punkt im Zusammenhang mit der Diskussion der möglichen Fehlerquellen bei Beobachtungen zurück (Abschnitt 3.2.3).

Durch die Beteiligung des teilnehmenden Beobachters am Geschehen, die ja auch emotionale Beteiligung bedeuten bzw. zur Folge haben kann, können zusätzlich auch Selektions- und Verzerrungseffekte bei den Beobachtern selbst auftreten (Mees & Selg, 1977). So weisen etwa Manns et al. (1987) darauf hin, daß teilnehmende Beobachter im allgemeinen unzuverlässiger sind, weil sie durch ihre „Doppelrolle" gegebenenfalls zu widersprüchlichen oder unvereinbaren Handlungen gezwungen bzw. Emotionen veranlaßt werden.

> Auch dafür ein Beispiel: Margolin, Hattem, John und Yost (1985) ließen 30 Ehepaare Filmaufnahmen von sich und fremden Paaren einschätzen. Die Übereinstimmung der Beobachtungen der Ehepartner mit trainierten externen Beobachtern war im Falle der fremden Paare wesentlich höher. Die Übereinstimmung zwischen den Ehepartnern war immer höher als zwischen einem Ehepartner und einem externen Beobachter (sowohl bzgl. des eigenen wie der fremden Bänder). Diese fehlende Übereinstimmung bei den eigenen Bändern weist aber nach der Ansicht von Margolin et al. (S. 245) nicht unbedingt auf fehlende Genauigkeit in einem der Fälle hin. Man könnte vielleicht auch sagen, daß beide jeweils eine andere Wirklichkeit sahen.

Hinzu kommt, daß oft das Protokoll nicht unmittelbar sondern retrospektiv angefertigt wird, was zusätzliche Fehler durch verzerrte oder lückenhafte Erinnerung möglich macht (auch dazu mehr in Abschnitt 3.2.3).

Soll man nun teilnehmende Beobachtung empfehlen oder nicht? Darauf gibt es keine generelle Antwort. Es kommt erstens auf die Absicht an: Welche Nachteile kann man bei einer gegebenen Untersuchungsabsicht eher in Kauf nehmen? Es kommt zweitens darauf an, welche Störquellen im einzelnen für die jeweiligen Nachteile und Verzerrungen verantwortlich sind: Wie

kann man sie identifizieren, wie kontrollieren, ggf. sogar eliminieren? (Diesen Punkt werden wir noch ausführlicher diskutieren müssen; vgl. Abschnitt 3.2.2.)

Dies weist auf einen Punkt von grundsätzlicher Bedeutung hin. Eine Einteilung von Beobachtungen in teilnehmende und nicht-teilnehmende (und dieses Argument gilt - mutatis mutandis - auch für die anderen Taxonomien) bleibt gewissermaßen an der Oberfläche. Natürlich ist es möglich, verschiedene Studien nach diesem Gesichtspunkt zu ordnen (jede Beobachtungsstudie ist entweder teilnehmend oder nicht-teilnehmend). Viel wichtiger ist jedoch die Frage, warum dieses oder jenes Vorgehen gewählt wurde. War für diese teilnehmende Beobachtung die These ausschlaggebend, daß teilnehmende Beobachter weniger mit Reaktivitätseffekten zu kämpfen haben? War für jene nicht-teilnehmende Beobachtung der Einwand entscheidend, daß aus theoretischen oder empirischen Gründen bezüglich des interessierenden Verhaltens keine relevanten Verzerrungen durch Reaktivität erwartet werden mußten? Die Frage, ob teilnehmend, ob verdeckt, ob vermittelt oder nicht beobachtet werden soll, ist nur im Kontext der konkreten Forschungsfragestellung sinnvoll und beantwortbar.

Wir werden daher nicht das Vorgehen bei der Beobachtung, sondern den Grad der Reduktion durch die Beobachtung zur Grundlage der weiteren Gliederung machen: Wieviel vom Geschehen wird erfaßt? Diese Dimension macht im Gegensatz zu den anderen a priori (d.h. von vorneherein) einen Unterschied für das *Ergebnis* der Beobachtung. Anders ausgedrückt: Ob es einen bedeutsamen Unterschied zwischen teilnehmender und nicht-teilnehmender, zwischen offener und verdeckter oder zwischen vermittelter und unvermittelter Beobachtung hinsichtlich des Ergebnisses gibt, hängt selbst bei ein und derselben Fragestellung von vielerlei Umständen ab und ist jedenfalls bei unterschiedlichen Fragestellungen eine offene, empirisch zu klärende Frage. Hingegen ist von vorneherein klar, daß und worin sich zwei Beobachtungen im Ergebnis unterscheiden, die mit in diesem Sinne unterschiedlichen Reduktionsgraden arbeiten. Die Frage, welchen Grad an Reduktion man erzielen will bzw. in Kauf zu nehmen bereit ist, ist damit freilich nicht beantwortet.

(5) Klassifikation nach dem Grad der Reduktion

Mees und Selg (1977) unterscheiden grundsätzlich drei Arten (besser: Stufen) der Verarbeitung von Beobachtung: isomorphe Beschreibung, reduktive Beschreibung und reduktive Einschätzung (ähnlich unterteilen etwa Glück, 1971; Faßnacht, 1995; Schaller, 1980; Longabaugh, 1980). Diese Dimension kennzeichnet den Grad der Reduktion (bzw. der Vollständigkeit der Repräsentation) der Phänomene, d.h. wieviel Information ausgewählt, zusammengefaßt bzw. nicht beachtet wird. Hierbei sind zwei Aspekte unterscheidbar: (a) der Grad der Selektion, d.h. der Auswahl der registrierten Ereignisse, und (b) der Grad der Interpretation bzw. Bewertung, d.h. der Zusammenfassung verschiedener (gewichteter) Informationen zu einem Urteil. Eine Tonbandaufzeichnung beispielsweise läßt zwar einerseits alle optischen Reize unregistriert, ist aber natürlich andererseits wesentlich vollständiger als etwa ein Stenogramm, weil dieses, selbst wenn es ansonsten vollständig ist, alle paraverbalen Äußerungen (Äh's, Räuspern, Ki-

chern, Keuchen etc.) nicht registriert. Gewissermaßen am anderen Extrem dieses Kontinuums zwischen vollständiger Replikation und totaler Reduktion ist eine ganz allgemeine „grobe" Einschätzung des Beobachteten einzuordnen („Hat dieser Lehrer die Stunde kompetent geleitet?"). Die wichtigen und sehr häufig verwendeten Zeichen- und Kategoriensysteme (über die wir in den Kapiteln 4 und 5 ausführlicher sprechen werden) sind je nach Umfang mehr oder weniger selektiv und darüber hinaus in aller Regel deutlich interpretierend, weil sie typischerweise Interpretations- oder auch Bewertungsleistungen des Beobachters voraussetzen: War dieser Satz ein Fall von ..., war diese Armbewegung ein Zeichen für ..., war dieses Verhalten aggressiv oder offensiv, war diese Äußerung informativ oder bewertend?

Das bekannteste Beispiel für den Versuch, möglichst vollständig alles beobachtbare Verhalten wiederzugeben, sind die Studien von Barker und Wright (1971), deren Ansatz wir im folgenden Kapitel (Abschnitt 2.1) ausführlicher darstellen wollen. Wenn man ganz ausführlich und im Detail das Beobachtete wiedergeben will, hat man natürlich - selbst wenn einem das gut gelänge - mit der weiteren Verwendung dieser Datenmengen auch unabhängig von der verwendeten Kodierung (sei sie sprachlicher oder anderer Art) erhebliche Probleme. Beispielsweise findet sich bezüglich der Mikroanalysen non-verbaler Verhaltensweisen eine detaillierte Diskussion bei Scherer (1974), die durch seine anschauliche Schilderung den notwendigen Aufwand deutlich macht.

Wir beginnen im folgenden Kapitel 2 die Darstellung wissenschaftlicher Beobachtungsmethoden, indem wir den nächstliegenden Weg der Wiedergabe diskutieren: Die *sprachliche Beschreibung* dessen, was man beobachtet. Dieses Verfahren hat einen sehr niedrigen Reduktionsanspruch: Es geht darum, in einem vertrauten Medium möglichst vollständig und verständlich alles wiederzugeben, was man beobachtet hat. Wenn man diesen Weg wählt, ergeben sich jedoch verschiedene, z.T. grundsätzliche Schwierigkeiten.

Literaturempfehlungen

Übersichtsarbeiten: Faßnacht (1995) und – etwas allgemeiner – Grümer (1974). Eine neuere, eher praktisch orientierte Einführung ist das Buch von Martin und Wawrinowsli (1991). Ebenfalls informativ ist die Edition von Mees und Selg (1977). Die Monographie von Haynes (1978) ist ziemlich ausführlich, enthält viele Literaturhinweise, ist jedoch als Einführungslektüre nur bedingt geeignet. Empfehlenswerte Artikel sind z.B. die Arbeiten von Schaller (1980; guter einführender Überblick, weiterführende Literatur), Erdfelder (1994), Glück (1971), Graumann (1966), Feger und Graumann (1983) oder Hasemann (1983); eine kurze Einführung geben Bortz und Döring (1995, S. 240-254). Eine gute Einführung in die Wissenschaftstheorie gibt Chalmers (1989). Eine Reihe von Beispielen für kleine Untersuchungen, die man beispielsweise auch im Rahmen von Lehrveranstaltungen wiederholen kann, um praktische Erfahrungen zu sammeln, finden sich z.B. in Bungard und Lück (1974), auch in Grümer (1974).

Kapitel 2

Beschreiben mit natürlicher Sprache

Als „natürlichste" Wahl für eine Zeichenmenge, der wir unsere Beobachtungen zuordnen, scheint sich unsere Alltagssprache anzubieten. Mit der größten Selbstverständlichkeit beschreiben wir mit ihrer Hilfe in unserem Alltag all das, was wir wahrnehmen bzw. beobachten; warum nicht diese Gabe für wissenschaftliche Beobachtung nutzen? Tatsächlich wurden in verschiedenen Zusammenhängen derartige Beobachtungssysteme vorgeschlagen, die die Alltagssprache in ihrer ganzen Fülle und mit ihrer ganzen Syntax zur Beschreibung gebrauchen (Faßnacht, 1995). Wir wollen in diesem Kapitel die Vorteile, aber auch die Probleme dieses mächtigen „Instrumentes" diskutieren. Es wird sich dabei zeigen, daß die Alltagssprache zwar nicht so untauglich ist, wie ihr Ruf in vielen Diskussionen es befürchten lassen könnte, aber daß man sich, wenn man sich ihrer für Beobachtungsstudien bedient, eine Reihe von Implikationen „einkauft", die sich unter Umständen als Nachteil herausstellen können. Wir wollen dies insbesondere am Beispiel der Handlungsbegriffe diskutieren. Zur Veranschaulichung und Einführung stellen wir zunächst das wohl bekannteste Beispiel einer Beobachtungswiedergabe mittels der Alltagssprache ausführlicher vor: die Studien von Roger G. Barker und Herbert F. Wright (orig. 1955, wir zitieren im folgenden den Reprint von 1971). Dieses berühmte Beispiel führt in das Vorgehen mit Hilfe eines „Verbalsystems" (Faßnacht, 1995) besser ein als lange Erläuterungen.

2.1 Das bekannteste Beispiel: Die Arbeiten von Barker und Wright

Ausgangspunkt des Forschungsansatzes von Barker und Wright (1971) war die Feststellung, daß in der Psychologie im Gegensatz zu anderen Wissenschaften (Biologie, Physik etc.) keine Phase des „Sammelns" der grundlegenden Phänomene stattgefunden habe (vgl. auch Abschnitt 1.4.2): sie wisse einfach zuwenig von ihrem Gegenstand. Fragen wie beispielsweise „Wie oft lachen Kinder im Alltag?" oder „Werden Kinder heute häufiger frustriert als Kinder vor 50 Jahren?" könnten nicht beantwortet werden. Aus diesem Grund richteten Barker und Wright Ende der 40er Jahre in einer amerikanischen Kleinstadt (Oskaloosa) die „Midwest Psychological Field Station" ein. Dabei bestand das Ziel vornehmlich in der Beschreibung von Umwelt und Verhalten der Kinder in dieser Gemeinde (s.o.: heuristische Beobachtung). Im folgenden werden die zentralen theoretischen Ausgangspunkte des Forschungsansatzes vorgestellt, soweit

sie für die Verhaltensbeobachtung von Bedeutung sind. Eine umfassende Darstellung findet sich bei Barker und Wright (1971).

(1) Basiseinheiten im ökologischen Ansatz

Barker und Wright unterscheiden drei Einheiten, die sie im Rahmen ihrer ökologischen Forschung als relevant erachten:

1. Verhaltenssettings,
2. Verhaltensobjekte,
3. Verhaltensepisoden.

Bei den Verhaltenssettings handelt es sich um konstante Verhaltensmuster („standing patterns of behavior"), die unabhängig von der einzelnen Person bestehen. Ein Beispiel dafür wäre etwa ein Tennismatch. Neben der raum-zeitlichen Bindung des Verhaltensmusters (hier: Tennisspielen) an eine bestimmte Umgebung, das sogenannte nonpsychologische Milieu (hier: Tennisplatz), besteht eine „synomorphe" Beziehung zwischen Verhaltensmuster und umgebendem Milieu. Der Begriff „synomorph" bezeichnet den Sachverhalt, daß Verhalten und Umgebung zueinander passen („kongruent" sind). Das Verhalten „Tennisspielen" wird auf die Umgebung „Tennisplatz" begrenzt; umgekehrt wurde der Tennisplatz nicht dazu angelegt, Fahrradrennen auf ihm auszutragen. Die Beziehung zwischen Verhalten und Umgebung definiert damit das Verhaltenssetting.

In ähnlicher Weise wie Verhaltenssettings lassen sich auch die Verhaltensobjekte als beständige Verhaltensmuster mit synomorpher Relation zur Umgebung betrachten. Sie unterscheiden sich jedoch von Verhaltenssettings in zweierlei Hinsicht:

1. Verhaltensobjekte werden von beständigen Verhaltensmustern umgeben (während bei Verhaltenssettings die Verhaltensmuster von einem Milieu umgeben werden). Um beim obigen Beispiel zu bleiben: Ein Verhaltensobjekt wäre ein Tennisschläger, der vom Verhaltensmuster Tennisspielen umgeben wird und nicht umgekehrt.
2. Verhaltensobjekte sind in Verhaltenssettings lokalisiert, indem sie diese ausstatten. Zur Umgebung Tennisplatz und zum Verhalten Tennisspielen gehört bekanntlicher- und notwendigerweise ein Tennisschläger.

Im Zusammenhang mit der Darstellung verschiedener Formen wissenschaftlicher Beobachtung ist insbesondere die von Barker und Wright definierte Kategorie der Verhaltensepisoden von Bedeutung. Verhaltensepisoden dienen der Unterteilung des Verhaltensstroms eines Individuums und stellen die kleinsten (individuellsten) Einheiten im ökologischen Forschungskontext dar. Der erste Schritt, um zu diesen Episoden zu gelangen, ist die Erstellung eines Verlaufsprotokolls („specimen record"). Hierbei geht es ganz konkret um die detaillierte, beschreibende Darstellung des individuellen Verhaltens mit Hilfe der Umgangssprache.

(2) Was soll beobachtet und protokolliert werden?

Betrachten wir erneut das Tennis-Beispiel. Eine Möglichkeit, das Verhalten des Tennisspielers Becker bei einem Match zu beschreiben, wäre etwa die folgende: „Becker schlägt auf, stürmt ans Netz und spielt den anschließenden Volley an seinem Gegner vorbei die Linie entlang". Das gleiche Verhalten könnte auch wie folgt beschrieben werden: „Becker fährt sich mit der Zunge über die Lippen, fixiert dabei den Gegner, wirft den Ball mit der linken Hand hoch, holt mit dem rechten Arm aus und schmettert den Ball in die gegnerische Spielhälfte. Den Rückhand-Volley plaziert er nach einer halben Körperdrehung mit der Rückhand. Anschließend ballt er die Faust." Worin unterscheiden sich die Beschreibungen? Offensichtlich in der Beobachtungsebene, das heißt der „Auflösung" der Beschreibung. Im ersten Fall wird auf einer molaren Ebene versucht, Beobachtetes zu beschreiben (Aktionen); im Verhältnis dazu erfaßt die zweite Verhaltensbeschreibung deutlich elementarere, d.h. molekulare Verhaltenseinheiten (Aktone) (vgl. Abb. 4).

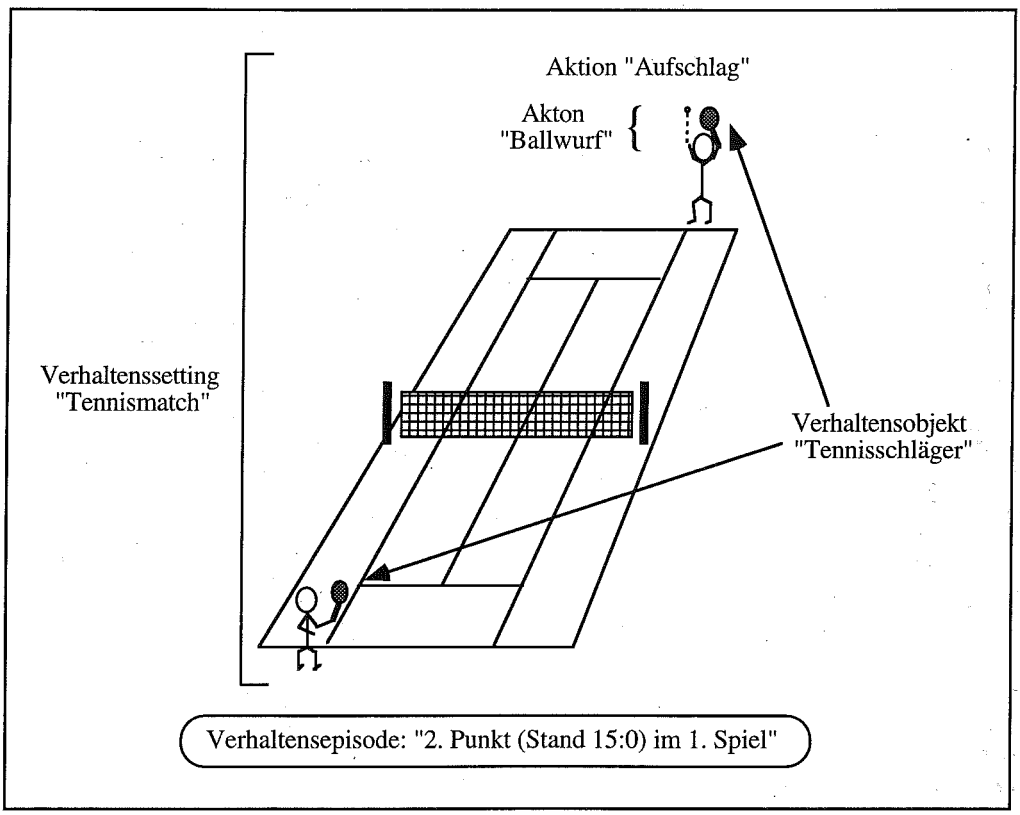


Abbildung 4: Tennismatch als Beispiel eines „Verhaltenssettings"

Während Aktionen sowohl von der Person als auch von der Umgebung als Ganzes ausgehen, sind demgegenüber Aktone als Einheiten konzipiert, die sich auf untergeordnete Aspekte von Person und Umgebung beziehen und demnach keine diskret wahrnehmbaren Ereignisse für die Person darstellen. Kriterium für die Unterscheidung ist dabei nicht, wie man vielleicht meinen könnte, der Umfang des Verhaltens, sondern der Kontext, in dem sich das Verhalten abspielt. „Ballwurf" ist Akton im Verhältnis zur Aktion „Aufschlag", aber Aktion im Verhältnis zum Akton „Öffnen der Hand" (zu den näheren Unterscheidungsmerkmalen vgl. Barker & Wright, 1971, S. 178 ff.). Mit Faßnacht (1995, S. 278) läßt sich die Unterscheidung zwischen Aktionen und Aktonen wie folgt zusammenfassen: „Aktone beschreiben das 'Wie' des Verhaltensablaufs, während Aktionen sich auf das 'Was' beziehen."

Zurück zur Ausgangsfrage: „Was soll beobachtet und protokolliert werden?" Da zwischen Aktionen und Aktonen kein eindeutiger Zusammenhang derart besteht, daß sich eine Beobachtungsebene aus der anderen eindeutig ableiten oder erschließen ließe, fordern Barker und Wright die möglichst vollständige Aufzeichnung sämtlicher Aktionen und Aktone einer Person. Damit ist das Ziel des Protokollierens jedoch nur zur Hälfte beschrieben. Ein weiterer Aspekt betrifft den Standpunkt, von dem aus der Beobachter seine Beobachtungen anstellt.

(3) Wie soll beobachtet und protokolliert werden?

Barker und Wright führen hierzu den Begriff des psychologischen Habitats ein. Darunter verstehen sie den Kontext, in dem sich (molares) Verhalten abspielt, und zwar aus der Sicht der handelnden Person. Es handelt sich hier um die Schnittstelle zwischen Person und Umgebung (nonpsychological milieu). Beide zusammen bilden ein dynamisches System - eben das psychologische Habitat.

Dem Beobachter kommt die Aufgabe zu, den Standpunkt der Person einzunehmen, die beobachtet wird. Er erreicht dies, indem er sich die Frage stellt: „Was sieht die Person und wie sieht sie es?" Barker und Wright fordern also den Beobachter dazu auf, über die reine Verhaltensbeschreibung hinauszugehen und Schlußfolgerungen über die Absichten, die den Verhaltensweisen zugrunde liegen, in das Beobachtungsprotokoll mit aufzunehmen. Andernfalls wären die Verhaltensaufzeichnungen unvollständig. Das Sich-Eindenken in die psychologische Welt der Person, die beobachtet wird, ist nach Barker und Wright nichts anderes als das, was wir im Alltag ohnehin immer tun müssen und offenbar auch können, wenn wir mit anderen interagieren. Die Beobachtungsaufgabe erfordere von daher auch keine besondere Übung, sondern nur die Anwendung bereits vorhandener Fähigkeiten (S. 206). Dementsprechend geben die Autoren auch nur grobe Hinweise für die Erstellung der Verlaufsprotokolle (vgl. Barker & Wright, 1971, S. 216-219). Der Beobachter soll vor allem beschreiben, *wie* etwas geschieht. Darüber hinausgehende Interpretationen sollen zum besseren Verständnis die Verhaltensbeschreibung ergänzen, nicht ersetzen. Zusammenfassend kann das Ziel des Beobachtungsansatzes von Barker und Wright definiert werden als die Protokollierung der Aktionen und Aktone sowie des psychologischen Habitats einer Person.

(4) Wie werden die Beobachtungsdaten aufbereitet?

Um den Aufwand der Erstellung eines Verlaufsprotokolls zu verdeutlichen, wollen wir das Vorgehen kurz skizzieren. Eine Beobachtungsperiode dauert maximal 30 Minuten. Anschliessend werden die Beobachter ausgetauscht. Für die Erstellung eines vollständigen Tagesprotokolls einer Person werden 7 bis 9 Beobachter benötigt, deren Aufzeichnungen über 300 (!) Schreibmaschinenseiten füllen. Bevor es jedoch zu dieser endgültigen Abschrift des Verlaufsprotokolls kommt, wird das Rohmanuskript in zwei sich wiederholenden Korrekturphasen überarbeitet und ergänzt, anschließend mit einem anderen Beobachter in bezug auf Verständlichkeit und Vollständigkeit durchgesprochen.

Die endgültige Abschrift des Verlaufsprotokolls bildet den Ausgangspunkt für die Einteilung des Verhaltensstroms in Verhaltensepisoden. Eine Episode kann aus einer einzelnen ununterbrochenen Aktion bestehen; ebenso lassen sich aber auch Episoden ausmachen, die aus mehreren Phasen bestehen, weil zwischendurch andere oder untergeordnete Aktionen ausgeführt werden (Abbildung 5 zeigt zur Erläuterung einen Ausschnitt aus einem Verlaufsprotokoll).

Nach welchen Kriterien werden Verhaltensepisoden gebildet? Barker und Wright nennen zum einen Merkmale, die einen Wechsel zwischen zwei Episoden ankündigen. Zu diesen sogenannten proximalen Episodenmerkmalen (Barker & Wright, 1971, S. 236) gehört der Wechsel des Verhaltensbereichs, z.B. von verbalem zu körperlichem Verhalten. Wichtiger als diese Kriterien, die zum Teil auch unterschiedliche Aktone indizieren können, sind jedoch Merkmale, die unabhängig von möglichen Varianten in der Ausprägung die Episode „an sich" charakterisieren. Es handelt sich dabei um die drei primären Episodenmerkmale Richtungskonstanz, normale Verhaltensperspektive und gleichmäßige Stärke. Die *Richtungskonstanz* weist darauf hin, daß jedes Verhalten in einer Episode auf ein bestimmtes Ziel hin ausgerichtet ist: Änderungen in der Richtung des Verhaltens deuten einen Wechsel der Episode an. Die *normale Verhaltensperspektive* betrifft den Umfang einer Episode: Eine Episode liegt innerhalb dessen, was die Person von ihrem eigenen Verhalten wahrnimmt. Hilfestellung für die Episodierung gibt dabei Frage: Welches ist die *längste* Aktion, die die Person nennen würde, wenn man sie fragen würde, was sie *jetzt gerade* tut (vgl. Barker & Wright 1971, S. 247)? Die *gleichmäßige Stärke* schließlich bezieht sich auf die annähernde Konstanz jeder Verhaltensepisode während ihres Verlaufs, beispielsweise ausgedrückt durch eine gleichmäßige Aufmerksamkeit der Person.

Sobald das Verlaufsprotokoll in eine Episodenstruktur gebracht worden ist, können verschiedenste Fragestellungen bearbeitet werden. Beispielsweise untersuchten Barker und Wright die Häufigkeit und die Formen sozialer Interaktion zwischen den Kindern und verschiedenen Bezugspersonen (Eltern, Lehrer, Freunde). Als Datenmaterial dienten die Tagesverlaufsprotokolle von acht Kindern im Alter von zwei bis zehn Jahren, die insgesamt 7751 Episoden umfaßten. Über die Hälfte aller Episoden waren sozialer Natur, d.h. neben dem Kind war mindestens eine weitere Person anwesend; in 80% dieser Situationen fanden Interaktionen statt. Dabei interagierten die Kinder häufiger mit Erwachsenen, besonders mit der Mutter, seltener mit Gleichaltrigen.

Going inside

7:32 The father called good-naturedly, "Come on, Chuck. Come along, boy."

Chuck jumped down easily and quickly.

He trotted a few steps ahead of his father.

Mr. Thurston caught up at the back door of the house. There he said briskly, as he opened the door, "Come on; let's get inside."

Chuck bounded into the house.

He walked quickly through the kitchen and dining room into the living room, where his mother sat resting on a couch.

Taking off wraps

He started to peel off his jacket.

Commenting on cold

At the same time, he remarked companionably to his mother, "It's cold outside. It's really cold out there." He said this in a very adult way.

Only smiling pleasantly, his mother seemed to take it that way.

7:33 Chuck pulled his jacket down from his shoulders; but it stayed on because the sleeves jammed against the bulky gloves he was wearing.

Getting mittens off

Chuck demanded of no one in particular, "Mittens off!"

His mother said nothing and made no move to help.

Chuck resolutely walked over to his mother.

Standing before her with his coat sagging, he soberly held out his hands.

The mother reached toward him.

Then, while he helped by pulling back a little, she tugged off his gloves.

Chuck wriggled on out of his coat, letting it drop where he stood.

Putting away wraps

His mother said firmly, "Chuck, put your hat and coat away."

He just stood there.

His mother repeated her command, this time very firmly.

Chuck asked, looking impish, "Shall I get the ruler, Mommie?"

> Earlier in the day Chuck had refused to put away his wraps, whereupon his mother had said threateningly, "Now, where is my ruler?" So, in asking now if he should get the ruler, Chuck evidently was just beating his mother to the draw.

Chuck did not press the question about the ruler. Before his mother could answer it, he picked up his hat and coat. Then he carried the wraps into the bedroom. He was smiling. He returned to the living room at once.

Abbildung 5: Ausschnitt aus dem Verlaufsprotokoll (mit Episodeneinteilung) von Dutton (Chuck) Thurston (Barker & Wright, 1971, S. 237)

Zur Klassifizierung der Interaktionen entwickelten Barker und Wright (1971, S. 331-339) ein Kategoriensystem, das verschiedene Handlungsmodalitäten beschreibt. Beispiele sind Kategorien wie „Dominanz", „Einwilligung" und „Aggression". Die Episoden der Verlaufsprotokolle wurden diesen Kategorien zugeordnet. Auf diese Weise konnten Aussagen über relative Auftretenshäufigkeiten von Verhaltensmustern in Abhängigkeit von den jeweiligen Interaktionspartnern gemacht werden. Beispielsweise war die Beziehung zwischen Kindern und Erwachsenen stärker von Komplementarität geprägt als bei Kindern untereinander. Dominanz im Sinne von Autorität der Erwachsenen ging häufig mit einwilligendem Verhalten der Kinder einher, während Kinder untereinander mehr Dominanzverhalten und - als Reaktion darauf - mehr Widerstand zeigten. Hilfsbereitschaft war nach Dominanz die zweithäufigste Kategorie bei den Erwachsenen, während die Kinder am häufigsten Bitten an die Erwachsenen richteten. Aggressionen traten sowohl auf seiten der Erwachsenen als auch auf seiten der Kinder relativ selten auf.

(5) Fazit: Sammeln, ja - aber alles auf einmal, nein!

Es dürfte deutlich geworden sein, welcher enorme Aufwand hinter dem Beobachtungsansatz von Barker und Wright steckt. Allein der Beschreibung von „One boy's day" haben die Autoren ein ganzes Buch gewidmet (Barker & Wright, 1961). Von einem ausgeglichenen Kosten-Nutzen-Verhältnis kann bei dieser Vorgehensweise in bezug auf die gewonnenen wissenschaftlichen Erkenntnisse kaum die Rede sein. Für praktische Anwendungszwecke ist das Verfahren jedenfalls viel zu unhandlich.

Der eingangs erwähnte Anspruch, die Verteilung der *relevanten* psychologischen Bedingungen zu identifizieren, unter denen Verhalten stattfindet, kann nicht dadurch eingelöst werden, daß zunächst einmal soviel wie irgend möglich gesammelt wird. Vielmehr scheint es angesichts der Datenmenge schwierig, Wesentliches von Unwesentlichem zu trennen. Damit soll nichts gegen das Ziel des Sammelns an sich gesagt werden, sondern lediglich gegen den Anspruch, das gesamte Verhalten erfassen zu wollen. So wie in der Auswertung von Verlaufsprotokollen immer nur bestimmte Verhaltensaspekte untersucht werden, sollte auch bereits bei ihrer Erstellung ein Zugang gewählt werden, der einzelne Aspekte berücksichtigt. Theoriefreie, „natürliche" Daten, wie sie das Verlaufsprotokoll liefern soll, und die Beschränkung der Verhaltensbeobachtung auf spezifische Bereiche schließen sich *bis hierher* nicht aus.

2.2 Mächtigkeit und Grenzen der Alltagssprache: Beschreibung und Deutung

Wir sollten nun aber die „Natürlichkeit" der Wahl der Alltagssprache als Beschreibungsmedium etwas kritischer unter die Lupe nehmen. Zunächst: Heißt „Wissenschaft betreiben" nicht immer (auch) „Einzelphänomene unter starker Abstraktion auf allgemeine Gesetze reduzieren"? Wie

kommt dann jemand dazu, mit einem derart komplexen System wie der Alltagssprache zu arbeiten? Zum zweiten: Heißt „Wissenschaft betreiben" nicht auch und vor allem, in der Genauigkeit und Präzision über die alltägliche Beschreibung, über den intuitiven Eindruck, über den „common sense" deutlich und wesentlich hinauszugehen? Wie ist das möglich, wenn in beiden Fällen dieselbe Sprache benutzt wird? Und schließlich: Welche Annahmen, welche Implikationen sind – vielleicht unbemerkt – in unserer Alltagssprache enthalten, und damit auch in den Beschreibungen, die wir mittels der Alltagssprache machen?

Beginnen wir mit dem letzten Punkt. Es erfordert offenbar eine ganze Wissenschaft, die Regeln zu identifizieren, die wir implizit beim Gebrauch unserer Sprache verwenden. Und damit sind nicht nur Untersuchungen über die Syntax (das formale „Gerippe"), sondern auch über die Semantik (die Bedeutungsstrukturen) angesprochen. Diese Untersuchungen, die im übrigen nicht nur Sprachwissenschaftler, sondern mit etwas anderer Zielsetzung auch Philosophen und (wir werden sofort sehen, warum) Psychologen durchführen, stoßen jedoch immer wieder an Grenzen. Mindestens die semantischen Strukturen sind offenbar nicht so ohne weiteres scharf und eindeutig zu identifizieren. Anders gesagt: Die Alltagssprache hat den Ruf notorischer Unschärfe. Beck (1987) beispielsweise kritisiert heftig, daß man sich auf sie im Zusammenhang wissenschaftlicher Beobachtung verlassen wolle: sie sei einfach zu unpräzise dazu. Man kann, so scheint es auf den ersten Blick, einfach nicht ausschließen, daß Menschen dieselben Wörter möglicherweise völlig unterschiedlich benutzen.

Eine Antwort auf dieses Problem ist sicher nicht einfach und in knapper Form zu geben. Ein wichtiger Punkt ist jedoch der folgende: Natürlich gibt es interindividuelle Unterschiede, aber sie können nicht wirklich dramatisch sein, weil es sonst keine Verständigung geben könnte (und weil nicht klar ist, wie eine solche Sprache, die in dieser Form sozusagen nur ich privat spreche, überhaupt vorstellbar, erlernbar und dann vermittelbar wäre[8]). Natürlich gibt es Mißverständnisse, aber sie fallen als Ausnahmen auf und sie lassen sich prinzipiell klären (den guten Willen der Beteiligten vorausgesetzt). Die Alltagssprache ist offenbar exakt genug, uns handlungsfähig zu erhalten, den Umgang miteinander sicherzustellen, Mißverständnisse hinreichend zu vermeiden und Informationen genau genug weiterzugeben. Wer dies bestreiten wollte, müßte konsequenterweise diese Sprache immer dann meiden, wenn er verstanden werden will - z.B. bei der Diskussion der Genauigkeit der Alltagssprache. Aber *ohne* die Möglichkeit der Verständigung könnte man überhaupt nicht argumentieren, also auch nicht für oder gegen die Möglichkeit des Argumentierens.

Zwar kann (und im Rahmen einer exakten Wissenschaft: soll) man Fachwörter oder sogar eine spezielle Terminologie einführen, um lange Erläuterungen und mögliche Mißverständnisse möglichst zu vermeiden. Der Haken liegt hier aber in dem unscheinbaren Wörtchen „einführen". Es ist eben unumgänglich, daß man sich zum Einführen einer neuen Sprache bereits einer

[8] Hierher gehört eigentlich eine Zusammenfassung über die entsprechende philosophische Debatte zur Möglichkeit und Denkbarkeit einer solchen „Privatsprache". Statt dessen hier nur der Hinweis: Wer sich für diese Problematik interessiert, sollte die klassische Auseinandersetzung bei Ludwig Wittgenstein (1984) nicht versäumen. Man lernt dabei ganz unversehens auch sehr viel über andere psychologische Fragen.

Sprache bedient, die der andere versteht. Die einzige Sprache, bei der man das letztlich voraussetzen kann (und daher muß), ist eben die Alltagssprache. Etwas technischer ausgedrückt: Die Alltagssprache ist immer die letzte Metasprache (d.h. die Sprache, in der wir uns über alle anderen Sprachen verständigen). Dieser Einwand trifft nun tatsächlich alle vorgeblich exakteren Verständigungsmöglichkeiten. Sie alle, jedes noch so komplexe Beobachtungssystem, jede noch so technische Sprache (auch mathematische Symbole), stehen letztlich auf den angeblich so schwachen Beinen der Alltagssprache. Würden sie das nicht, wären sie schlicht unverständlich: eine bedeutungslose Menge von Zeichen. Wir können der Unschärfe der Alltagssprache nicht dadurch entgehen, daß wir eine andere als die Alltagssprache verwenden, weil sie letztlich jeder neuen Sprache zugrunde liegt.

Dies wirft allerdings zwei schwierige Fragen auf:

1. Wie lernen wir denn unsere *erste* (Alltags-, Mutter-) Sprache?
2. Wie können wir sicher sein, daß andere Menschen, die diese Sprache ebenfalls zu sprechen scheinen, uns auch tatsächlich verstehen?

Eine ganz grobe Antwort darauf lautet: Die Praxis ist Korrekturinstanz und damit Kriterium. Bei Mißverständnissen interveniert die Umwelt – teils aktiv fördernd (Eltern, Lehrer), teils als passives Korrektiv: Wir scheitern einfach, wenn wir uns nicht verständlich machen können. Und überdies: Wir können schon deshalb kaum bezweifeln, daß wir uns verständlich machen können, weil wir unseren Zweifel dann ja ebenfalls nicht artikulieren könnten.

Es ist also festzuhalten, daß wir uns zwar einerseits über die korrekte Verwendung aller (semantischen) Regeln kaum jemals vergewissern können, andererseits aber gar nicht anders können, als unsere Begriffe korrekt zu verwenden: Ein großer Teil unserer individuellen Entwicklung besteht darin, zum „kompetenten Sprachbenutzer" zu werden. Wir wachsen sozusagen in ein vorgegebenes „Gebrauchsmuster" der Sprache hinein, welches uns die Verständigung ermöglicht. Natürlich gibt es Auseinandersetzungen, ob ein Begriff diese oder jene Bedeutungsnuance umfaßt; diese Auseinandersetzung ist wiederum aber nur vor dem Hintergrund eines Konsenses über die Bedeutung sehr vieler anderer Begriffe möglich! An diesen Konsens appelliert jeder, der Verbalsysteme in einer Beobachtungsstudie einsetzt – und so explizit ja auch Barker und Wright.

Allerdings ist es richtig und wichtig, darauf hinzuweisen, daß es in der Umgangssprache unscharfe Begriffe gibt, die für ihren Alltagszweck taugen mögen, für die angestrebte Exaktheit der wissenschaftlichen Theorie oder Hypothese aber nicht. Ein schönes Beispiel dafür sind Begriffe wie „selten", „häufig", „manchmal", „oft", „meist", „ab und zu" etc. Den Test kann jeder mit einer kleinen Gruppe Personen machen: Geben Sie einfach eine Reihe solcher Begriffe vor, und lassen Sie (z.B. in Prozentangaben) schätzen, „wieviel" damit gemeint ist („Wieviel Prozent ist 'häufig'?"). Sie werden mindestens für einige der Begriffe drastische Unterschiede zwischen den Personen finden. Die Lösung dieses Problems liegt jedoch auf der Hand: Verwende immer die exakteste Beschreibung, die möglich ist.

Mees und Selg (1977) weisen auf einen weiteren Nachteil der Beschreibung von Beobachtungen in der Alltagssprache hin: Sie suggeriert leicht eine Unvoreingenommenheit und Unbefangenheit („Theoriefreiheit"), die es so nicht gibt. Dieser Suggestion nicht zu erliegen, sind

alle Wissenschaftler aufgefordert (wir kommen auf das Problem der „theoriefreien" Beobachtung und Beschreibung im folgenden Kapitel 3 zurück).

Es stellt sich aber noch ein ganz anderes Problem, das - mutatis mutandis - auch alle anderen Anzeigesysteme (Zeichen-, Kategorien- und Ratingsysteme; vgl. Kap. 4 und 5) betrifft. Unsere Alltagssprache ist sehr „mächtig": Manche ihrer Begriffe besitzen die „mysteriöse" Eigenschaft, mehr zu besagen, als man wahrnehmen kann. Ich sehe ein paar Muskelbewegungen von Arm und Gesicht, sage aber: „Willi grüßt Otto"! Dürfen wir (oder müssen wir vielleicht sogar) solche Begriffe bei der Beschreibung von Beobachtungen verwenden?

2.3 Menschliche Handlung: Das Problem „mentaler" Begriffe

Dieser Punkt, der uns bereits bei Barker und Wright begegnet war, ist gut durch das Problem der „Handlungsbegriffe" zu erläutern (siehe zum folgenden auch Greve, 1994). Nehmen wir das Beispiel von oben: „Willi grüßt Otto". *Sehen* können wir nur die Muskelbewegungen von Arm und Gesicht, die Beschreibung „grüßt" enthält aber offensichtlich mehr. Wir unterstellen bei Willi eine bestimmte Absicht, und vor allem unterstellen wir diese *ihm* (und nicht seinem Arm) als einer Person, die sich der Handlung bewußt ist und sie freiwillig unternimmt: Er hätte auch anders handeln können. Zu den „Gebrauchsmustern" unserer Sprache gehört häufig also auch ein Schließen von rein physischen Ereignissen auf Psychisches, *Mentales.*[9] Auf der einen Seite haben wir reine beobachtungsbezogene Verhaltensbegriffe, auf der anderen reden wir von Handlungen.

Das Problem ist leider sogar noch ein wenig komplexer. Aus ein und derselben Bewegungsfolge kann ich in aller Regel mit gleicher Berechtigung auf mehrere verschiedene Handlungen schließen: War diese Handbewegung ein Gruß, das Verscheuchen einer Fliege, ein unwillkürlicher Muskelreflex oder etwas anderes? Und umgekehrt kann sich ein und dieselbe Handlung in ganz unterschiedlichen beobachtbaren Bewegungen äußern (erinnern wir uns an das Beispiel des Aufschlags). Und schließlich: Offenbar können wir in gewissem Sinne auch mehrere Handlungen zugleich ausführen. Das macht das Beispiel des rasenmähenden Herrn Schmitt deutlich (Abb. 6).

Was tut Herr Schmitt? Er mäht den Rasen, er geht dabei seiner Frau aus dem Wege, er ärgert dadurch seinen Nachbarn (Herrn Ifrabrumlitz), er trainiert auf diese Weise nebenbei seinen Körper und fördert seine Gesundheit, indem er seine Muskeln bewegt, er pflegt seinen Garten, dadurch daß er den Rasen schneidet, was letztlich auch seinem Ziel dient, den Wert seines Anwesens zu erhöhen, und bereitet schließlich (ohne daß er daran gedacht hat) die Klee-Ernte vor. Je nach Beschreibung bzw. Beschreibungsebene bedeutet die beobachtbare Bewegung etwas

[9] Was sich an dieser Stelle andeutet, ist nichts anderes als das altehrwürdige „Leib-Seele-Problem". Auch hier anstatt eines langen Exkurses nur ein kurzer Hinweis: Zur Einführung eignet sich die lesenswerte Arbeit von Bieri (1981; sollte sich jemand für unsere Gedanken hierzu interessieren, gibt es dazu eine kleine Arbeit [Greve & Wippermann, 1990], die man bei uns anfordern kann).

anderes, und diese Beschreibungen schließen sich auch nicht notwendigerweise gegenseitig aus.

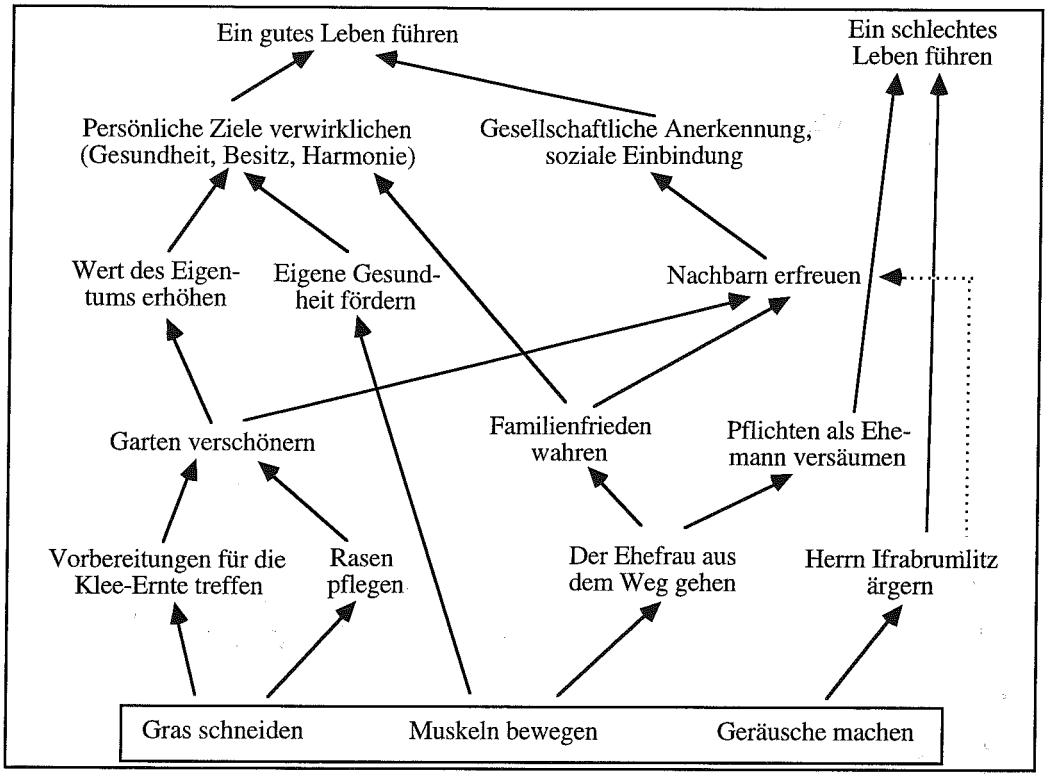


Abbildung 6: „Was tut Herr Schmitt hinter seinem Rasenmäher?" (Beispiel modifiziert nach Rommetveit, 1980, S. 118)

Wir stehen offenbar vor einem Dilemma. Einerseits scheint für die Beschreibung auf der Ebene der Handlungsbegriffe (oder allgemeiner: der Beschreibung unter Verwendung *mentaler* Begriffe) eine *Interpretationsleistung* des Beobachters gefordert zu sein, die sich nicht unter Rekurs auf das, was er beobachtet hat, absichern läßt. Andererseits scheint uns die Pointe dessen, was wir beobachten, zu entgehen, wenn wir auf diese Begriffe verzichten. Das psychologisch Interessante ist ja nicht, daß sich diese oder jene Muskelgruppen an Willis Arm bewegt haben, sondern daß er Otto gegrüßt hat (und nicht ihn verletzen wollte). Wir können also auf die Leistungsfähigkeit des menschlichen Beobachters (d.h. in gewissem Sinne eben: Interpreten) nicht verzichten. Wo immer möglich, sollten wir uns dabei vergewissern, daß der Handelnde selbst nicht doch auf Nachfragen eine andere Absicht enthüllt. Seine Ansicht dazu, was er gerade getan und beabsichtigt hat, ist natürlich von ganz besonderem Interesse. Wo diese Information nicht ohne weiteres zu bekommen oder nicht glaubhaft ist, müssen wir mindestens sicherstellen

bzw. überprüfen, daß die „externen" Interpreten übereinstimmend geurteilt haben. Wie und unter welchen Voraussetzungen wir das prüfen können, werden wir in späteren Kapiteln sehen (4 und 5).

Es ist aber wichtig, sich klarzumachen, daß man dieser Problematik nicht durch die Wahl eines Beobachtungssystems mit höherem Reduktionsniveau (Zeichen-, Kategorien- oder Ratingsysteme) entgehen kann. Auch für diese müssen die Zeichen, Kategorien oder Ratingdimensionen eindeutig gekennzeichnet und beschrieben werden. Auch mit ihnen sollen häufig psychologisch relevante Verhaltensweisen (Handlungen) erfaßt werden. Der Versuch, eine Psychologie völlig ohne Rekurs auf mentale Phänomene (wie Absichten, Überzeugungen oder Gefühle) zu realisieren, kann als gescheitert betrachtet werden (wir haben darauf schon im Abschnitt 1.1 hingewiesen). Barker und Wright waren ihrer Zeit mit ihrer handlungstheoretischen Sichtweise in gewissem Sinne voraus. Selbstverständlich sind aber viele neuere Beobachtungsstudien diesem Ansatz verpflichtet, und eben oft auch dann, wenn sie nicht in natürlicher Sprache protokollieren, sondern ein Zeichen- oder Kategoriensystem einsetzen (vgl. z.B. Kalbermatten & von Cranach, 1981; Humpert & Dann, 1988; Mees, 1988).

Bevor wir nun aber Beobachtungssysteme mit höherem Reduktionsanspruch (Zeichen-, Kategorien- und Ratingsysteme) ausführlicher darstellen, ist es wichtig, die Frage zu diskutieren, wie zuverlässig und genau menschliche Beobachtung überhaupt sein kann. Es wird sich zeigen, daß wir mit verschiedenen Fehlern rechnen müssen. Erst wenn wir sie etwas genauer kennengelernt haben, können wir die Beobachtungssysteme, die in wissenschaftlich-psychologischer Beobachtung in aller Regel eingesetzt werden, wirklich einordnen, denn sie stellen auch eine Antwort auf diese Probleme dar.

Literaturempfehlungen

Eine Einführung in Verbalsysteme (und eine zusammenfassende Darstellung der Studien von Barker und Wright) findet sich in Faßnacht (1995), die klassische Arbeit von Barker und Wright (1971) ist sicher lesenswert. Die Schwierigkeiten sprachlicher Beschreibungen werden besonders kritisch – vielleicht zu kritisch – bei Beck (1987) hervorgehoben. Eine gute Einführung in die prinzipiellen Probleme einer handlungstheoretischen Perspektive bietet Moya (1990; vgl. auch Greve, 1994); aus Sicht der Beobachtungsmethoden diskutiert werden sie etwa bei Mees (1988) oder bei Kalbermatten und von Cranach (1981).

Kapitel 3

Wie zuverlässig und genau kann Beobachtung sein?

Wir werden in diesem Kapitel sehen, daß bei Beobachtung neben den impliziten Bedeutungsstrukturen der Alltagssprache noch weitere Einflüsse eine Rolle spielen. Das Beobachtungsprotokoll, das wir über den Beobachtungsgegenstand schließlich anfertigen, wird nicht nur durch das, was beobachtet wird, und die Sprache, in dem die Beobachtung wiedergegeben wird, bestimmt. Es kommen vielmehr zahlreiche weitere Einfluß- und damit potentielle Fehlerquellen hinzu. Diese müssen wir mindestens kennen, besser kontrollieren und am besten eliminieren, damit das Beobachtungsprotokoll dem Beobachtungsgegenstand möglichst ähnlich wird.

Bevor wir jedoch den praktisch wichtigeren Schritt tun - d.h. die konkreten Fehlerquellen diskutieren - müssen wir uns noch über einen grundsätzlichen Punkt klar werden. Selbst wenn wir nämlich die Implikationen ignorieren, die in der Beschreibung des Beobachteten – genauer gesagt: in der hierfür gewählten Sprache – liegen, ist Beobachtung - Wahrnehmung überhaupt - nicht nur passive Rezeption. Anders ausgedrückt: „Wahrnehmung" beschreibt nicht nur das Auftreffen von Lichtkonfigurationen auf der Netzhaut, sondern auch ihre weitere Be- und Verarbeitung durch den Wahrnehmenden bzw. durch den „Wahrnehmungsapparat".

3.1 „Voraussetzungsfreie" Wahrnehmung: Prinzipielle Fragen

Wir finden in zahlreichen Experimenten Belege dafür, daß die Repräsentationen unserer Umwelt in unserem Gehirn nicht unverbunden einfach als eine „Sammlung" von Tatsachen „abgelegt" werden, sondern miteinander in verschiedenster Art und Weise verbunden werden und so ein vielfach vernetztes System von Abhängigkeiten bilden. Dieses System wirkt sich in komplexer Weise auf unsere gesamten Wahrnehmungs- und Denkprozesse aus. Nehmen wir als Beispiel eine sogenannte Kippfigur: In Abbildung 7 kann man einen Hasen oder auch eine Ente erkennen, aber nicht beide gleichzeitig (vgl. auch Martin & Wawrinowski, 1991, S. 15ff.).

Einerseits ist natürlich klar, daß die objektiven Gegebenheiten (also in diesem Fall die Striche auf dem Papier) unsere Wahrnehmung steuern: Man sieht nichts anderes als einen Hasen oder eine Ente. Andererseits sehen wir hier aber deutlich, wie diese objektiven Gegebenheiten mit Erwartungsmustern, unseren Repräsentationen bestimmter Konzepte quasi „abgeglichen" werden. Auch die zahlreichen Phänomene optischer Täuschungen verdeutlichen diesen Punkt sehr schön; auch dafür ein Beispiel (Abb. 8).

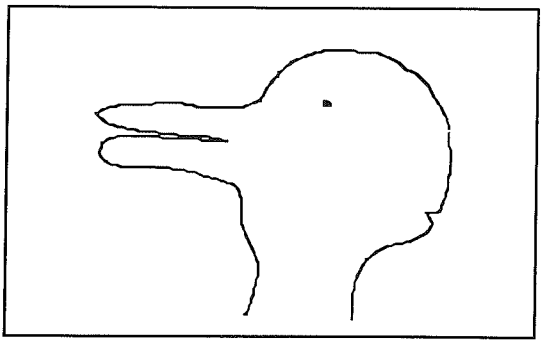


Abbildung 7: Ente oder Hase?

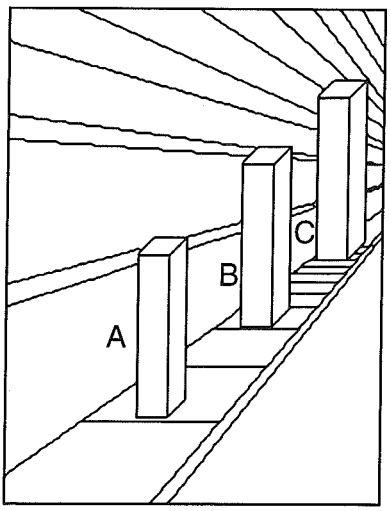


Abbildung 8: Größentäuschung durch Hintergrundgestaltung

Auch in diesem Fall sehen wir sozusagen „eigentlich" drei gleich große Quader. Die „Hintergrundinformation" geht aber in die Wahrnehmung dieser Objekte derart mit ein, daß sie in charakteristischer Weise „gedeutet" oder „interpretiert" werden. Wir nehmen die Quader A, B und C als unterschiedlich groß wahr; daran ändert nicht einmal die Information etwas, daß hier wirklich nur eine optische Täuschung vorliegt.

Die kognitive Psychologie belegt diese Prozesse mit dem Begriffspaar „bottom up" versus „top down processing" bzw. „datengetriebene" versus „konzeptgeleitete Verarbeitung" (Lindsay & Norman, 1977/1981). Unser kognitiver „Apparat" hat offensichtlich die bemer-

kenswerte Fähigkeit, die Konzepte, die er gespeichert hat, mit den Reizen, die seine Sinnesorgane aufnehmen, zu verbinden (Abb. 9 gibt für diese Fähigkeit ein weiteres Beispiel).

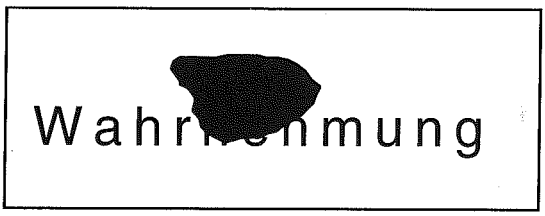


Abbildung 9: Musterergänzung

Auch in Fällen wie diesem sind wir typischerweise ohne Schwierigkeiten in der Lage, das Wort „Wahrnehmung" sozusagen direkt zu „sehen", d.h. die fehlenden Teile „automatisch" zu ergänzen und die irrelevanten Aspekte (Klecks) eben als irrelevant einzuordnen. Da wir im Falle derart komplexer Wahrnehmungsaufgaben bis auf weiteres nicht auf gleichwertige Maschinen und Apparate zurückgreifen können, müssen wir vorläufig uns selbst als Meßinstrumente einsetzen. Den Vorteilen der ungeheuren Leistungsfähigkeit (und halbwegs guten Verfügbarkeit) stehen dabei natürlich verschiedene Nachteile gegenüber. Der erste ergibt sich aus dem soeben noch als Vorteil angepriesenen Phänomen: Wir arbeiten und funktionieren nach Regeln, die uns (a) z.T. selbst nur ausschnitthaft bekannt sind, und die (b) für uns - sogar dann, wenn sie uns bekannt sind - oft nicht in der wünschenswerten Weise kontrollierbar sind (wir hatten darauf kurz bereits in Abschnitt 1.4.1 hingewiesen). Zur Erläuterung zwei „Klassiker" der Psychologie:

Beispiel 1: Bartlett (1932; zitiert nach Glass, Holyoak & Santa, 1979) legte seinen Versuchspersonen eine alte indianische Sage vor: „The war of the ghosts" war eine kurze Geschichte, die ein wenig fremd für unsere Ohren klingt; z.B. sind einige Übergänge in der Geschichte in ihrer Logik unklar. Bartlett bat nun seine Versuchspersonen, die Geschichte später nachzuerzählen. Natürlich fehlten - wie zu erwarten - sehr viele Einzelheiten; interessanter waren aber die Hinzufügungen: Die Versuchsteilnehmer erzählten eine logisch abgerundete Geschichte.

Beispiel 2: Loftus und Palmer (1974) baten Versuchspersonen, sich einen Film über einen Verkehrsunfall anzuschauen. Später wurde ein Teil der Personen mit der Frage: „Wie schnell fuhren die Autos, als sie aneinanderstießen?", ein anderer mit der Frage: „Wie schnell fuhren die Autos, als sie ineinander krachten?" (sinngemäße Übersetzung G/W) um eine Einschätzung der Geschwindigkeiten der beteiligten Wagen gebeten. Das Ergebnis war kurz gesagt, daß die mittlere Schätzung der Geschwindigkeit in der ersten Gruppe geringer war als in der zweiten. Darüber hinaus wurde den Versuchspersonen nach einer Woche ein Fragebogen zu dem Film vorgelegt, der unter anderem die Frage enthielt, ob splitterndes Glas zu sehen gewesen sei. Die zweite Gruppe (die sich also lediglich durch die etwas andere Wortwahl in der eine Woche zuvor gestellten Frage von der ersten unterschied) berichtete signifikant häufiger von splitterndem Glas, obwohl dieses gar nicht im Film zu sehen war!

Bei genauerem Hinsehen zeigt sich, daß hier nicht nur Probleme unserer aktuellen, gewissermaßen je persönlichen Unvollkommenheit, sondern grundlegendere Schwierigkeiten lauern. Ist so etwas wie „reine" („theoriefreie") Wahrnehmung überhaupt eine sinnvolle Vorstellung, ein kohärenter Begriff? Der sogenannte „hermeneutische Zirkel", die wechselseitige Abhängigkeit von „Theorie" (d.h. Voreinstellungen im wörtlichen und metaphorischen Sinne) und „Erfahrung" tritt keineswegs nur in den sogenannten Geisteswissenschaften, sondern in völlig analoger Weise auch in den Naturwissenschaften auf (vgl. hierzu z.B. Stegmüller, 1979). Jede Wahrnehmung setzt Schemata voraus, ist immer Kategorisierung und Konfigurierung, auch Filterung, eben ein gewissermaßen aktiver Vorgang: Unzusammenhängendes wird zusammengesetzt, d.h. als Zusammenhang gedeutet. Eine „Blume" wird als Blume wahrgenommen, sie ist dabei aber so wenig eine Blume „an sich", wie sie eine Farbpalette oder eine Kohlenwasserstoffkolonie (wo wäre übrigens bei *dieser* Sicht die Grenze zwischen „Blume" und Umwelt genau?) oder was immer ist. Aber durch die Wahrnehmung sind nicht nur sozusagen die Grenzen aktiv gezogen, auch Sinnzusammenhänge werden aktiv gebildet. Dieses macht Wahrnehmung im vollen Wortsinn erst aus und gilt für alles, was wir „beobachten": menschliche Handlungen und Verhaltensweisen, Ereignisse, Dinge usw. Dies alles wird von uns im Licht von Theorien als die jeweilige Beobachtung konstruiert: „Beobachtungssätze sind unvermeidlich theorieimprägniert oder 'theoriegeladen'" (Lenk, 1987, S. 26).[10] Von allen diesen Objekten der Erkenntnis sind - logisch gesehen - grundsätzlich stets mehrere Beschreibungen möglich, und das heißt: Selektionen und Deutungen im Sinne dieser „top-down"-Prozesse. Wenn wir das hier so formulieren, muß aber klar sein, daß wir hierbei nicht (kontrolliert und absichtlich) handeln: wir können uns - im Gegensatz etwa zu einer Interpretation eines Musikstückes - nicht einfach für oder gegen eine „Interpretation" entscheiden. Wir *sehen* einfach eine Blume. Die These, Wahrnehmung sei ein „aktiver" Zugriff auf „die Welt", darf nicht in dem Sinne mißverstanden werden, daß wir als Wahrnehmende sozusagen willkürlich modellieren. Gemeint ist nur, daß Wahrnehmung ein Prozeß ist, an dem Wahrnehmender und Wahrgenommenes interaktiv beteiligt sind, genauer gesagt: den beide zusammen erst *konstituieren.*[11] „Sehen" ist nicht einfach der physiologisch beschreibbare Prozeß des Auftreffens von Reizen auf der Netzhaut. „People, not their eyes, see. Cameras, and eyeballs, are blind" (Hanson, 1958/1972, S. 6).

---

10 Man kann sich allerdings fragen, ob hier der Begriff der „Theorie" nicht seinerseits überdehnt wird (Faßnacht, 1995, S. 43ff.).

11 Für philosophisch bzw. wissenschaftstheoretisch Interessierte: Die „Kritik der reinen Vernunft" von Immanuel Kant (1968; orig. 1781) ist hier eine zwar schwierige, aber für dieses (und nicht nur dieses) Problem einschlägige Lektüre. Zeitgenössischer Klassiker ist das Buch von Hanson (1958/1972); eine ausführliche Diskussion - auch mit Bezug auf empirisch-psychologische Befunde - findet sich z.B. bei Bohnen (1972) oder Kriz, Lück und Heidbrink (1990).

3.2 Beobachtungsfehler und Fehlerquellen: Konkrete Effekte

Stellen wir uns eine Szene am Strand vor: blaue See, blauer Himmel, weiße Schäfchenwölkchen (nur ein paar!), Kinder bauen eine Burg, bunte Badeanzüge, ein roter Ball, Sonnenreflexe auf dem Wasser. Ein Photo! Nur so kann dieser schöne Augenblick unverfälscht festgehalten werden. Ein Photo: in der Tat das anscheinend ideale Medium der Speicherung. Aber leider wären keine zwei Photos dieser Szene identisch: Verschiedene Ausrüstungen (Kameras, Linsen, Filter oder Filme), verschiedene Standorte und Aufnahmewinkel, eine Wolke, die den Glitzereffekt der Sonne auf den Wellen einen Moment unterbricht oder auch nur die Lichtverhältnisse ein wenig verändert, ein Wackeln des Photographen beeinflussen die Aufnahme selbst. Hinzu kommen Wiedergabeprobleme: kleine Unterschiede beim Entwickeln des Films oder beim Herstellen des Abzugs beeinflussen das Bild, das man schließlich vor sich liegen hat. Bei all dem ist noch nicht einmal berücksichtigt, daß das Photo natürlich doch nur einen Ausschnitt des gesamten Strandpanoramas liefert, ganz abgesehen davon, daß natürlich weder das Rauschen der Wellen noch das Lachen der Kinder oder das Kreischen der Möwen und vor allem nicht das rundum zufriedene Gefühl des Beobachters auf das Photo gebannt werden können. Das ideale Speichermedium?[12]

Das Ergebnis einer Beobachtung (das „Protokoll") wird nicht nur durch das bestimmt, was in ihm protokolliert werden soll. Daran ist wohl nicht zu rütteln. Vielmehr wird sein tatsächlicher Inhalt zusätzlich durch verschiedene Einfluß- und Fehlerquellen bestimmt. Kent und Foster (1977) weisen hier nachdrücklich darauf hin, daß wir über eine Vielzahl von Faktoren, die beim Zustandekommen des Ergebnisses einer Beobachtung eine Rolle spielen, derzeit noch gar nichts wissen (geschweige denn von Art, Umfang und Bedeutung ihres Einflusses). Obwohl diese Einschätzung nun bald anderthalb Jahrzehnte alt ist, wird man ihr im wesentlichen noch beipflichten müssen. Bei Hasemann (1964) findet sich der Hinweis, daß erste Forschungen über diese Fehler aus der Kriminalistik stammen: Es geht dort natürlich immer wieder um die Zuverlässigkeit und Glaubwürdigkeit von Zeugen. Die Untersuchung von Loftus und Palmer (1974; s.o. Abschnitt 3.1) gehört in diesen Bereich. Wildman und Erickson (1977) behaupten dagegen, das Interesse an der Zuverlässigkeit von Beobachtungen sei in den 20er und 30er Jahren im Zusammenhang mit Forschungen an Kindern erwacht. Wie auch immer, die Zahl empirischer Studien zu verschiedenen Beobachtungsfehlern und -verzerrungen ist mittlerweile recht umfangreich (Hinweise auf einige Überblicksartikel finden sich am Ende dieses Kapitels).

Die Probleme dürfen in der Tat nicht unterschätzt werden. Etwa Frenz und Frey (1981) zeichnen folgendes gern zitierte Bild: „Beobachter, so lehrt die Literatur, verkürzen, vereinfachen, ignorieren, akzentuieren, kontaminieren, beschönigen, verzerren das Verhalten ... in Abhängigkeit von ihren jeweiligen Einstellungen, von ihrer jeweiligen Gefühlslage, von ihren

12 Frei nach Kent und Foster (1977).

jeweiligen Erwartungen ... Hinzu kommt, daß Beobachter ihre Beurteilungsmaßstäbe im Verlaufe der Beobachtung - als Folge der Beobachtung - verändern ... Manche Autoren sind zudem überzeugt, daß allein die bloße Anwesenheit des Beobachters das zu beobachtende Geschehen wesentlich modifiziert" (S. 73f.). Ähnlich vernichtend urteilt Beck (1987) mit Bezug auf sein (durchaus typisches) Beispiel der Unterrichtsforschung: „Wer ... an Erkenntnissen interessiert ist, die zuverlässiger sind als die alltagspraktische Unterrichtserfahrung, die ja auf ... unkontrollierten Sachverhaltswahrnehmungen beruht [gemeint ist hier die interne Verarbeitung einer hoch komplexen Situation durch den menschlichen „Datenverarbeitungsapparat" Gehirn, G/W], der muß ... auf den Einsatz des Menschen als 'Meßgerät' verzichten" (S. 40).

Eine erste wichtige Frage bei diesem vernichtenden Urteil über die Zuverlässigkeit von Beobachterdaten muß natürlich lauten, ob die Kritik wirklich in allen Punkten überhaupt ihre sachliche Berechtigung hat. Wir werden später sehen, daß hier empirische Befunde in einigen Punkten zu Revisionen bzw. zu Differenzierungen auffordern. Frenz und Frey weisen übrigens in ihrer Arbeit selbst darauf hin, daß die genannten Probleme „in der Praxis" vermutlich weniger gravierend seien. Man könne die Randbedingungen der Untersuchung soweit verbessern, „daß man statt des idealen Beobachters auch den realen Beobachter verwenden" könne (S. 76). Relevanter als diese Fehlerquellen sei die *Planung* der Untersuchung (siehe Abschnitt 3.3).

Die abfällige Kennzeichnung des „letztlich undurchschaubaren, aber irgendwie funktionierenden 'Datenverarbeitungsapparates Mensch' " (Beck, 1987, S. 60) ist aber in mancher Hinsicht auch grundsätzlich unangebracht. Man könnte zunächst behaupten, eine bessere Möglichkeit gebe es nicht, denn auch die Ergebnisse präziserer Instrumente müssen für uns *erkennbar* sein (über diesen Punkt haben wir kurz bereits in Abschnitt 1.4 gesprochen). Dieser Einwand scheint auf den ersten Blick etwas dürftig, denn ein Mikroskop stellt doch wohl tatsächlich eine Verbesserung gegenüber dem unbewaffneten Hinsehen dar. Andererseits: Welchen Grund haben wir denn für die Annahme, daß die kleinen Partikel, die wir mit seiner Hilfe sehen, „wirklich" so aussehen, wie sie durch das Mikroskop aussehen? Wir haben Theorien über die Funktion der beim Mikroskop verwendeten Elemente (z.B. Linsen). Diese Theorien mögen sich in anderen Kontexten bewährt haben. Wie wir aber bereits angedeutet haben (Abschnitt 1.4, Exkurs), ist damit die Richtigkeit dieser Theorien nicht bewiesen, zumal nicht ihre Gültigkeit in anderen als den getesteten Bereichen. Außerdem war in diesen Testsituationen eben die unmittelbare Einsicht Kriterium für das Bestehen des Testes, d.h. in aller Regel: das unmittelbare Hinsehen (Beobachten). Dieser Prozeß kann über verschiedene Zwischenschritte gelaufen sein: Instrumente, die durch Instrumente geprüft wurden, die durch Instrumente ... usw. Das macht deswegen keinen grundsätzlichen Unterschied, weil das „erste" dieser Instrumente eben doch durch unmittelbare Beobachtung geprüft werden mußte. Aber auch wenn wir dem Mikroskop unmittelbar etwas unterlegen, von dem wir wissen, wie es aussieht (um zu prüfen, ob das Mikroskop es unverzerrt genau so wiedergibt), ist uns nicht geholfen. Denn woher wissen wir, wie dieses Objekt „wirklich" aussieht? Wiederum durch die unmittelbare Anschauung. Wir sind letztlich offenbar immer dazu gezwungen, unserer „normalen" Beobachtung oder Wahrneh-

mung zu trauen. Und normalerweise dürfen wir das auch. Menschen sind im Alltag ganz offenbar hinreichend gute Beobachter: Sie können das, worauf es ankommt, mit hinreichender Genauigkeit, Zuverlässigkeit und Gültigkeit erfassen und anderen vermitteln (bereits Barker und Wright hatten das ja zur Grundlage ihrer Untersuchungen gemacht; vgl. z.B. 1971, S. 206).[13] Zwar unterlaufen uns mitunter Fehler, aber die Tatsache, daß sie uns auffallen, sagt etwas über ihre Häufigkeit, und die Tatsache, daß wir sie überhaupt *als Fehler* registrieren können, bedeutet, daß wir offenbar eine Vorstellung davon haben, wie es „richtig" gewesen wäre.

Das heißt natürlich nicht, daß wir die Fehlermöglichkeiten bei einer konkreten Beobachtung ignorieren dürfen. Im Gegenteil: Es ermutigt uns, nach ihnen zu suchen, sie zu untersuchen und möglichst zu vermeiden, jedenfalls zu kontrollieren. Mit diesen oder ganz ähnlichen Fehler- und Einflußquellen, darauf sollte man vielleicht ausdrücklich hinweisen, müssen sich im übrigen auch andere Methoden der Datenerhebung (Meßinstrumente, Fragebögen etc.) herumschlagen. Auch sie selegieren und modifizieren die „Wirklichkeit" auf verschiedenste Weise. Es wäre also ganz falsch, aus der ausführlichen Diskussion von Problemen psychologischer Beobachtung den Schluß zu ziehen, wir wollten gegen diese Methode der Datenerhebung plädieren. Wir meinen vielmehr, daß effektive Maßnahmen mit dem Ziel einer hohen Qualität von Beobachtungsstudien eine möglichst klare Diagnose voraussetzen. Diese Diagnose ist das Anliegen dieses Kapitels (nur sicherheitshalber: Ein Vollständigkeitsanspruch verbietet sich dabei natürlich von selbst).

3.2.1 Gütekriterien: Reliabilität, Validität und Generalisierbarkeit

Bevor wir jedoch die wichtigsten Fehler- und Einflußquellen ordnen und ausführlicher darstellen, müssen einige grundlegende Konzepte und Begriffe wenigstens kurz eingeführt werden. Die folgenden Abschnitte behandeln daher zunächst die Konzepte der Reliabilität und der Genauigkeit, sehr knapp das Konzept der Validität und schließlich einige Fragen zur Generalisierbarkeit von Beobachtungsdaten. Diese Konzepte sind für die empirische Psychologie von grundlegender Bedeutung.

(1) Reliabilität und Genauigkeit

Am Anfang steht traditionell die formale Definition: *Reliabilität* (Zuverlässigkeit) meint das *Verhältnis von wahrer Varianz einer Variablen zu ihrer Gesamtvarianz.*[14] Da sich die Gesamtvari-

---

13 Block (1961) greift dies für einen recht interessanten Einwand gegen den üblichen Einsatz menschlicher Beobachter auf, indem er gewissermaßen umgekehrt argumentiert: *Wenn* menschliche Beobachter in diesem Sinne sehr zuverlässige Instrumente sind, dann sind die üblichen simplen Verhaltensbeobachtungen, schlichten Häufigkeitszählungen, das, wie Bischof (1989, S. 203) es formuliert hatte, „kuhäugige Anfertigen von Strichlisten" und dergleichen eine sehr schlechte Abbildung ihres Wissens, eine sehr schlechte Ausnutzung ihrer Fähigkeiten.

14 'Varianz' ist ein Kennwert, der die Variabilität einer Verteilung (z.B. von Meßwerten) abbildet (zur formalen Definition siehe Bortz, 1993, S. 41).

anz aus wahrer Varianz und Fehlervarianz zusammensetzt, meint Feger (1983) sinngemäß dasselbe, wenn er formuliert: „Reliabilität kann ... allgemein als das Verhältnis von systematischer, d.h. theoretisch erklärter Varianz zu Fehlervarianz, d.h. durch die jeweils betrachteten Bedingungen nicht erklärte Varianz konzipiert werden" (Feger, 1983, S. 22f.). Soweit die technische Bestimmung des Begriffs. Was soll das aber nun bedeuten? Jede Messung ist fehlerbelastet: Störende Einflüsse der Umwelt, zu grobe Instrumente, Fehler im konkreten Instrument und so weiter tragen zu dem tatsächlichen Ergebnis der Messung bei. Uns interessiert aber in einem konkreten Fall eigentlich nur, wie lang beispielsweise dieses konkrete Werkstück (unter normalen Umständen) ist. Wir wollen nicht mit erfassen, daß Metermaße nicht immer genau gleich lang sind (manche sind verzogen, manche sind nicht korrekt hergestellt worden) oder daß dieses Werkstück ein wenig länger wird, wenn man es erhitzt. Eine Messung ist nun einfach umso reliabler, je geringer der Anteil solcher Fehler am Ergebnis ist. „Generell beschreibt also Reliabilität die *Reproduzierbarkeit* von Beobachtungen unter theoretisch für das Auftreten des Beobachteten *äquivalenten* Bedingungen bei *Unterschieden* in theoretisch *irrelevanten* Bedingungen" (Feger, 1983, S. 24; Hervorhebungen G/W). Welche Bedingungen für Beobachtungen könnten in diesem Sinne irrelevante Bedingungen sein? Nun, beispielsweise verschiedene Beobachter oder verschiedene Beobachtungszeitpunkte. Verschiedene Situationen bzw. Umstände können dagegen durchaus theoretisch relevant sein. Ein Beispiel zur Erläuterung: Das Verhalten von denselben Kindern in der Schulklasse und auf dem Fußballplatz wird sich selbstverständlich unterscheiden, ohne daß wir auf den Gedanken kämen, *deshalb* der Beobachtung dieser Kinder eine mangelnde Reliabilität zu unterstellen. Oder, um es in der üblichen Terminologie zu formulieren: Bei *verschiedenen* Beobachtungen in diesem Sinne kommen zusätzliche und unterschiedliche Varianzquellen ins Spiel (vgl. dazu auch Manns et al., 1987). Wenn dagegen das Verhalten derselben Kinder in derselben Klasse bei demselben Lehrer in demselben Fach von zwei verschiedenen Beobachtern völlig unterschiedlich wahrgenommen (bzw. protokolliert) wird, werden wir genau dies vermuten: Eine starke Fehlerbelastung - und das heißt: eine geringe Reliabilität - (mindestens einer) dieser Beobachtungen. Prinizipiell könnte man sogar denselben Beobachter dasselbe Verhalten mehrfach beobachten lassen, indem man einfach das Verhalten auf Video aufzeichnet und dieses Video mehrfach beobachten läßt. Dieses Verfahren ist natürlich am besten geeignet, die Zuverlässigkeit eines Beobachters abzuschätzen: Man vergleicht einfach seine Beobachtungen bei verschiedenen Sitzungen miteinander. Dieses Verfahren wird jedoch für die allermeisten praktischen Fälle zu aufwendig sein.

Im Grunde müßte man auch hier wieder eine Einschränkung machen: Durch dieses Verfahren wird genaugenommen nur die *Konsistenz* dieses Beobachters, nicht aber seine "Fehlerbelastetheit" geprüft: Er könnte ja immer denselben Fehler machen, etwa aufgrund individueller Kodierungsgewohnheiten (Manns et al., 1987). Wir müßten dieses Verfahren also bei mehreren Beobachtern anwenden und zusätzlich deren Ergebnisse untereinander vergleichen. Dazu gleich mehr.

Man hilft sich bei diesem Problem grundsätzlich dadurch, daß man eine Messung mehrfach unter theoretisch äquivalenten Bedingungen wiederholt. Wenn dabei die Unterschiede zwischen den verschiedenen Messungen groß sind, ist das Meßinstrument offenbar stark fehleranfällig, d.h. unreliabel. Diese Meßwiederholung kann man nun dadurch erhalten, daß man zu verschiedenen Zeitpunkten dasselbe Instrument einsetzt (Retest-Reliabilität).[15] Man kann aber auch verschiedene Instrumente *derselben Art* bei der gleichen Gelegenheit einsetzen. Dies ist der Gedanke, der zu der Berechnung von Übereinstimmungsmaßen zwischen verschiedenen Instrumenten (z.B. Beobachtern) bei einer Meßgelegenheit führt: Man schätzt die Reliabilität dieser Instrumente, indem man sie miteinander vergleicht (wie das im einzelnen vor sich geht, werden wir für die verschiedenen Beobachtungssysteme in den entsprechenden Kapiteln erklären). Je ähnlicher ihre Ergebnisse, d.h. je größer die Übereinstimmung zwischen verschiedenen Beobachtern – so die Annahme – desto reliabler sind die einzelnen Instrumente (d.h. hier: die Beobachter). Maße für die Beobachterübereinstimmung gibt es mittlerweile in reicher Zahl. Cone und Foster (1982) zitieren Arbeiten dazu, in denen von über 20 Maßen gesprochen wird. Andererseits hat etwa Kelly (1977) bei einem Überblick über die in einer einschlägigen Zeitschrift berichteten Beobachtungsstudien gefunden, daß häufig ganz einfache Maße angewendet werden („%-Übereinstimmung"). Wir werden diese Problematik ausführlich erläutern und die wichtigsten Maße vorstellen (vgl. dazu Abschnitt 4.5, 5.5 und den Anhang).

Übrigens könnte man so auch die *Objektivität* der Beobachtung bestimmen, wenn man unter Objektivität - wie dies häufig der Fall ist - einfach die intersubjektive Übereinstimmung versteht. Meint man dagegen mit „Objektivität" die Unabhängigkeit des Ergebnisses von subjektiven Verzerrungen, ist sie auch bei hoher intersubjektiver Übereinstimmung nicht garantiert, weil alle Instrumente denselben Fehler machen können (vgl. dazu Abschnitt 3.2.3).

Die Einschätzung der *Genauigkeit* einer Messung bedeutet den Vergleich dieser Messung mit einem *Standard*. In einigen (eher seltenen) Fällen könnte dieser Standard ein mechanischer Apparat sein. Foster und Cone (1980) erläutern dies am Beispiel des „out of seat-behavior", des Aufstehens vom Stuhl, das durch eine druckempfindliche Apparatur gemessen wird.

An diesem Beispiel wird die zugrundeliegende Problematik deutlicher, als Foster und Cone dies bewußt zu sein scheint. Wie wird denn beurteilt, ob der mechanische „out of seat-recorder" auch ordnungsgemäß funktioniert? Ganz einfach, möchte man sagen: Wir setzen jemanden auf den Stuhl und schauen, ob der Automat das anzeigt. Was bedeutet das aber? Es bedeutet, daß die Genauigkeit des Apparates ihrerseits durch Beobachtung geprüft wird (die Beobachtung nämlich, daß jetzt jemand auf dem Stuhl sitzt, der Apparat das also anzeigen müßte), bevor dieser Apparat seinerseits die Genauigkeit von derartigen Beobachtungen prüfen kann. Merke: Es gibt keine unabhängige Evidenz.

---

15 Eine Schwierigkeit ergibt sich dabei dann, wenn man auch für das zu messende Phänomen (aus inhaltlichen Gründen) eine Veränderung erwartet. Bei Längenmessung ist das z.B. der Fall, wenn wir das Wachstum einer Pflanze oder eines Menschen messen wollen. Hier werden sich also auch die wahren Werte verändern. In der Psychologie gilt dies häufig etwa bei entwicklungspsychologischen Fragestellungen.

Die Genauigkeit eines Beobachters wird üblicherweise anhand eines vorbereiteten (konstruierten) Videobandes geschätzt, beispielsweise der Aufnahme eines nach Drehbuch ablaufenden Dialoges. Vom Beobachter des Videos muß dann zum Beispiel die Anzahl bestimmter Wörter beobachtet werden. Da anhand des Drehbuches die wahre Anzahl feststeht, liegt hier ein eindeutiger Standard vor. In vielen praktischen Fällen wird das zu aufwendig sein (und ja überdies auch keine Gewißheit für die Genauigkeit der in diesem Sinne „nicht standardisierten" Beobachtungen hergeben, die ja die eigentlich interessanten sind). Man behilft sich hier oft mit einem „Eich-Beobachter" (oder „Kalibrierungsbeobachter"), d.h. einem besonders erfahrenen oder geschulten Beobachter, dessen Beobachtungen als Standard fungieren (vgl. hierzu z.B. Foster & Cone, 1980; Mees, 1977a). Dieses Vorgehen muß sich dann natürlich die Frage gefallen lassen, warum man denn nicht gleich den Eichbeobachter nimmt, da ja mit anderen Beobachtern höchstens ebenso genaue Ergebnisse wie mit ihm erzielt werden können (Foster & Cone, 1980; Mees, 1977a); dieser Einwand trifft natürlich erst recht zu, wenn ein mechanischer Standard zur Verfügung steht. Ein wichtiger Grund hierfür ist die Sparsamkeit: Man kann mit nur einem Eichbeobachter zahlreiche gleichzeitig ablaufende Beobachtungen stichprobenartig (und damit in aller Regel hinreichend, freilich nicht vollständig) auf Genauigkeit prüfen, indem dieser Beobachter zwischen den Beobachtergruppen wechselt.

(2) Validität

Block (1961) hat die Frage der Reliabilität von Beobachtern ausführlich diskutiert. Ein wichtiges Argument *für* den Einsatz menschlicher Beobachter ist für ihn neben der sachlichen Feststellung, daß in der Praxis oft nichts besseres, ja nicht selten überhaupt keine Alternative verfügbar sei, vor allem die Behauptung, daß menschliche Beobachtungen häufig eine hohe Validität hätten, jedenfalls meist „näher" an den zu untersuchenden Phänomenen seien als „objektivere" Verfahren. Zwei Fragen drängen sich auf: (1) Was ist mit Validität überhaupt gemeint, und (2) was hat sie mit der Reliabilität zu tun?

Zur ersten Frage zuerst nur eine kurze Antwort: Mit der Frage nach der *Validität* (Gültigkeit) einer Messung ist die Frage gemeint, ob und inwieweit tatsächlich das gemessen wurde, was gemessen werden sollte. Trifft das eingesetzte Meßinstrument überhaupt den Punkt? Mißt z.B. eine mündliche Prüfung nicht viel mehr die soziale Sicherheit und Sprachgewandtheit des Kandidaten als sein fachliches Wissen oder sogar vielleicht einfach nur seine Fähigkeit, Streß auszuhalten?

Um etwas genauer zu verstehen, wie wir feststellen können, wann eine Messung valide ist, werfen wir zunächst einen Blick auf die zweite Frage. Dazu läßt sich folgendes sagen: Eine hohe Reliabilität ist notwendige Bedingung für eine hohe Validität (d.h.: „ohne hohe Reliabilität gibt es keine hohe Validität"). Ein unzuverlässiges und ungenaues Meßinstrument mißt sozusagen gar nichts richtig, also auch das nicht, was es messen soll. Daraus folgt umgekehrt: Eine hohe Validität ist hinreichende Bedingung für eine hohe Reliabilität (Also: „immer, wenn hohe Validität tatsächlich vorliegt, dann muß auch eine hohe Reliabilität vorliegen"). Das nützt uns

natürlich solange nichts, wie ungeklärt ist, woran wir denn nun *erkennen*, ob eine Messung valide ist. Und dies erkennen wir jedenfalls nicht an der Reliabilität, denn leider ist eine hohe Reliabilität zwar eine notwendige, aber keine hinreichende Bedingung[16] für eine hohe Validität: Ein Meßinstrument kann äußerst zuverlässig das Falsche messen.

Wir können natürlich nicht einfach „irgendetwas" nehmen, um zu messen, was wir messen wollen. Das klingt trivial, ist es aber nicht, wenn man bemerkt, daß mitunter Beobachtungssysteme - einmal eingeführt - offenbar auch (und gelegentlich *nur*) wegen ihrer Verfügbarkeit eingesetzt werden. Das mag im Einzelfall begründet sein, ist es aber wohl nicht immer (Foster & Cone, 1980, S. 327). Die mittlerweile leichter zugänglichen und häufiger vorgelegten Sammlungen von Verfahren (z.B. Manns et al., 1987; Simon & Boyer, 1974) erleichtern dies natürlich, wobei sie immerhin den Vorteil eines umfassenden Überblicks bieten, durch den man eher ein passendes Verfahren auswählen kann. Das Argument, das Manns et al. (1987) zugunsten dieser Sammlungen führen, ist dabei natürlich ebenfalls zutreffend und beachtenswert: Die Alternative, daß jede Untersuchung ihr eigenes, eigens zurechtgeschneidertes Verfahren oder System verwendet, führt zu Zersplitterung und Unvergleichbarkeit, d.h. zu einer unproduktiven und destruktiven Heterogenität in der Forschung.

Es gibt eine Reihe unterschiedlicher Konzeptionen von Validität. Wir verzichten an dieser Stelle aber zunächst darauf, sie ausführlicher zu besprechen (dies bietet sich in einem späteren Abschnitt besser an; siehe 4.4.2).

(3) Generalisierbarkeit

Es ist hier zunächst notwendig, sich klarzumachen, daß die Generalisierbarkeit einer (Beobachtungs-) Studie häufig gar nicht angezielt ist. Wenn wir im Sinne der deduktiven Beobachtung Daten erheben, geht es uns vielmehr darum, eine Theorie oder Hypothese zu prüfen, indem wir nach falsifizierender Information suchen. Wie wir gesehen haben (Abschnitt 1.4.2, Exkurs), genügt hierzu ein einziger Fall. Zur Widerlegung der Hypothese „Alle Brillenträger sind größer als 1,55 Meter" genügt eine einzige Beobachtung (eines 1,53 Meter großen Brillenträgers), mag er für Brillenträger repräsentativ sein oder nicht. Freilich gibt es Beobachtungen mit generalisierender Absicht. Die meisten heuristischen Beobachtungen werden einen entsprechenden (mehr

---

16 Sicherheitshalber wollen wir diese Begriffe kurz erläutern.
Der Regen ist *hinreichende Bedingung* dafür, daß der Rasen naß ist, aber nicht *notwendige Bedingung*, denn der Rasen wird auch dann naß, wenn ein Rasensprenger ihn tränkt. Der Rasensprenger ist, wenn es nur leicht nieselt, vielleicht eine *unterstützende Bedingung*, d.h. das Ergebnis (nasser Rasen) wird mit ihm schneller und gründlicher erreicht als ohne ihn.
An diesem Beispiel kann man auch erkennen, daß man niemals sicher sein kann, daß eine bestimmte Bedingung wirklich hinreichend ist. Genaugenommen ist der Regen nämlich z.B. dann nicht hinreichend, wenn der Rasen abgedeckt ist (z.B. weil es sich um den Rasen des „Center Court" in Wimbledon handelt, der bei einer Regenunterbrechung des Finales im Dameneinzel mit Planen geschützt wird). In diesem Falle ist es bei Regen für das Naßwerden genaugenommen *notwendige Bedingung*, daß der Rasen *nicht* abgedeckt ist. Allgemeiner kann man sagen: Der Regen und die Abwesenheit störender Bedingungen sind *zusammen hinreichend* dafür, daß der Rasen naß ist. Diese Floskel („die Abwesenheit...") nennt man die „ceteris paribus-Klausel": Unter „sonst gleichen" Bedingungen ist X hinreichend für Y. Unangenehmerweise ist mit ihr *jede* Vorhersage wahr: Wenn A dann B (A ist hinreichend für B), außer es tritt irgendetwas dazwischen, was das verhindert. Dies ist dasselbe wie: „Wenn A dann B oder Nicht-B"; und das stimmt immer.

oder weniger weitreichenden) Anspruch haben. So wollten Barker und Wright (1971; vgl. Abschnitt 2.1) natürlich nicht nur etwas über die acht von ihnen untersuchten Kinder herausfinden, und sie gingen ja auch ausdrücklich nicht von einer expliziten Fragestellung aus, die sie anhand dieser Kinder überprüfen wollten. Wieweit freilich ihre Befunde zu generalisieren sind (ob sie etwa auch für Kinder in Brooklyn zutreffen), ist eine schwierige Frage, die sie im übrigen auch selbst diskutieren.

Das Ergebnis einer Untersuchung (z.B. einer Beobachtungsstudie) kann aus mehreren Gründen nicht generalisierbar (d.h. für andere Personen als die untersuchte Stichprobe gültig) sein. Das Ergebnis kann zum einen fehlerbehaftet sein und daher nicht einmal replizierbar, geschweige denn auf andere Personen (-gruppen) generalisierbar sein. Verschiedene Fehlermöglichkeiten werden im folgenden Abschnitt (3.2.2) ausführlich diskutiert. Daneben könnte auch die gewählte Stichprobe nicht repräsentativ für andere Personengruppen sein, auf die generalisiert werden soll. Auch dieser Punkt wird kurz zur Sprache kommen. Die Generalisierbarkeitsdebatte insgesamt ist überdies kein spezielles Problem von Beobachtung als Datenerhebungsmethode (vgl. zur Übersicht z.B. Foster & Cone, 1980; Cone & Foster, 1982; Manns et al., 1987, S. 43ff.); sie gilt vielmehr ganz generell für an bestimmten Stichproben mit bestimmten Methoden gewonnene Daten.

Es gibt verschiedene Untersuchungen (zusammengestellt etwa bei Foster & Cone, 1980), die darauf hinweisen, daß unterschiedliche „settings" (d.h. Umgebungen im weitesten Sinne des Wortes, also inklusive aller sozialen und psychologischen Aspekte) unterschiedliches Verhalten produzieren. Dies ist jedoch kein Einwand gegen die Erhebungs*methode*. Um dies zu behaupten, müßte man zeigen, daß dieselben Beobachter in unterschiedlichen Settings unterschiedlich (und nicht: Unterschiedliches!) beobachten. Nach unserem Kenntnisstand gibt es dazu bislang wenig empirische Hinweise. Auf den ersten Blick wäre ein solcher Effekt (situations- und settingabhängige Beobachterfehler) im allgemeinen auch nicht sehr plausibel, wenn man von zwei Ausnahmen absieht:

1. Die Umstände beeinträchtigen die Beobachtung direkt (beispielsweise durch Geräusche, Beleuchtung, schlechte Sicht o.ä.; s.u., Abschnitt 3.2.2), oder
2. der Beobachter wird in einigen dieser Settings zum teilnehmenden Beobachter, in anderen nicht, bzw. allgemeiner: Der Beobachter hat in verschiedenen Settings verschiedene Rollen.

Der letzte Punkt weist einfach darauf hin, daß die Vergleichbarkeit der Methode gewährleistet bleiben muß (das ist bei teilnehmender versus nicht-teilnehmender Beobachtung nicht der Fall). Dieser Punkt gilt natürlich ebenso für die von Cone und Foster (1982) sogenannte „cross-laboratory" Generalisierbarkeit. In all diesen Fällen können empirische Untersuchungen, wie Cone und Foster (1980) sie berichten, nur dann Unterschiede (oder keine Unterschiede) erbringen, die man inhaltlich interpretieren darf, wenn man die Kontextunabhängigkeit der Datenge*winnung* voraussetzt. Das trifft insbesondere den von Cone und Foster hervorgehobenen Punkt unterschiedlicher Trainingsmethoden in verschiedenen Studien. Diese bilden möglicherweise

eine bedeutsame Varianzquelle (d.h. sie sind eine wichtige Ursache für Unterschiede in den Ergebnissen), aber mit ihnen ist eben die verwendete Methode nicht mehr dieselbe. Dies macht Replikationsstudien schwierig. Daraus folgt nun aber wiederum nicht, daß die Forderung nach Replizierbarkeit falsch wäre, sondern dies weist im Gegenteil auf die Notwendigkeit hin, Replikationen zu versuchen. Wenn sie trotz sorgfältiger Beachtung der sonstigen Randbedingungen scheitern, kann das ein Hinweis auf Methodenartefakte sein (d.h. auf Effekte und Ergebnisse, die nicht der untersuchten Wirklichkeit, sondern den Untersuchungsmethoden zugeschrieben werden müssen).

3.2.2 Konkrete Fehler: Eine Systematik

Wir wollen nun versuchen, die wichtigsten Fehler ausführlicher darzustellen und dabei etwas Ordnung in die vielen verschiedenen Quellen für Differenzen zwischen dem tatsächlichen Geschehen und dem Beobachtungsprotokoll (zwischen Wahrheit und Wiedergabe) zu bringen. Wie kann eine Systematik hierfür aussehen? Als Ordnungsgesichtspunkt schlagen wir (im Grundsatz hierin z.B. Beck, 1987, folgend) den „Ort" (man kann genausogut sagen: den „Zeitpunkt") in dem *Prozeß* vor, durch den die „Realität" über Wahrnehmung, Deutung, Speicherung und Wiedergabe schließlich zu einem intersubjektiv zugänglichen Protokoll (einer „Abbildung") verarbeitet wird. Abbildung 10 versucht, die möglichen Fehlerquellen in dieser Hinsicht geordnet darzustellen.

Beginnen wir den Prozeß (sozusagen erkenntnistheoretisch naiv) mit einer konkreten Tatsache, zum Beispiel einem bestimmten Verhalten, das möglichst genau beobachtet und wiedergegeben werden soll. Was kann nun alles zum letztlich vorliegenden Beobachtungsprotokoll beitragen, d.h. dieses Protokoll beeinflussen - *außer* der Tatsache selbst?

(1) Fehler zu Lasten des Beobachters

Die entscheidende Verarbeitung der „Information" – die eigentliche Beobachtung – findet natürlich im Beobachter statt. Hier lassen sich im Grundsatz mindestens drei unterschiedliche Prozeßabschnitte unterscheiden, in denen die jeweils „eintreffenden" Daten verarbeitet (modifiziert) oder einfach nur weitergereicht werden: Die Wahrnehmung, die Interpretation und die Erinnerung. Entsprechend können wir drei Fehlerquellen identifizieren: (1a) Wahrnehmungsfehler, (1b) Deutungs- und Interpretationsfehler, (1c) Erinnerungsfehler (Kapazitätsprobleme und Selektions- bzw. Modifikationsfehler). Hinzu kommt ein vierter Punkt, der in systematischer Hinsicht den vorigen Punkten zwar deutlich nachgeordnet ist, aber gleichwohl innerhalb der Person des Beobachters greift: (1d) Wiedergabefehler. Jedoch gibt es hier scharfe Grenzen eigentlich nicht. Es ist oft ganz unklar, ob ein bestimmter Verarbeitungsvorgang noch Wahrnehmung, schon Deutung oder gar Erinnerung ist. Bereits die Wahrnehmung selbst ist, wie wir gesehen haben, nicht eindimensional.

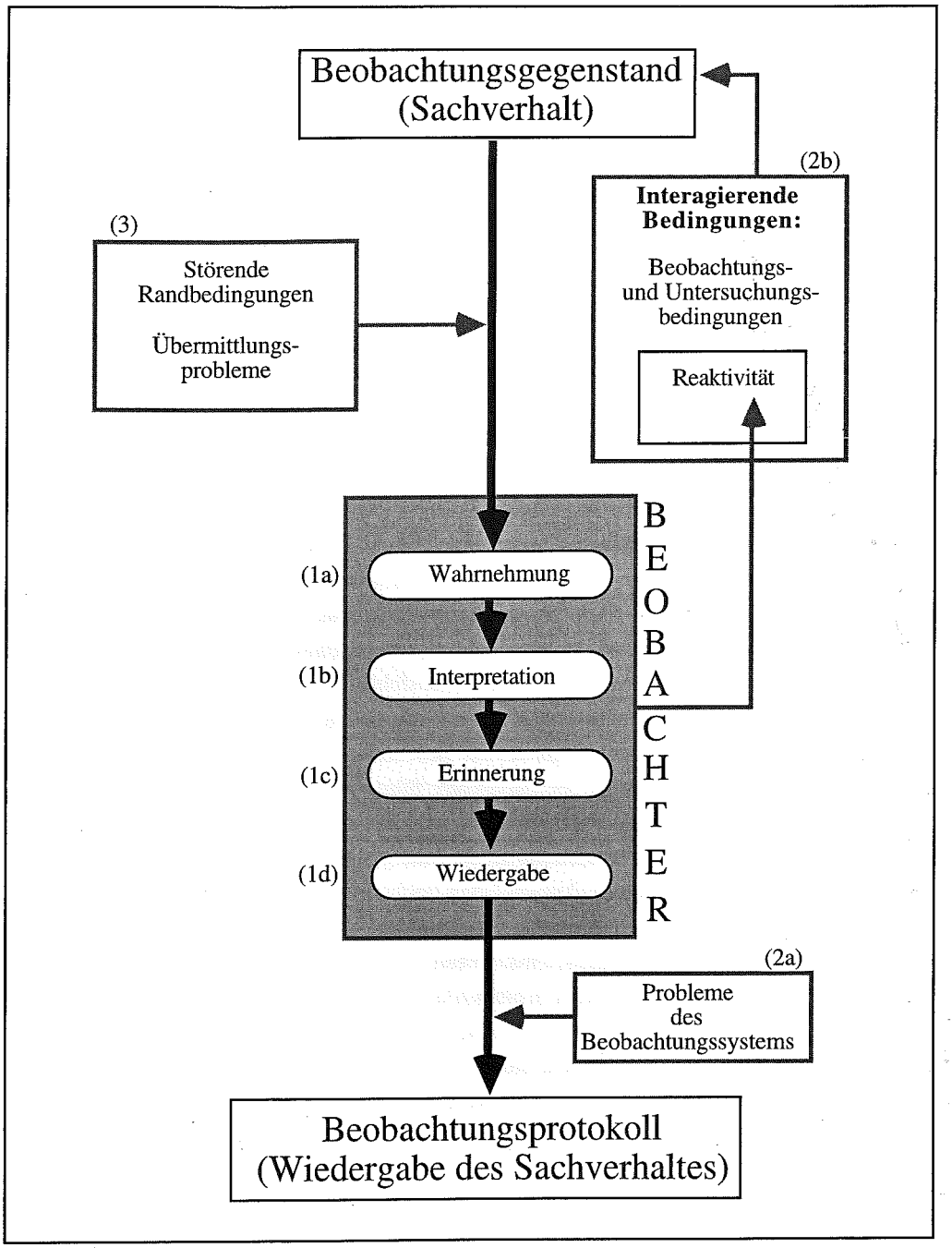


Abbildung 10: Beobachtungsfehlerquellen (Erläuterung im Text; die Numerierung entspricht der Gliederung in den folgenden Abschnitten)

(2) Fehler zu Lasten der Beobachtung

Hier müssen zwei Hauptgruppen unterschieden werden, hinter denen sich zwei völlig unterschiedliche Fehler- bzw. Einflußtypen verbergen. Zum einen kann es aufgrund des gewählten Vorgehens bei der Beobachtung zu Problemen kommen, das Beobachtete (und durch den Beobachter Verarbeitete) adäquat wiederzugeben. Zum anderen ist daran zu denken, daß die Tatsache der Beobachtung einen Einfluß auf den Beobachtungsgegenstand ausübt. Insbesondere der letztere ist ein außerordentlich wichtiger und vieldiskutierter Punkt.

(2a) Probleme des Beobachtungssystems. Durch das falsche Beobachtungssystem wird möglicherweise die trotz aller Modifikation durch den Beurteiler noch halbwegs erhalten gebliebene Genauigkeit und Zuverlässigkeit der Beobachtung vollends verschenkt. Für das grundsätzliche Verständnis dieses Problems ist die Vertrautheit mit den verschiedenen Beobachtungssystemen, die in den folgenden Kapiteln (4 und 5) diskutiert werden, gar nicht notwendig. Es genügt, wenn man sich vor Augen hält, daß in aller Regel das Beobachtete in vorgegebene Kategorien notiert bzw. eingetragen wird (erinnern wir uns an das Beispiel der Studie von Bandura im ersten Kapitel; siehe Abschnitt 1.1). Erzwungene und künstliche (zu enge oder zu weite) Kategorien verhindern, vermindern oder sabotieren Differenzierungen, die der Beobachter leisten könnte (bzw. die der Sache „angemessen" wären). Das „verschenkt" nicht nur u.U. wertvolle Information, sondern verzerrt möglicherweise auch die Darstellung (Faßnacht, 1995, spricht von einem „Klassifikationsfehler"; vgl. auch Pinther, 1972). Beck (1987) weist mit Bezug auf Feger (1972) darauf hin, daß erzwungene Kategorisierung, die nicht den bei freier Beschreibung gewählten Beschreibungsdimensionen des Beobachters entspricht, zu unreliablen Urteilen der Beobachter führt. Kurz gesagt: Ein unangemessenes Beobachtungssystem beeinflußt das Ergebnis.

Ein weiterer Punkt ist die Auswahl der Beobachtungsobjekte. Manns et al. (1987) weisen etwa darauf hin, daß eine falsche Auswahl der Beobachtungsgegenstände die Beobachtung verzerren kann (vgl. hierzu auch Boice, 1983). Das hängt natürlich von dem Interesse ab, mit dem die Beobachtung unternommen wird. Wenn es uns etwa um die („normalen") Kommunikations- und Interaktionsprozesse in einer Kleingruppe geht (eine typische Beobachtungsfragestellung in der Tradition von Bales, 1950a; vgl. Abschnitt 5.3), könnte ein allzu dominantes Gruppenmitglied in unserer Beobachtungsstichprobe unseren Eindruck von diesen typischen Prozessen und Rollen in Kleingruppen verzerren. Allgemeiner ausgedrückt: Nicht immer ist das individuelle Verhalten das, was uns interessiert. Aber natürlich geht es uns häufig, beispielsweise im Zusammenhang mit klinischen oder therapeutischen Diagnosen, tatsächlich genau darum, das Verhalten dieser konkreten Person(en) genau kennenzulernen, gerade wenn es von der Norm deutlich abweicht. Es handelt sich hier demnach um einen „sampling"- oder Stichprobenfehler (vgl. hierzu auch von Cranach & Frenz, 1969, S. 293ff.); er besagt nichts anderes, als daß wir uns möglicherweise eine für unsere Untersuchungsabsicht ungeeignete Stichprobe von Personen ansehen. (So gesehen handelt es sich also um einen „Fehler zu Lasten

des Untersuchungsleiters"; gelegentlich findet sich dafür auch die irreführende Kennzeichnung als „Fehler zu Lasten des Beobachteten".)

In einem gewissen (etwas weiten) Sinne könnte man diesem Punkt auch Auswertungsfehler zuordnen. Dies deswegen, weil das eigentlich interessierende Ergebnis der Beobachtung ja nicht der Berg von Strichlisten oder anderen „Roh-Protokollen", sondern etwas allgemeiner „der Befund" ist, der sich aus der Auswertung des Rohmaterials ergibt. (Dies ist ja ein Einwand gegen extrem umfangreiche Materialsammlungen im Sinne etwa von Barker und Wright gewesen; siehe Abschnitt 2.1.) Auswertungsfehler verzerren jedoch nicht eigentlich die Beobachtung. Die Auswertung der Beobachtungsdaten ist überdies - abgesehen von der Berechnung von Beobachterübereinstimmungswerten - nicht unser Thema. Wir wollen es daher hier mit dem Hinweis bewenden lassen, daß man auch in der Weiterverarbeitung der mühsam gewonnenen (und vielleicht sogar nur wenig „verunreinigten") Daten gravierende Fehler begehen kann, die das Ergebnis vollständig verfälschen (Literaturhinweise dazu finden sich zum Beispiel bei Manns et al., 1987).

(2b) Interagierende Bedingungen. Ein viel wichtigerer Punkt ist jedoch die zweite Kategorie von „Fehlern zu Lasten der Beobachtung", die Frage nämlich, ob die Beobachtung selbst möglicherweise das beeinflußt, was beobachtet werden soll. Wir kommen darauf im nächsten Abschnitt ausführlich zu sprechen (Abschnitt 3.2.3).

(3) Fehler zu Lasten äußerer Bedingungen

Schließlich können natürlich auch andere, d.h. von der zu beobachtenden Tatsache und der Tatsache der Beobachtung unabhängige „reale" Einflüsse die Beobachtung stören. Durch ungünstige äußere Beobachtungsbedingungen (z.B. schlechte Lichtverhältnisse, störende bzw. überlagernde Geräusche) kommt die Wirklichkeit beim Beobachter sozusagen gar nicht richtig oder vollständig an. Eine zweite Einflußquelle, an die hier gegebenenfalls zu denken wäre, sind Verzerrungen und Selektionen, die durch vermittelnde technische Geräte auftreten (z.B. Verzerrungen des Tonfalles bei Bandaufnahmen, Unbemerkbarkeit kleiner Reize durch handwerklich oder technisch schlechte Filmaufnahmen, Probleme der Wiedergabegeräte etc.; s.o., Abschnitt 1.5).

Diese Fehlerquellen sind durch sorgfältige Planung der Beobachtung vermeidbar: Sorgfalt bei der Wahl der Beobachtungsbedingungen, natürlich auch Sorgfalt bei der Wahl der Beobachter selbst, ist - wenn irgend möglich - unbedingt erforderlich. Ausnahmen sind beispielsweise naturalistische Beobachtungen (unter den natürlichen Bedingungen, die eben nicht verändert werden sollen) oder etwa Einschränkungen, die in Kauf genommen werden müssen, wenn verdeckt beobachtet werden soll. Diese Fehlerquellen werden wir im folgenden nicht einzeln und ausführlich diskutieren.

3.2.3 Konkrete Fehler: Exemplarisch vertiefende Diskussionen

Dieser Abschnitt liefert die angekündigte inhaltliche Darstellung und Diskussion der wichtigsten Beobachtungsfehler und -quellen. Er folgt dabei der soeben erläuterten Systematik (Fehler zu Lasten äußerer Bedingungen werden dabei, wie bereits oben angekündigt, nicht mehr eigens diskutiert). Da die Zahl dieser Fehler recht groß ist, wollen wir zur besseren Übersicht vorab die (Unter-) Gliederung dieses Abschnitts explizit auflisten (Tab. 2).

Tabelle 2: Übersicht über die wichtigsten konkreten Fehlerquellen

1	Fehler zu Lasten des Beobachters
(1a)	Wahrnehmung
	1. Konsistenzeffekte
	2. Einfluß vorangehender Informationen
	3. Projektion
	4. Erwartungseffekte
	5. Emotionale Beteiligung
	6. Logischer oder theoretischer Fehler
	7. „Observer drift"
(1b)	Interpretation
	1. Zentrale Tendenz
	2. Persönliche Tendenzen oder Dispositionen
(1c)	Erinnerung
	1. Kapazitätsgrenzen
	2. Erinnerungsverzerrungen und -selektionen
(1d)	Wiedergabe
2	Fehler zu Lasten der Beobachtung
	1. Reaktivitäts- und Erwartungseffekte
	2. Beobachtungs- und Untersuchungsbedingungen
	3. Probleme des Beobachtungssystems

(1) Fehler zu Lasten des Beobachters

In den folgenden Abschnitten wollen wir nun zunächst die Fehlerquellen etwas genauer betrachten, die durch den bzw. „in" dem Beobachter auftreten. Wir folgen dabei weiterhin der Gliederung des vorigen Abschnittes bzw. der Abbildung 10.

(1a) Wahrnehmung

Wie bereits oben angedeutet, umfaßt dieser Punkt eine Vielzahl recht heterogener Fehlerquellen.[17] Sie sind hier unter einer Überschrift subsumiert, weil in allen Fällen das bewußt Wahrgenommene bereits durch sie verzerrt ist, oder anders gesagt: das, was der Beobachter beobachtet, bereits durch sie mitbestimmt ist.

1. Konsistenzeffekte

Damit ist die Tendenz gemeint, in seinen Äußerungen, Meinungen, also auch seinen Urteilen im großen und ganzen konsistent, d.h. widerspruchsfrei zu bleiben. Der wichtigste, jedenfalls der vermutlich meistzitierte und -diskutierte Effekt ist der sogenannte *Halo-Effekt* (zur Diskussion vgl. z.B. Zapf, 1989; zum Überblick vgl. etwa von Cranach & Frenz, 1969; Schaller, 1980). Damit ist die Tendenz gemeint, einzelne Urteile in Abhängigkeit von einem bestehenden Gesamteindruck oder einem besonders hervorstechenden Merkmal zu fällen, wobei „in Abhängigkeit" in aller Regel bedeutet: in *Übereinstimmung* mit diesen.

> Ein Beispiel dafür bietet etwa eine Studie von Nisbett und Wilson (1977). Die Beobachter schätzen hier die äußere Erscheinung oder den Akzent einer Person dann eher als ansprechend ein, wenn sich diese Person als „warme" und freundliche Person gab, und als eher befremdlich, wenn sie sich „kalt" und distanziert verhielt.

Dieser Effekt ist früh entdeckt worden (Fisicaro, 1988, verweist auf eine Arbeit von Wells aus dem Jahre 1907); seinen Namen verdankt er Edward L. Thorndike (1920), der damit das „Ausstrahlen" des hervorstechenden Merkmals auf andere (und eines allgemeinen Eindrucks auf spezielle Bereiche) umschreiben wollte. Fisicaro hat erst kürzlich (1988) die traditionelle Auffassung, daß der Halo-Fehler sich in einer verringerten Genauigkeit niederschlägt, gegen einige neuere (scheinbar empirisch begründete) Zweifel überzeugend verteidigt (zu Details dieses Phänomens vgl. etwa die Experimente von Klauer, 1991).

Generell kann das Bemühen um Konsistenz aber auch das Bestreben betreffen, von einem ersten, vielleicht vorschnell und unüberlegt abgegebenen Urteil im weiteren Verlauf der Beurteilungen nicht mehr allzusehr abzuweichen[18], oder anders gesagt: die Tendenz, frühere Äußerungen oder Überzeugungen im Zweifel lieber zu bestätigen als zu widerlegen (vgl. hierzu im Überblick auch Schaller, 1980; Pinther, 1972). Häufig werden hier verschiedene Fehler unterschieden; Beck (1987) nennt etwa jeweils einzeln: Generalisierungsfehler, Nachbarschaftseffekt sowie den ebenfalls teilweise hierher gehörenden Vorwissensfehler.

---

17 Genaugenommen müßte man in diesem Zusammenhang auch physiologische Einschränkungen der Wahrnehmungsfähigkeit des Beobachters (Sehfehler, Hörschwächen etc.) diskutieren. Wir verzichten hier deswegen darauf, weil sie psychologisch relativ wenig interessant sind; das bedeutet freilich nicht, daß sie nicht ernst zu nehmen seien oder gar unberücksichtigt bleiben könnten.

18 Dieses Phänomen wird häufig auch mit „primacy-effect" beschrieben, das komplementäre Phänomen, daß der letzte Eindruck nachträglich die vorherigen „einfärbt" mit „recency-effect". Wir verwenden dieses Begriffspaar hier jedoch für einen Erinnerungsfehler (s.u.).

2. Einfluß vorangehender Informationen

Dies ist ein sehr heterogener Punkt, der im Grunde nur deswegen hier separat behandelt wird, weil er oft in entsprechenden Übersichtsartikeln einzeln genannt wird (z.B. Pinther, 1972; Schaller, 1980; Beck, 1987). Diese vorhergehenden Informationen können nämlich sein:

- die Information, was die anderen Beobachter gesehen (berichtet) haben („Konformität"),
- die Information, was der Untersuchungsleiter erwartet oder was er gerade wie kommentiert hat („Erwartungseffekt"),
- die Information, was man selbst vorher behauptet hat, gesehen zu haben („Konsistenzeffekt"), bzw.
- die Information, was man tatsächlich zuvor oder gleichzeitig noch gesehen hat (Beck, 1987, spricht von einem „Nachbarschaftseffekt", der häufig ebenfalls ein „Konsistenzeffekt" ist).

3. Projektion

Gelegentlich finden sich Hinweise auf die Tendenz von Beobachtern, in den Beobachteten das wiederzuerkennen, was sie bei sich selbst sehen, sehen wollen oder gerade nicht sehen wollen. In diese Fehlerkategorie gehört auch die ungerechtfertigte Wahrnehmung (und der entsprechende Einfluß) von einer „Nähe" oder Ähnlichkeit zum Beobachteten. Der Begriff der „Projektion" (und ebenso der „Übertragung") stammt aus der psychoanalytischen Denktradition. Handfeste empirische Studien zu ihrer Bedeutung für die Verhaltensbeobachtung sind uns nicht bekannt. Es ist mehr als unklar, inwieweit überhaupt und falls ja, in welcher Richtung und in welchem Ausmaß sich solche Effekte auf konkrete Verhaltensbeobachtungen auswirken bzw. sich in entsprechenden Protokollen niederschlagen. Generell wird insbesondere für Fehler dieser Art (mit unklarer Richtung und unklarer Stärke) zutreffen, daß sie zwischen verschiedenen Beobachtern annähernd zufallsverteilt sind, d.h. durch den Einsatz von mehreren Beobachtern aufgefangen werden.[19]

4. Erwartungseffekt des Beobachters: „Wer sucht, der findet"

Die Frage, die sich hinter diesem Punkt verbirgt (vgl. zum Überblick: Kazdin, 1977; Mees, 1977a; Pinther, 1972; Cone & Foster, 1982; Kent & Foster, 1977; Johnson & Bolstad, 1973), ist die folgende: Sieht der Beobachter, was er sehen will bzw. was er glaubt sehen zu sollen, d.h. neigt er zu hypothesenkonformen Einschätzungen? Es ist wichtig, sich dabei vor Augen zu halten, daß das, was der Beobachter in derartigen Fällen zu sehen glaubt, nicht wirklich zu sehen ist: Ein „unbefangener" Beobachter würde etwas anderes sehen! Dieser Erwartungseffekt ist in gewissem Sinne ein Spezialfall der Beurteilung im Sinne (in der Tendenz) einer impliziten, d.h. unausgesprochenen, vielleicht unbewußten „Theorie" (Überzeugung) (s.u.). Generell

---

19 Das gilt nicht für alle der hier diskutierten Fehler. Fehler, die durch eine spezielle Auswahl, ein spezielles Training oder gemeinsame Veränderungen der Beobachter („drift"; siehe jeweils dort) auftreten, können so nicht kontrolliert werden.

gilt jedoch bei diesen Erwartungseffekten, daß die Beobachtungen (ggf. die Beurteilungen) sämtlich *in einer bestimmten Richtung* verzerrt werden.

Einschlägig sind hier vor allem die Arbeiten von Robert Rosenthal (1976; Rosenthal & Rosnow, 1969; eine Zusammenstellung findet sich in Rosenthal, 1977). Er hat - sogar für experimentelle Untersuchungen - in einer Reihe aufsehenerregender Studien nachzuweisen versucht, daß Versuchs- bzw. Untersuchungsleiter die Tendenz haben, das zu finden, was sie zu finden erwarten. Cone und Foster (1982) weisen darauf hin, daß ein solcher Erwartungseffekt bei teilnehmenden Beobachtern möglicherweise stärker ausgeprägt sein kann, weil diese häufig ein persönliches Interesse an den beobachteten Personen haben. Dieser Punkt trifft natürlich auf zahlreiche Beobachterfehler zu (s.o.: Vorurteile, Milde, Ähnlichkeit etc.).

Das kann dann längerfristig allerdings auch zur Folge haben, daß sich das Verhalten der Beobachteten in der entsprechenden Weise verändert (s.u.: Erwartungseffekt des Beobachteten). Berühmt geworden ist hierfür die Studie von Rosenthal und Jacobsen (1968) zum „Pygmalion-Effekt", über die wir im ersten Kapitel bereits berichtet haben (Abschnitt 1.3). Die Meinung, die man über jemanden hat, beeinflußt natürlich das Verhalten gegenüber dieser Person in vielfältigster und subtilster Weise, und dies beeinflußt wiederum das Verhalten der Person selbst.

Dieser Erwartungseffekt spielt jedoch möglicherweise bei Beobachtungsstudien keine so große Rolle. Es gibt mittlerweile mehrere Studien, die darauf hinweisen, daß die konkreten einzelnen Verhaltensbeobachtungen von den Erwartungen der Beobachter nicht oder nur wenig beeinflußt werden (Kent, O'Leary, Diamet & Dietz, 1974; Shuller & McNamara, 1976). Das Resümee von Kent und Foster (1977) – „behavioral observation procedures in the majority of cases seem to be largely unobtrusive, unaffected by the expectations of the examiner" (S. 323) – erscheint allerdings etwas zu optimistisch. So findet sich auch in den genannten Studien ein Erwartungseffekt relativ zuverlässig immer dann, wenn die Beobachter außerdem nach ihren globalen (summarischen) Einschätzungen dessen, was sie beobachtet hatten, gefragt wurden: *Diese* Einschätzungen wurden deutlich von den (zuvor induzierten) Erwartungen beeinflußt. Entsprechend plädiert etwa Beck (1987) dafür, sich von globaleren Kategoriensystemen grundsätzlich abzuwenden und verstärkt wieder elementarere Verhaltenseinheiten zu verwenden. Diese seien reliabler, und die verschiedenen (weiteren) Verarbeitungsschritte, die beim Ratingsystem das Endergebnis in ganz unkontrollierbarer Weise bestimmten, könnten explizit gemacht und auseinandergehalten werden. Eine wichtige, empirisch zu beantwortende Frage ist demnach, ob induzierte Erwartungen auch dann keinen Effekt auf konkrete Verhaltensbeobachtungen haben, wenn sich auch die Erwartungen auf konkrete Verhaltensweisen und nicht nur auf globale Hypothesen beziehen.

Eine Art „Erziehung" der Beobachtungsleistungen in Richtung auf die erwünschten Hypothesen scheint sich dann zu ergeben, wenn der Untersuchungsleiter zwischen den verschiedenen Beobachtungsdurchgängen die Zwischenergebnisse mit den Beobachtern bespricht und (eindeutig) kommentiert: „Sie sehen dieses Phänomen offenbar schlecht", „Dieser Punkt wird von Ihnen ganz offensichtlich stark überschätzt" o.ä. (vgl. dazu Cone & Foster,

1982). Kent und Foster (1977) berichten von einer Studie von O'Leary, Kent und Kanowitz (1975), derzufolge Kommentare des Untersuchungsleiters, die er bei Besprechung der Zwischenergebnisse abgibt, die folgenden Beobachtungsergebnisse deutlich beeinflußten.

Von diesem Erwartungseffekt sorgfältig zu unterscheiden ist ein Erwartungseffekt des Beobachteten. Bei ihm geht es um die Frage, ob die Beobachteten sich *tatsächlich* in einer Weise verhalten, die (ihrer Vermutung nach) der Beobachter erwartet. Dieses Verhalten würde dann auch ein unbefangener Beobachter sehen. Wir werden diesen Erwartungseffekt („2. Art") unter dem Abschnitt diskutieren, in den er in systematischer Hinsicht gehört: Fehler zu Lasten der Beobachtung (hier: zu Lasten interagierender Bedingungen).

5. Emotionale Beteiligung

Auf die Gefahr, daß die Beobachter ein persönliches Interesse an den Beobachteten haben könnten und dadurch in ihrem Urteil beeinflußt werden, wurde bereits hingewiesen. Beck (1987) spricht von einem Wertungseffekt. Mit Hinweis auf diesen Punkt wird beispielsweise oft vor teilnehmender Beobachtung gewarnt (Cone & Foster, 1982). Auch Erwartungseffekte können über emotionale Beteiligung der Beobachter vermittelt sein. So weisen etwa Manns et al. (1987) darauf hin, daß etwa Psychologiestudenten, wenn sie als Beobachter verwendet werden (ein für viele Studien gängiges Verfahren), sich z.B. daran stören, daß (1) in der Regel das offene Verhalten, nicht aber die Gefühle der Beobachteten berücksichtigt würden, oder daß man (2) nur zusehen, nicht aber sich einmischen (z.B. helfen, klären oder schützen) dürfe.

6. Logischer oder theoretischer Fehler

Damit ist die Tendenz gemeint, die Wirklichkeit im Sinne bzw. „im Lichte" bestimmter „naiver" Annahmen, Vorurteile oder „Theorien" zu beurteilen bzw. zu sehen. Über den grundsätzlichen Einfluß der seitens des Beobachters vorliegenden Strukturen haben wir oben ausführlicher gesprochen (Abschnitt 3.1). Hier ist - außer den grundsätzlichen, wahrnehmungskonstitutiven semantischen, konzeptuellen und syntaktischen Strukturen - vor allem auch an inhaltliche „implizite Persönlichkeitstheorien" und Vorurteile gedacht („Brillenträger sind intelligent"), die das Urteil, ja mitunter schon die Wahrnehmung (z.B. durch Selektion) bestimmter Verhaltensweisen beeinflussen können.

7. „Observer drift"

Damit ist die allmähliche Veränderung des „Standards" eines Beobachters gemeint. Dafür kann es natürlich mehrere Ursachen geben, die sich auch keineswegs ausschließen: Ein zunehmendes Vergessen der seinerzeit erlernten und geübten Kriterien, eine zunehmende Ermüdung, eine nachlassende Motivation, der Erwerb von „störenden" Gewohnheiten, aber natürlich auch eine mit steigender Übung verbesserte Wahrnehungsschärfe oder eine durch zunehmende Vertrautheit mit dem Beobachtungsgegenstand veränderte Einstellung (vgl. z.B. Boice, 1983; Manns et al., 1987).

Dieser Fehler fällt aber normalerweise bei regelmäßiger Kontrolle der Beobachterübereinstimmung sofort auf. Besonders gefährlich ist daher das Phänomen des „consensual drift", d.h. der *gleichsinnigen* allmählichen Veränderung der „Standards" in einer Gruppe von Beobachtern über eine längere Zeitspanne hinweg. Diese gesamte Beobachtergruppe revidiert sozusagen unbemerkt die ursprünglich angewendeten Kriterien. Wenn das innerhalb der Gruppe gleichsinnig geschieht, schlägt sich diese Veränderung natürlich *nicht* in veränderter Beobachterübereinstimmung nieder (vgl. dazu im Überblick: Boice, 1983; Johnson & Bolstad, 1973; Kazdin, 1977; Foster & Cone, 1980; Cone & Foster, 1982; Wildman & Erickson, 1977; Kent & Foster, 1977; Manns et al., 1987). Dazu paßt der häufige Befund, daß die Übereinstimmung innerhalb einer Gruppe von Beobachtern in der Regel größer zu sein pflegt als zwischen Beobachtern verschiedener Gruppen, sogar dann, wenn sie ursprünglich gemeinsam trainiert wurden (Kent, 1972; zitiert nach Kent & Foster, 1977). In die gleiche Richtung weisen auch Befunde aus Untersuchungen mit teilnehmenden Beobachtern: Die Übereinstimmung der Beobachtungsleistungen zwischen Ehepartnern ist höher als die der einzelnen Partner mit externen Beobachtern (Margolin et al., 1985; s.o., Abschnitt 1.5) In einer Untersuchung von O'Leary und Kent (1973; zit. bei Kent et al., 1974) zeigten sich ebenfalls entsprechende Effekte.

Eine praktikable Empfehlung ist, immer wieder Eichbeobachter einzusetzen oder die Beobachter immer wieder mit Filmen o.ä. zu eichen (Kent & Foster, 1977; Johnson & Bolstad, 1973). Eine andere naheliegende (vorbeugende) Möglichkeit wäre, feste Beobachtergruppierungen über längere Zeiträume zu vermeiden (Mees, 1977a). Dieses Verfahren setzt allerdings eine hinreichend große Auswahl von Beobachtern voraus, die man immer wieder neu gruppieren kann.

(1b) Interpretation

Unter diesem Punkt sollen die Fehlerquellen zusammengefaßt werden, die sozusagen das Wahrgenommene (und dabei ggf. durch die im vorigen Abschnitt diskutierten Fehler Verzerrte) beeinflussen.

1. Zentrale Tendenz

Mit dem Fehler der zentralen Tendenz (vgl. dazu Zapf, 1989, S. 45ff.; von Cranach & Frenz, 1969; Pinther, 1972) sind zwei ganz unterschiedliche Effekte angesprochen: Ein Urteilsfehler und ein statistisches Artefakt. Im Zusammenhang psychologischer Beobachtung (und in der entsprechenden Literatur) geht es in aller Regel um den ersteren. Mit diesem Urteilsfehler ist einfach gemeint, daß *Personen* die Tendenz haben, beispielsweise bei sogenannten Schätz- oder Ratingskalen extreme Urteile zu vermeiden.

Was ist überhaupt eine *Ratingskala* (oder Schätzskala)? Wir verschieben die ausführliche Erläuterung noch etwas (Abschnitt 5.4) und beschränken uns statt dessen einstweilen auf ein kleines Beispiel:
Wie aggressiv war das Verhalten des Beobachteten? Bitte kreuzen Sie einen Wert auf der folgenden Skala an:

Man könnte die Extrempunkte („0" bzw. „6") beispielsweise deswegen meiden, weil man sich zum Rand der zur Verfügung stehenden Urteilsskala für möglicherweise noch zu erwartende extremere Fälle einen Spielraum bewahren möchte. Beck (1987, S. 186) ordnet diesen Fehler der Kategorie „output" zu, d.h. er unterstellt, daß es sich um ein Problem handelt, das erst beim Ankreuzen, nicht schon beim Einschätzen selbst entsteht. Im Lichte der vorliegenden Befunde sind derartige Fragen (Verarbeitungs- oder Wiedergabefehler?) selten begründet zu beantworten.

Davon zu unterscheiden ist eine Verwendung des Begriffs einer „zentralen Tendenz" zur Bezeichnung eines Artefaktes im Kontext statistischer Analysen von Meßwertfolgen. In diesem Zusammenhang ist damit das Phänomen gemeint, daß bei einer Wiederholungsmessung der zweite Wert einer Person näher am Stichprobenmittelwert erwartet wird. Diese „Tendenz" ist umso stärker, je weiter der erste Wert vom Mittelwert entfernt ist. Diese sogenannte „Regression zur Mitte" wird außerdem umso größer, je geringer der statistische Zusammenhang der Messungen ist (mit dem ihre Reliabilität geschätzt wird). Eine genauere Erläuterung dieses statistischen Artefaktes kann man etwa bei Bortz und Döring (1995, S. 517ff.) nachlesen.

Warum erwähnen wir das hier überhaupt? Stellen wir uns den naheliegenden Fall vor, daß von mehreren Personen bestimmte Verhaltensweisen zu mehreren aufeinanderfolgenden Zeitpunkten beobachtet und eingeschätzt werden (beispielsweise hinsichtlich der Aggressivität). Wenn nun bei der Auswertung auffällt, daß extreme Werte einer Person (bei der ersten Beobachtung) in der folgenden Beobachtung näher am Mittelwert liegen, könnte man einen inhaltlichen Fehler (z. B. eine „Tendenz zur Abschwächung" o.ä.) bei den Beobachtern vermuten. Dieser Effekt ist in diesem Fall jedoch nur eine Konsequenz der einfachen Tatsache, daß Beobachtungen nicht perfekt reliabel sind.

2. Persönliche Tendenzen und Dispositionen

Ungeachtet der obigen Einschränkung kann es aber auch auf individueller Ebene eine stabile Tendenz zur Milde (bzw. entsprechend: eine Tendenz zur Strenge) geben. Damit ist eine personenspezifische Neigung zu generell milden oder generell strengen Urteilen in Abhängigkeit einerseits von der persönlichen Neigung und andererseits von den wahrgenommenen Eigenschaften des Beurteilten gemeint (vgl. etwa Zapf, 1989; Beck, 1987; Pinther, 1972). Neben dieser gibt es weitere generelle Urteilstendenzen, die hier wenigstens erwähnt werden sollen:

- Tendenz zur Zustimmung: „Ja-sage"-Tendenz auf eine gerichtete Auswertungsfrage (z.B.: „War das Kind kooperativ?"; vgl. hierzu z.B. Zapf, 1989, S. 45ff.);

- Tendenz zur Kontrastbildung: Das aktuell geforderte Urteil wird hierbei mit einem anderen (z.B. dem vorigen) oder mit dem Selbstbild des Beobachters kontrastiert (vgl. z.B. Schaller, 1980);
- Tendenz zur „sozialen Erwünschtheit": Damit ist die Tendenz gemeint, seine Urteile in einer Richtung zu modifizieren, die man für die allgemein sozial gewünschte hält (z.B. „Kinder sind doch nicht wirklich bösartig, so ein Urteil 'darf' man doch nicht fällen!").

Diese Tendenzen machen sich vor allem bei verbalen Protokollen bemerkbar, aber auch bei Ratingskalen, z.B. zu Persönlichkeitseigenschaften (von Cranach & Frenz, 1969). Die differentiellen Effekte (d.h.: Wann neigt wer bei wem zu unangebrachter Milde bzw. Strenge?) sind natürlich außerordentlich komplex. Wichtig ist, daß man derartige Fehlerquellen im Auge behält und möglichst durch entsprechende Beobachtervariationen kontrolliert.

(1c) Erinnerung

Es ist zunächst wichtig, darauf hinzuweisen, daß Fehler durch unvollständige oder verzerrte Erinnerungen nur dann auftreten, wenn die Beobachtung nicht unmittelbar protokolliert wird, sondern nachträglich aufgezeichnet wird (eine minimale Verzögerung wird es aber in aller Regel geben, daher ist dieser Schritt in Abb. 10 nicht nur optional). Wie wir gesehen haben, ist dies insbesondere bei verdeckter Beobachtung mitunter notwendig, aber auch bei teilnehmender Beobachtung mag es sinnvoll sein, nicht simultan mit dem gezeigten Verhalten, sondern mit einer gewissen Verzögerung zu protokollieren, um einen störenden Einfluß auf das beobachtete Verhalten zu verringern. Vor allem bei Ratingverfahren (vgl. Abschnitt 5.4) wird die geforderte Einschätzung häufig für größere Zeitabschnitte retrospektiv vorgenommen.

1. Kapazitätsgrenzen

Es ist - bei Beobachtungsstudien wie auch allgemein - wichtig zu beachten, daß Menschen nur begrenzte Informationsmengen (gleichzeitig) aufnehmen und verarbeiten (d.h. u.a. behalten) können. In diesem Zusammenhang wird häufig der sogenannte „primacy-recency"- Effekt (vgl. z.B. Bredenkamp & Wippich, 1977) genannt, der auf das Phänomen hinweist, daß man sich von einer Reihe von Reizen (z.B. Zahlen, Silben, Wörtern, aber auch Gesten etc.) insbesondere die ersten und die letzten besonders gut merken kann.

2. Systematische Erinnerungsverzerrungen und -selektionen

Das menschliche Gedächtnis speichert nicht nur einfach die wahrgenommenen Sachverhalte „neutral" und mit hoher Wiedergabetreue, sondern „baut" sie gewissermaßen in ein vorhandenes Netz von Vorwissen und Erfahrungen ein. Dabei und bei dem späteren „Abrufen" dieser Informationen werden sie - ganz analog zu den diskutierten Prozessen bei der Wahrnehmung und Interpretation - modifiziert. Wird also eine Beobachtung nachträglich aufgezeichnet, können die unter (1; Wahrnehmung) und (2; Interpretation) diskutierten Effekte nochmals beim Er-

innern auftreten: Modifizierte Wahrnehmungen werden so bei bzw. durch Speicherung und „Abruf" möglicherweise nochmals modifiziert. Hierher gehören etwa Effekte wie die „abgerundete" Erinnerung, wie sie in den Experimenten von Loftus und Palmer (1974; Abschnitt 3.1) auftrat.

Zur Verdeutlichung wiederum ein „Klassiker" der experimentellen Psychologie: Carmichael, Hogan und Walter (1932) zeigten ihren Versuchspersonen einfache Figuren (ein Beispiel zeigt Abb. 11).

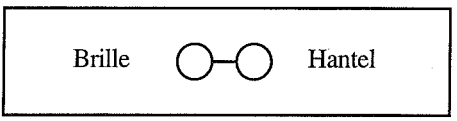


Abbildung 11: Versuchsmaterialbeispiel (orig. vgl. Carmichael, Hogan & Walter, 1932, S. 75)

Die Figuren waren immer als stilisierte Abbildungen mindestens zweier Objekte zu interpretieren. So kann man (mit etwas Wohlwollen) die Figur in Abb. 11 als Brille ebenso deuten wie als Hantel. Der Versuchsplan bestand nun darin, einer Gruppe von Versuchspersonen die Figuren in Kombination mit jeweils einem der beiden Begriffe vorzulegen. Bei der anderen Gruppe wurde ebenso verfahren; es wurde lediglich der jeweils andere Begriff mit der Figur gezeigt. Nachdem sich alle Versuchspersonen die Figur-Begriff-Kombinationen eingeprägt hatten, wurden sie später gebeten, die Figuren, die sie gesehen hatten, zu zeichnen. Diese Zeichnungen der Versuchspersonen waren nun - wie zu erwarten - ein mehr oder weniger verzerrtes Abbild der tatsächlichen Bilder. Das interessante Ergebnis dieses kleinen Versuchs bestand nun aber darin, daß sich für die beiden Gruppen ganz charakteristische Verzerrungen ergaben. So war ein typisches Ergebnis der ersten Gruppe (die Abb. 11 in Kombination mit dem Begriff 'Brille' gesehen hatte) ein etwas in Richtung Brille modifiziertes Bild (Abb. 12 links), während ein typisches Ergebnis der anderen Gruppe ('Hantel') in der Abbildung 12 rechts wiedergegeben wird.

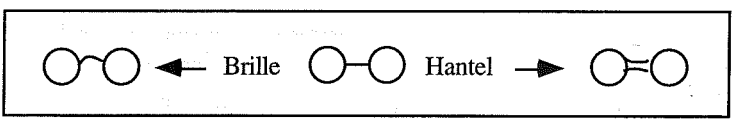


Abbildung 12: Ergebnisbeispiel (orig. vgl. Carmichael, Hogan & Walter, 1932, S. 80)

Begriff und Figur waren also in bestimmter Weise zusammen abgespeichert; beim Abruf wurde dann die Figur - in charakteristischer Weise verzerrt - „re-konstruiert".

(1d) Wiedergabe

Bereits 1909 ermahnt Whipple seine Kollegen: „we fail to keep in mind that the observer not only observes, but also reports, and that it is not only possible, but practically certain, that the report is only a partial often a misleading statement of the real experience" (Whipple, 1909, S. 153). So ist zum Beispiel nicht von vorneherein sicher, daß die Beobachter ihre Urteile - wie immer sie zu ihnen gekommen sind, wie verzerrt oder unverzerrt sie auch immer sein mögen -

genauso weiter- bzw. wiedergeben, wie sie sie „empfinden". Absehen werden wir hier natürlich von der Möglichkeit, daß sie ihre Protokolle bewußt „schönen", weil sie ein bestimmtes Interesse an den Ergebnissen (oder an der Täuschung des Versuchsleiters oder Geldgebers etc.) haben (vgl. dazu z.B. Kent & Foster, 1977).[20] Wichtig ist z.B. ein Effekt, den man mit „Konformitätsdruck" bezeichnen könnte. Klassisch ist hier ein berühmtes Experiment von Asch.

Asch (1955) bat Versuchspersonen, zu schätzen, welche der drei Linien (A, B oder C; vgl. Abb. 13) der Länge der Vergleichslinie (X) am nächsten kam. In der „Kontrollbedingung" gaben die Teilnehmer ihre Schätzung allein und unbeobachtet ab, in der „Experimentalbedingung" dagegen saßen sie mit sieben Personen zusammen in einem Raum, die *vor* ihnen ihre Einschätzung abgaben. Die Pointe besteht nun darin, daß diese anderen Personen „Verbündete" des Leiters waren (was die achte, die eigentliche Versuchsperson, nicht wußte) und absichtlich und konsistent falsche Antworten gaben (z.B. sagten alle sieben, die Linie A käme X am nächsten). Es zeigte sich, daß viele Versuchspersonen sich in ihrem Urteil der „vorherrschenden" Meinung anpaßten; sei es, weil sie nun tatsächlich anders „wahrnahmen", sei es, weil sie in ihrem Urteil nicht vom Gruppentrend abweichen wollten. Selbst bei scheinbar einfachen und offensichtlichen Beobachtungen kann es also unter bestimmten Randbedingungen zu massiven Verzerrungen und Verfälschungen des letztlich abgegebenen Urteils kommen. Eine zusammenfassende Darstellung mit weiterführenden Literaturhinweisen nicht nur dieses klassischen Experimentes findet sich in Schwartz (1987/88; S. 175ff.).

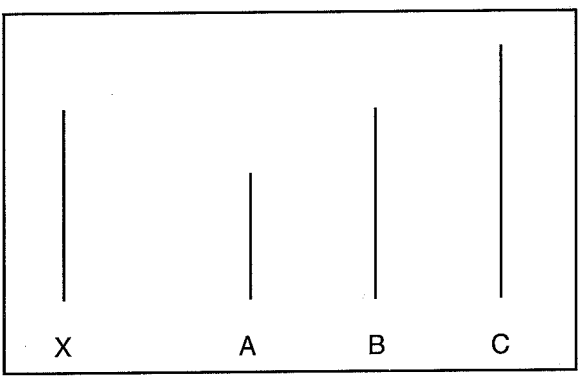


Abbildung 13: Linienvorgaben des Experimentes von Asch

(2) Fehler zu Lasten der Beobachtung

In den folgenden Abschnitten wollen wir vor allem diejenigen Fehler diskutieren, auf die wir bei der Vorstellung der Systematik (Abschnitt 3.2.2) nur kurz hingewiesen hatten, nämlich die Fehler, die durch die Tatsache der Beobachtung auftreten („Fehler zu Lasten interagierender Bedingungen"; vgl. Abb. 10).

20 Diese Möglichkeit ist leider nicht absurd. Es ist offenbar doch nicht so selten, daß Wissenschaftler „das Publikum" bewußt täuschen. Wer sich dafür interessiert, sei auf das Buch von Broad und Wade (1984) hingewiesen.

1. Reaktivität

Mit „Reaktivität" ist allgemein gemeint, daß die Beobachteten sich aufgrund der Beobachtung anders verhalten, als sie es unbeobachtet normalerweise tun würden bzw. getan hätten (zum Überblick vgl. z.B. Mees, 1977a; Cone & Foster, 1982; von Cranach & Frenz, 1969; Pinther, 1972; Schaller, 1980; Wildman & Erickson, 1977; Kent & Foster, 1977; mit etwas anderer Schwerpunktsetzung auch Bungard & Lück, 1974). Das Wissen der Person, beobachtet zu werden, verändert unter Umständen genau das Verhalten, um das es in der Beobachtung geht. Wenn man einen potentiellen Dieb beschattet, wird er, solange er das bemerkt, nicht klauen. Daraus zu schließen, daß er also kein Dieb ist (d.h. niemals klaut oder gar: niemals klauen würde), ist natürlich falsch.

> Auch dazu wieder ein Klassiker der Psychologie: der *Hawthorne-Effekt*. In den 30er Jahren untersuchten Psychologen in einem Werk der Western Electric Company in Hawthorne, Chicago, ob die Beleuchtungsstärke in einer Werkhalle einen Einfluß auf die Produktivität der dort arbeitenden Angestellten hatte. Sie mußten zu ihrer großen Überraschung feststellen, daß sich die Leistungen zwar in der Gruppe unter der Bedingung erhöhter Helligkeit, aber auch in Gruppen mit der Bedingung verminderter Helligkeit steigerten. Erst wenn die Beleuchtung auf Mondhelligkeit verringert wurde, sank die Effizienz der Arbeiter. Die plausible Erklärung: Die bloße Tatsache, daß sich die Arbeiter unter Beobachtung wußten, löste Effekte aus, die alle Effekte durch die experimentellen Manipulationen überdeckten. Eine Schilderung dieser und einer Reihe weiterer in diesem Werk durchgeführte Studien (und mancher Kritik an ihnen) findet sich in Schwartz (1987/88; S.213-226).

Die Reaktivitätseffekte bei Beobachtungsstudien sind oft untersucht worden; die Befundlage ist uneinheitlich (Kent & Foster, 1977). Häufig ist ziemlich unklar, was genau das Verhalten der Beobachteten beeinflußt (hat). In verschiedenen Übersichtsartikeln (z.B. Cone & Foster, 1982; Johnson & Bolstad, 1973) findet sich der Hinweis, daß Erwachsene für derartige Effekte anfälliger sind. Niemand, der schon einmal Kinder selbstversunken einen Turm hat bauen sehen, wird sich darüber wundern. Johnson und Bolstad (1973) diskutieren weitere potentielle Einflußfaktoren: Neben dem Alter oder anderen interindividuellen Unterschieden bei den Beobachteten („Empfindlichkeit", Ablenkbarkeit) könnten hier etwa die „Aufdringlichkeit" der Beobachtung (teilnehmend vs. nicht-teilnehmend; offen vs. verdeckt; etc.) oder persönliche Eigenschaften und Attribute des Beobachters (Hautfarbe, Geschlecht, äußere Erscheinung etc.) eine Rolle spielen. Veränderungen im Verhalten der Beobachteten aufgrund dieser „Reaktivität" werden in aller Regel kaum in ihrer Richtung vorhersagbar sein. Es gibt hier allerdings eine wichtige Ausnahme, die wir etwas näher betrachten müssen.

Der wichtigste und vermutlich meistdiskutierte Reaktivitätseffekt ist der bereits angedeutete Erwartungseffekt des Beobachteten (zum Überblick vgl. z.B. Kazdin, 1977; Mees, 1977a; Pinther, 1972; Cone & Foster, 1982; Kent & Foster, 1977; Johnson & Bolstad, 1973). Damit ist der Effekt gemeint, daß die Beobachteten sich in einer Weise verhalten, die (ihrer Vermutung nach) der Beobachter erwartet (Kriz, 1979, nennt das die Erwartungserwartung: die Erwartung

des Beobachteten über die Erwartungen des Beobachters).[21] Im Unterschied zum Erwartungseffekt des Beobachters würde dieses Verhalten auch ein anderer (unbefangener) Beobachter sehen. Dieser Erwartungseffekt ist in der Tat ein paradigmatischer Fall von Reaktivität: Der Beobachtete reagiert in bestimmter Weise auf die Beobachtung. Ein klassisches Beispiel, die Untersuchung des „klugen Hans", soll auch diesen Punkt verdeutlichen.

Diese Studie des Psychologen Pfungst (1977, das Jahr der Originalveröffentlichung wird in dieser Neuauflage nicht genannt; eine englische Übersetzung erschien erstmals 1911) ist in zweifacher Hinsicht klassisch. Zum ersten ist sie eine der frühesten Studien zu der Frage, ob Beobachter nicht häufig - wenigstens auch - Erzeuger der Daten sind, die sie lediglich zu registrieren vermeinen. Die Arbeit von Pfungst ist aber vor allem auch deshalb klassisch und lesenswert, weil sie die Gründlichkeit und Systematik einer wissenschaftlichen Untersuchung anhand eines hübschen Falles beispielhaft vorführt. Auch dies also eine Pflichtlektüre für Psychologen!

Der „kluge Hans" war ein Pferd, das nach fester Überzeugung seines Besitzers, Herrn van Osten, und auch allem äußeren Anschein nach rechnen, buchstabieren und auf verschiedene Fragen inhaltlich korrekt antworten konnte. Es übermittelte seine Antwort dabei durch Klopfen mit dem rechten Huf, indem es bei nicht numerischen Antworten einen festgelegten Code verwendete. Diese Antworten gab Hans auch dann noch erstaunlich korrekt, wenn jemand anderer als Herr van Osten ihn befragte. Hans hatte jedoch in Wahrheit, so stellte Pfungst bei seiner gründlichen Analyse fest, nur äußerst sensibel auf verschiedene Hinweise seitens der Fragesteller zu reagieren gelernt. Er begann mit dem Klopfen, sobald er das Stellen einer Frage am Tonfall „erkannte", und unterbrach es sofort, wenn er ein „OK-Signal" des Fragenden registrierte (z.B. ein unbeabsichtigtes und unbewußtes leichtes Heben des Kopfes, feinste Variationen der Mimik etc.), der die korrekte Antwort ja kannte. Die Anzahl richtiger „Antworten" sank z.B. dann auf Zufallsniveau, wenn er den Fragenden nicht sehen konnte.

Der oben geschilderte „Hawthorne"-Effekt ist vermutlich häufig ebenfalls ein Erwartungseffekt des Beobachteten. Die beobachteten Arbeiter eines Werkes könnten etwa vermuten, die Beobachtung diene der Feststellung ihrer Leistungsfähigkeit und -bereitschaft, von der wiederum ihre Weiterbeschäftigung abhänge. Auf die Untersuchung von Rosenthal und Jacobson (1968) haben wir bereits mehrfach hingewiesen. Die methodische Konsequenz, die aus dem Phänomen der Erwartungseffekte zu ziehen ist, liegt auf der Hand. Der Beobachtete muß „blind" gegenüber den Erwartungen des Untersuchungsleiters sein, und der Beobachter (oder Experimentator) muß „blind" gegenüber der konkreten Untersuchungsbedingung sein.

Wie kann das aussehen? Das bekannteste Beispiel hierfür sind Untersuchungen, die die Wirksamkeit von Medikamenten mit sogenannten „Placebos" vergleichen, deren Wirkung allein auf der Überzeugung des Patienten beruht, daß er ein wirksames Medikament einnimmt. Diese Studien werden im „Doppel-Blind"-Verfahren durchgeführt. Allen Patienten wird mitgeteilt, sie bekämen ein hochwirksames Medikament verabreicht, und allen Ärzten wird mitgeteilt, sie verabreichten dieses Medikament. Tatsächlich verabreicht aber ein Teil der Ärzte (ohne daß sie oder ihre Patienten das wissen) nur Zuckerpillen. Wenn sich bei die-

[21] Diese „zweite Art" des Erwartungseffektes kann den ersteren zur Folge haben, und zwar dann, wenn *die Beobachter selbst* sich durch den Untersuchungsleiter beobachtet fühlen bzw. bestimmte Erwartungen bei ihm wahrnehmen oder unterstellen. Auf diesen Punkt weisen z.B. Johnson und Bolstad (1973) hin.

sen Patienten die gleiche Besserung einstellt wie bei den anderen, ist offenbar ein „Placebo-Effekt" eingetreten.

Unter der Überschrift „Reaktivität" wird häufig nur ein Spezialproblem dieses allgemeinen Phänomens diskutiert: Das Problem nämlich, daß auch Beobachter sich anders verhalten, wenn sie beim Beobachten beobachtet werden (Überblick dazu: Kazdin, 1977; Cone & Foster, 1982; Kent & Foster, 1977; Johnson & Bolstad, 1973).[22] Sie arbeiten dann genauer und sorgfältiger, als wenn sie sich unkontrolliert fühlen. Dieser Effekt erhöht sich, wenn der Versuchsleiter („Supervisor") den Beobachtern persönlich bekannt ist (Kent & Foster, 1977). Das leuchtet unmittelbar ein: Wir alle sind etwas fleißiger, wenn der Chef in der Nähe ist, unsere Einstellung ist seiner ein wenig näher, wenn er uns selber fragt oder dabei ist, wenn wir sie äußern, und wir sind genauer, wenn er uns zusieht, als wenn das, was wir tun, nicht überprüft werden kann (Kent & Foster, 1977, S. 306). Die gängigste Empfehlung hierzu lautet: Öfter „blind" checken (ohne daß die Beobachter das merken), z.B. indem man eine Videokamera installiert, die die Beobachter (und die Beobachteten) von Zeit zu Zeit aufnimmt. Wenn man dies außerdem ankündigt, ohne den Zeitpunkt zu nennen, hat das den Effekt, daß die Beobachter sich durchgängig beobachtet und kontrolliert fühlen, was im allgemeinen ihre Aufmerksamkeit, Genauigkeit und ihre Übereinstimmung erhöht (Cone & Foster, 1982; Kent & Foster, 1977). Eine Gefahr besteht allerdings darin, daß eine zu strikte Überprüfung manchmal auch das Gegenteil bewirkt: Die so gemaßregelten und geknechteten Beobachter verweigern dann die Kooperation oder verlieren einfach die Lust.

Wie ernst ist die Gefahr der Verzerrung durch Reaktivitätseffekte bei Beobachtungsstudien zu nehmen? Kent und Foster (1977, S. 288) weisen in ihrem Überblick darauf hin, daß zwar verschiedene Studien (z.B. Mercatoris & Craighead, 1974) derartige Reaktivitätseffekte gefunden haben, diese aber häufig nur einige wenige Verhaltensbereiche betrafen. Die Konsequenz, teilnehmende Beobachter in der Hoffnung einzusetzen, bei ihnen seien entsprechende Effekte geringer, werde dennoch häufig gezogen. Jedoch fanden Hay, Nelson und Hay (1980) in einer Studie mit Lehrern und Studenten, daß es auch bei teilnehmender Beobachtung Reaktivitätseffekte gibt; in ihrer Studie betraf dies zudem nicht nur das Verhalten der beobachteten Studenten sondern auch das der beobachtenden Lehrer (vgl. zu diesem Punkt auch den Abschnitt 1.5 zur teilnehmenden Beobachtung).

Reaktivität ist im übrigen für sich genommen kein hinreichender Einwand gegen die Durchführung von Beobachtungen. Cone und Foster (1982) argumentieren, daß die Reaktivität ein Maß nicht automatisch nutzlos mache: Auch der Intelligenztest sei ein in diesem Sinne reaktives Maß, was für seine Nützlichkeit und (unter bestimmten Bedingungen) Aussagekraft zunächst nichts besagt (vgl. auch Foster & Cone, 1980). Einschränkend muß man hier natürlich einwenden, daß es wesentlich darauf ankommt, zu welchem Zweck man beobachtet: Will man Vorher-

22 Man könnte hier von Beobachtereffekten „1. Ordnung" (Reaktion der Beobachteten auf die Tatsache der Beobachtung) und „2. Ordnung" (Reaktion der Beobachter auf das Beobachtet-werden beim Beobachten) sprechen.

sagen treffen (von Verhalten unter bestimmten Bedingungen), oder will man Informationen über bestimmte (normale) Verhaltensweisen sammeln? Kurz: Es ist in jedem konkreten Fall zu bedenken und ggf. empirisch zu prüfen, ob die Beobachtung mit dem gesuchten Verhalten interagieren könnte. Die Antwort wird freilich nicht immer so offensichtlich auf der Hand liegen wie bei der Beobachtung eines Diebes (vgl. Abschnitt 3.2.3).

Eine Strategie, mit der Reaktivitätseffekte üblicherweise untersucht werden (zum Überblick: Johnson & Bolstad, 1973; Baum, Forehand & Zegiob, 1979; Dubey, Kent, O'Leary, Broderick & O'Leary, 1977), besteht neben dem Vergleich verschieden „aufdringlicher" Beobachtungen (z.B. Mercatoris & Craighead, 1974; Weinrott, Garrett & Todd, 1978) in der Prüfung der Stabilität der beobachteten Verhaltensweisen über die Dauer der Beobachtung hinweg. Wenn sich sonst nichts ändert, sollte, so wird unterstellt, eine allmähliche Gewöhnung an die Beobachtung eintreten, und – wenn es einen Reaktivitätseffekt gegeben hat – nach bzw. wegen dessen Nachlassen eine Veränderung des zu beobachtenden Verhaltens auftreten (vgl. z.B. Masling & Stern, 1969). Unglücklicherweise kann aber zum einen nicht ohne weiteres sichergestellt werden, daß diese Veränderungen tatsächlich nur auf die Gewöhnung an den Beobachter (an das Beobachtet-werden) zurückgehen und nicht etwa auf nicht-reaktive, *tatsächliche* Veränderungen (vgl. auch Cone & Foster, 1982). Zum anderen aber entsteht bei derartigen Untersuchungen folgendes Dilemma: Einerseits ist dieses Untersuchungsparadigma überhaupt nur unter dieser Voraussetzung (der Gewöhnung) sinnvoll, andererseits ist es unter dieser Voraussetzung beinahe uninteressant, diesen Effekt zu untersuchen. Johnson und Bolstad (1973, S. 39) zitieren beispielsweise die Einschätzung von Barker und Wright (1971), die annehmen, daß Reaktivitätseffekte relativ kurzlebig sind und daß sich etwa Familienmitglieder schnell an die Anwesenheit von Beobachtern gewöhnt haben. In ähnlicher Weise habe sich auch Bales (1950a) geäußert (vgl. dazu auch von Cranach & Frenz, 1969, S. 308). Boice (1983) zieht daraus nur die Konsequenz, wenn er darauf hinweist, daß Reaktivitätseffekte, wenn sie denn vorliegen, oft nur das Ergebnis einer mangelnden oder zu wenig sorgfältigen Vorbereitung seien. Man könne die Beobachtungs„objekte" an das Beobachtet-werden eben auch gewöhnen; nur müsse man das eben rechtzeitig und lange genug vorher tun. In der Biologie und Verhaltensforschung sei das völlig üblich. Unangenehmerweise taucht hier jedoch das Problem auf, woran man denn nun erkennen kann, daß die Gewöhnung eingetreten (d.h.: die Reaktivität abgeklungen) ist. Offenbar muß man dazu das haben, was man ja eigentlich erst sucht: das Wissen über das normale (unbeobachtete) Verhalten der Beobachtungsobjekte. Die Konsequenz ist wiederum: Wo immer ethisch vertretbar, muß man also verdeckt beobachten.

2. Beobachtungs- und Untersuchungsbedingungen

Diese Fehlerquelle erscheint relativ trivial; wir können uns daher darauf beschränken, sie nur ganz kurz anhand eines fiktiven Beispiels zu behandeln. Wenn wir etwa das Verhalten eines Elternpaares ihrem Kind gegenüber untersuchen wollen, das als auffällig zur Therapie angemeldet wurde, werden wir vielleicht auch eine typische Interaktion zwischen den drei beteiligten Per-

sonen beobachten wollen. Wenn wir dazu die Familie ins psychologische Labor bitten und sie auffordern, beispielsweise ein Gesellschaftsspiel zu spielen, ist damit zu rechnen, daß typische Kommunikationsmuster schon aufgrund der äußeren Bedingungen schwächer oder gar nicht auftreten. Es fehlt der zu Hause stets laut laufende Fernsehapparat, das schreiende Kleinkind, die an die Wände klopfenden Nachbarn, die räumliche Enge des heimischen Wohnzimmers, der Zeitdruck, unter dem der Vater normalerweise steht etc.

3. Probleme des Beobachtungssystems

Die Fehler, die durch unangemessene Gestaltung des der Beobachtung zugrundegelegten Beobachtungssystems entstehen, können erst dann ausführlicher diskutiert werden, wenn wir etwas mehr über die verschiedenen Systeme und ihre Besonderheiten gesagt haben (siehe Abschnitt 4.1). Wir stellen diese Diskussion bis dahin zurück.

3.3 Lösungen: Training, Beobachtungssysteme, Kontrolle

Anscheinend stehen wir nun also vor einem Dilemma: Einerseits benötigen wir uns selber als „Meßinstrument", andererseits müssen wir anerkennen, daß die Zuordnungen, die dieses Instrument trifft, offenbar nicht immer korrekt sind. Sollen wir also das ganze Unternehmen aufgeben? Das ist offenbar nicht nötig; man kann diesen Fehlern vorbeugen, und man kann, wenn sie schon nicht zu verhindern sind bzw. waren, versuchen herauszufinden, ob und inwieweit sie in der konkreten Beobachtung aufgetreten sind. Die Vorbeugung betrifft im wesentlichen zwei Bereiche. Der wichtigste ist die Planung und Durchführung der Beobachtung, insbesondere bezüglich des der Beobachtung zugrundeliegenden Verfahrens: des Systems nach dem bzw. mit dessen Hilfe beobachtet wird. Wir werden diese Systeme in den folgenden Kapiteln dem Prinzip nach und anhand von Beispielen vorstellen. Ein weiterer in diesem Zusammenhang möglicher und vermeidbarer Fehler ist bereits angesprochen worden: die Wahl der Stichprobe (Abschnitt 3.2.2). Ein sehr wichtiger, häufig angesprochener, aber eher selten ausführlich diskutierter Punkt ist die Auswahl und das Training der eingesetzten Beobachter.

Der zweite Weg, mit den Fehlerquellen umzugehen, ist der Weg der Kontrolle. Auch hier sind im Prinzip zwei Strategien zu unterscheiden. Der übliche Weg, den man dabei beschreitet, ist - auch dies haben wir bereits erläutert - ein Vergleich der Ergebnisse verschiedener Beobachter, d.h. die Prüfung, inwieweit sie übereinstimmen. Wir werden die wichtigsten Maße in den folgenden Kapiteln (Abschnitte 4.5, 5.5 und Anhang) einführen und erläutern. Da diese Berechnung natürlich quantitative Daten („Zahlen") voraussetzt, entfällt eine entsprechende Debatte für die Protokollierung der Beobachtung mit Hilfe der Alltagssprache. Zwar können diese verbalen Protokolle selbstverständlich weiterverarbeitet und dabei auch quantifiziert werden, so daß die so erzeugten quantitativen Daten dann auch entsprechenden Berechnungen unterzogen werden können. Aber erstens handelt es sich dann um Daten, die hinsichtlich des

Grades der Reduktion in eines der folgenden Kapitel gehören, und zweitens ist - wenn man mit diesen Daten dann Übereinstimmungsmaße berechnete - nicht mehr sicher zu entscheiden, ob sich vorhandene Unterschiede und Diskrepanzen bereits bei der Beobachtung oder erst bei der Weiterverarbeitung eingeschlichen haben.

Daneben können und sollten natürlich die Beobachter bei der Arbeit kontrolliert werden. Wir haben oben gesehen, daß die Reliabilität allein dadurch ansteigen kann. Die einzelnen Strategien (Stichwort z.B.: „Eichbeobachter") sollen hier nicht nochmals diskutiert werden, auch nicht mögliche unerwünschte Effekte der Beobachtung der Beobachter (Stichworte: Reaktivität, Erwartungseffekte). Es ist freilich nochmals vor dem Effekt zu warnen, daß die Motivation der Beobachter durch zu strenge Kontrollen so stark leiden kann, daß ein gegenteiliger Effekt eintritt.

3.3.1 Vorbeugung

„Je weniger Kategorien, je präziser definiert, je weniger zur Klassifikation an Schlußfolgerung und Interpretation notwendig, desto größer wird die Reliabilität der Daten sein" (Gellert, 1955, S. 184). Neben dieser generellen Empfehlung und dem obligatorischen Hinweis, Beobachtungen so unauffällig wie eben möglich durchzuführen, finden sich in der Literatur (vgl. z.B. Kent & Foster, 1977; Boice, 1983; Manns et al., 1987) eine Reihe weiterer konkreter Hinweise, um Beobachtungsfehler von vorneherein zu vermeiden oder zu verringern. Wir zählen sie einfach kurz auf; ihre Begründung liegt unmittelbar auf der Hand oder ergibt sich aus den oben geführten Diskussionen.

- Erstellen Sie das Beobachtungsschema sorgfältig (theoriegeleitet; Vergleich mit vorliegenden Alternativen; in Vorstudien erprobt; nicht zu umfangreich).
- Trennen Sie die Kategorien so genau und eindeutig wie möglich, kennzeichnen Sie sie mit eindeutigen und vertrauten Begriffen.
- Vorsicht bei der Interaktion von Beobachtern mit dem Versuchsleiter.
- Vorsicht bei der Interaktion von Beobachtern mit den Beobachteten.
- Bereiten Sie die Beobachtung gut vor; gewöhnen Sie, wenn das möglich ist, die Beobachteten an das Beobachtet-werden.
- Beobachten Sie möglichst verdeckt, solange es ethisch vertretbar ist.
- Beobachten Sie möglichst nicht nur unvermittelt.
- Lassen Sie die Beobachter Methode und Beobachtungsobjekte vorher kennenlernen.
- Beobachten Sie geeignete Objekte.
- Lassen Sie den Beobachter möglichst den Handlungskontext des Beobachteten überblicken, damit er Handlungen und Reaktionen adäquat einschätzen kann.
- Lassen Sie dem Beobachter genügend Zeit.

3.3.2 Auswahl und Training der Beobachter

Der Gedanke, es könne „gute" und „schlechte" Beobachter geben, ist - wieder einmal - recht alt. Als Beispiele für besondere „Beobachtungsbegabungen" werden dann Charles Darwin, Sigmund Freud und natürlich der unvermeidliche Aristoteles angeführt, gewissermaßen, wie Boice (1983) es formuliert, als die „Mozarts der Beobachtung". Erstmals systematisch wurde dieses Problem in der vielleicht ältesten Arbeit innerhalb der (akademischen) Psychologie über Fragen der Beobachtung diskutiert und untersucht: in einer Arbeit von Alfred Binet (1897). Binet unterscheidet dort (anhand von Bildbeschreibungen)

- einen *beschreibenden* Typ, der auf auffallende Objekte achtet, nichts über Bedeutungen und Beziehungen sagt und in seinem Bericht keine Phantasie oder Emotionen erkennen läßt,
- einen *beobachtenden* Typ, der Bedeutungen (das *Thema* des Bildes!) und Beziehungen beschreibt, der bewertet und interpretiert[23],
- einen *emotionalen* Typ, der Emotionen zuschreibt, aber im Schnitt weniger Objekte beschreibt,
- einen *gelehrten* Typ, der die Fabel erzählt, anstatt das Bild zu beschreiben, der berichtet, was er über das, was er sieht, denkt, und
- einen *idealistischen*, dichterischen oder imaginativen Typ, bei dessen Berichten Phantasie, persönliche Erinnerungen und emotionale Reaktionen überwiegen.

Das Interesse an diesen Gedanken und Untersuchungen war seinerzeit nach der Einschätzung von Boice (1983) eher gering, nicht zuletzt wohl auch wegen des noch hohen Ansehens der Methode der *Intro*spektion (vgl. Abschnitt 1.2), das die Frage nach der Beobachtung *anderer* Personen (und also auch nach der Qualität dieser Beobachtungen) in den Hintergrund treten ließ. Boice (1983) weist in seinem lesenswerten Überblick aber darauf hin, daß diese Frage dennoch nicht völlig aus den Augen verloren wurde; Allport etwa habe 1925 die Frage aufgeworfen, ob „naive" oder trainierte Beobachter bessere Ergebnisse erbrächten. Diese Frage wird kaum verbindlich zu beantworten sein. Fieguth (1977b) verweist auf Studien (z.B. Patterson & Cobb, 1973), die Laien erfolgreich als Beobachter eingesetzt haben. Einmal mehr wird hier alles von der Abwägung der Erfordernisse der Fragestellung einerseits und der Umstände andererseits abhängen.

Immerhin konnte Taft in seinem Überblicksartikel (1955) bereits auf eine Reihe empirischer Studien zurückgreifen. Er kommt nach Sichtung der bis zu diesem Zeitpunkt vorliegenden Literatur zu dem Ergebnis, daß es eine solche generelle Urteilsfähigkeit gibt. Er diskutiert mehrere mögliche Kennzeichen (beispielsweise Alter, Geschlecht, familiärer Hintergrund, Intelligenz, ästhetische Sensibilität oder Selbstaufmerksamkeit der Beobachter) und kommt zu dem Schluß,

---

23 Interessant und vielleicht nachdenkenswert ist, daß für Binet die Interpretation und Bewertung zu einem (guten) Beobachter gehört.

daß es auf eine hinreichende soziale Distanz zwischen Beobachter und Beobachteten und vor allem auf die *Motivation* des Beobachters zu genauer Beobachtung ankomme.

Der Gedanke, eine Differentialpsychologie des Beobachtens zu treiben, ist offenbar jedoch mittlerweile aus der Mode gekommen. Jedenfalls findet sich in kaum einem der neueren Übersichtsartikel und Handbücher auch nur eine Erwähnung dieses Ansatzes und seiner Möglichkeiten. Diese Ablehnung ist vielleicht etwas vorschnell. Warum sollte es nicht Persönlichkeitsmerkmale geben, aus denen sich aussagefähige Vorhersagen guter Beobachtungsleistungen ableiten lassen? Beispielsweise gibt es zu der in Kapitel 3.2.3 diskutierten Tendenz zur Konsistenz eine lange Forschungstradition (vgl. Schmitt, 1990). Warum sollte es nicht möglich sein, Persönlichkeitseigenschaften, generalisierte Haltungen oder Einstellungen, emotionale Dispositionen oder intellektuelle Fähigkeiten zu identifizieren, die - nach angemessenem Training vielleicht - eine bessere Beobachtungsleistung bei Personen mit hohen Werten in entsprechenden Bereichen erwarten lassen als bei Personen, bei denen diese Voraussetzungen weniger oder gar nicht erfüllt sind? Warum sollten diese allgemeinen Effekte nicht bei verschiedenen Personen in verschiedener Weise wirken? Welche Fertigkeiten werden für welche Aufgabe benötigt? Wie können sie - unabhängig von der jeweils angezielten Beobachtungsaufgabe - erfaßt und eingeschätzt werden? Kann man sie gezielt trainieren und wenn ja: wie? Relativiert sich der häufige Befund, daß Beobachtertraining nicht immer viel hilft (Boice, 1983), wenn man spezifischer trainiert? Boice regt an, hier auch Ergebnisse anderer Forschungsbereiche heranzuziehen. So könnten zum Beispiel Studien darüber, wie Kinder das Zuhören oder genaue Hinsehen lernen, nützliche Hinweise auf die benötigten Fertigkeiten und ihren Erwerb liefern. Welche Möglichkeiten für einen Forscher, der auf der Suche nach spannenden Themen ist!

Die Empfehlung, auf das Beobachtertraining sehr viel Sorgfalt zu legen, findet man in nahezu jeder einschlägigen Arbeit (z.B. Boice, 1983; von Cranach & Frenz, 1969; Grümer, 1974; Hasemann, 1983, S. 472f.; Cone & Foster, 1982; Wildman & Erickson, 1977; Humpert & Dann, 1988). Auch fehlt in kaum einer Studie der Hinweis, die eingesetzten Beobachter seien trainiert worden. Aber man findet nur selten Hinweise darauf, *wie* man denn ein solches Training zu gestalten habe (Kent & Foster, 1977; eine der wenigen Ausnahmen ist Fieguth, 1977b; vgl. auch Boice, 1983). Das ist natürlich insofern nicht unerwartet, als die Inhalte des Trainings ja auch auf den Inhalt der Studie abgestimmt sein müssen. Ein allgemeiner (und wichtiger) Hinweis in diesem Zusammenhang ist sicherlich der, daß die Beobachter an standardisiertem Material üben und lernen sollten. Boice (1983) hat argumentiert, daß nicht nur bestimmte Aufgaben (Schemata, Fragestellungen, Instrumente etc.), sondern auch bestimmte Fähigkeiten (Sensibilität, Gedächtnis, Selbstaufmerksamkeit) berücksichtigt und möglichst trainiert werden könnten und sollten. Er ist der Ansicht, daß unter dieser Voraussetzung „well-trained observers can make a real contribution to research and practice" (S. 23). Es ist beim Stichwort „Training" jedoch auch nochmals an die Gefahr der Beeinflussung *durch* das Training zu erinnern. Dies betrifft etwa die oben diskutierten Erwartungseffekte des Beobachters, aber auch bestimmte Vorurteilsbildungen und Wahrnehmungsgewohnheiten. Fieguth (1977b)

gibt in diesem Zusammenhang die wichtige Empfehlung, daß der Trainer nicht mit dem Forscher identisch sein sollte, um Erwartungseffekte möglichst zu unterbinden.

Taft (1955) diskutiert den Vorteil von Expertise (bei Psychologen) bereits recht ausführlich. Seine Schlußfolgerung ist allerdings eine andere und etwas überraschend: Die Psychologieausbildung (seiner Zeit) hilft offenbar wenig: „There is also evidence that suggests that courses in psychology do not improve ability to judge others and there is considerable doubt whether professional psychologists show better ability to judge than do graduate students in psychology" (S. 12). Taft fragt sich natürlich, warum das so sein könnte. Seiner Vermutung nach sind Psychologen zum einen zu sehr an sozialen Beziehungen interessiert, um gute (objektive) Beurteiler zu sein. Es ist allerdings zu befürchten, daß diese Tendenz heute (in der Wissenschaft ganz sicher) mindestens verschwunden, wenn nicht ins Gegenteil verkehrt worden ist. Zum anderen leben nach Ansicht von Taft Professoren und Kliniker zu isoliert von allgemeinen Lebenserfahrungen und von den Personen, die zu beurteilen sie sich bemühen. Auch daran ist selbstverständlich kein Wort mehr wahr: Wissenschaftliche Psychologen leben heutzutage in ganz normalen Elfenbeintürmen und schreiben Bücher über psychologische Beobachtung – wie jeder normale Mensch auch!

Literaturempfehlungen

Zum Thema Beobachtungsfehler gibt es zahlreiche Überblicksartikel. Zu nennen wären hier etwa: Johnson und Bolstad (1973; sehr guter Überblick), Kent und Foster (1977; ebenfalls sehr guter Überblick); Cone und Foster (1982, S. 327ff.; viele Literaturbeispiele, Untersuchungen zu den verschiedenen Einfluß- und Fehlerquellen werden ausführlich geschildert), Zapf (1989, S. 49ff.; recht ausführlicher Überblick mit Literaturhinweisen), Beck (1987; liefert als einer der wenigen einen systematischen Überblick auch in tabellarischer Form: S. 184ff.), von Cranach und Frenz (1969; guter Überblick), Faßnacht (1995; überblick); Martin und Wawrinowski (1991; Überblick); Pinther (1972; für verschiedene Fehler einschlägig), Hasemann (1983; viele Hinweise auch auf ältere Literatur), Glück (1971; gut zur Einführung, nicht erschöpfend), Schaller (1980; einführender aber relativ grober Überblick, mit Literaturhinweisen), Kazdin (1977; zur einführenden Lektüre nicht ungeeignet, diskutiert aber nur einige Fehler, z.T. überholt), Kriz (1979; Thema „Artefakte"), Bungard und Lück (1974; berichten über Forschungsartefakte zwar aus Sicht von Experimenten, die Probleme sind aber entweder prinzipiell dieselben oder leicht übertragbar).

Kapitel 4

Die explizite Reduktion des Wahrgenommenen

In diesem Kapitel werden wir beginnen, Beobachtungssysteme vorzustellen, die kontrollierbarere Ergebnisse erlauben: In der Regel wird in ihnen deutlich weniger als in Verbalsystemen (Kapitel 2) auf das Vorverständnis des Beobachters zurückgegriffen, vor allem, weil hier die Zuordnungsvorschriften expliziter sind. Das führt zu Überlegungen über die Einheit, die wir unserer Beobachtung zugrunde legen wollen (Abschnitt 4.2), zur Art der Zuordnungsaufgabe, die dem Beobachter aufgegeben wird (Abschnitt 4.3), zum Problem des Zusammenhangs von theoretischen Konstrukten und den Beobachtungszeichen (Abschnitt 4.4) und schließlich zur Diskussion von Methoden zur Schätzung der Beobachterübereinstimmung (Abschnitt 4.5).

4.1 Einführung der Zeichensysteme

Wenn in der Überschrift dieses Kapitels von „Reduktion des Wahrgenommenen" die Rede ist, so müssen wir gleich einem Mißverständnis zuvorkommen: Natürlich bedeutet auch die Verwendung von Verbalsystemen Reduktion des (prinzipiell) Vorfindbaren auf das, was der Beobachter für wert hält, festgehalten zu werden (ganz abgesehen von den Einschränkungen, denen das „Meßinstrument" Beobachter unterliegt; siehe Kap. 3). Trotzdem besteht ein Unterschied zu den folgenden Systemen: Hinter der Nutzung von Verbalsystemen steht in der Regel der Anspruch (wie man es exemplarisch sehr schön am Beispiel von Barker und Wright sehen kann), zunächst möglichst unverfälscht (und eben möglichst unreduziert) die erklärungsbedürftigen Phänomene der Wissenschaft anzugeben; d.h. (etwas überspitzt) die Ergebnisse solcher Studien liefern erst das Material zur Konstruktion von Theorien.

In der Regel wird jedoch umgekehrt gearbeitet: Beobachtung wird eingesetzt, um Hypothesen zu testen (vgl. Abschnitt 1.4.2).

> Lovaas, Koegel, Simmons und Lang (1973; vgl. dazu auch Faßnacht, 1995, S. 180f.) entwickelten vor dem Hintergrund theoretischer Vorstellungen eine Therapie für autistische Kinder. Kindlicher Autismus ist u. a. durch „eine schwere zwischenmenschliche Kontaktstörung" (Nissen, 1989, S. 518) gekennzeichnet. Insofern erwarteten die Autoren, daß nach Abschluß der Therapie mehr „soziales Verhalten" von den Kindern gezeigt wird.
>
> Es dürfte klar sein, daß einerseits ein knapper „Über-Alles-Eindruck" des Behandlungseffektes kein geeignetes Mittel zur Beurteilung eines Heilverfahrens sein kann, aber andererseits die ausführliche Beschreibung des Verhaltens durch den Einsatz von Verbalsystemen

kaum eine deutliche, knappe Aussage erbringen wird. Was liegt nun näher, als einige Verhaltensklassen zu definieren, die entweder als symptomatisch oder als Zeichen für eine Besserung gelten, um dann zu schauen, ob sich in einer Gruppe von therapierten Kindern eine Verlagerung (gegenüber dem Zeitpunkt vor der Therapie) hin zu den „Besserungszeichen" ergeben hat, diese jedoch bei nicht-therapierten Kindern während des gleichen Zeitraumes ausblieb? Lovaas et al. (1973) gingen genau so vor und definierten damit ein sogenanntes Zeichensystem. Sie konnten nun Beobachtern die Aufgabe übertragen, während eines festgelegten Zeitraumes die definierten Verhaltensweisen zu registrieren: Ein Vergleich von Erwartung und realem Ergebnis war möglich.

Allgemein sprechen wir von einem *Zeichensystem*, wenn wir Verhaltensklassen definiert haben, in die wir konkret vorkommendes Verhalten einordnen können, und wenn wir das Vorkommen eines konkreten Verhaltens notieren, indem wir das für seine zugehörige Klasse vorgesehene Zeichen geben. Dabei machen wir zunächst keine Annahmen über die logische Struktur dieser Verhaltensklassen; es interessiert uns dabei nicht, ob die einzelnen Klassen und damit ihre Zeichen sich gegenseitig ausschließen oder ob etwa eine Verhaltensweise in mehrere Klassen „paßt". Wir machen außerdem keine Annahme darüber, ob die Liste vollständig ist, d.h. ob *jede* Verhaltensweise einem Zeichen zuzuordnen ist. Die Bedingungen für ein Zeichensystem in diesem technischen Sinne sind also wenig streng. Damit ist noch nichts darüber gesagt, ob dieses Zeichensystem aus *diagnostisch relevanten* Verhaltensklassen besteht, also aus Verhaltensklassen, die „Zeichen *für* etwas" sind. In diesem diagnostischen Sinn ist etwa „Echolalie" ein Zeichen für Autismus.

Lovaas et al. (1973) wählten fünf Zeichen, von denen zwei für symptomatisches (autistisches) Verhalten standen und drei als „Besserungszeichen" interpretiert wurden. Die *Bezeichnung* der fünf Zeichen sei hier in der Übersetzung von Faßnacht (1995, S. 181) wiedergegeben:

1. Selbst-Stimulation
2. Echolalie
3. Angepaßte Sprache
4. Soziales nicht-verbales Verhalten
   a) Aufforderndes Verhalten: z.B. das Kind erfaßt die Hand des Erwachsenen und führt ihn zur Türe
   b) Einwilligungen: z.B. das Kind befolgt die Aufforderung „Setze dich hin"
5. Angemessenes Spiel

Die Hervorhebung des Wortes „Bezeichnung" soll andeuten, daß selbstverständlich umfangreiche Definitionen zu den Verhaltensklassen mitgeliefert werden; Interessierte seien auf Lovaas et al. (1973) verwiesen.

An diesem Beispiel wird sehr schön deutlich, was es heißt, einen Kompromiß zwischen den Aussagen einer Theorie über die Zusammenhänge komplexer Begriffe und den Leistungsmöglichkeiten menschlicher Beobachter zu finden. Natürlich ist auch dem Laien klar, daß bei einer Störung, die sich im Fehlen sozialen Verhaltens manifestiert, „soziales Verhalten" des Kindes eine Besserung andeutet. Die Instruktion an einen Beobachter „Notiere immer die Dauer 'sozialer Verhaltensweisen' des Kindes" würde ihn aber wahrscheinlich überfordern und zu nicht re-

produzierbaren Ergebnissen führen. So ist es geboten, diese Verhaltensklasse entweder durch Beispiele (wie sie etwa in der Aufgliederung der Kategorie 4 bei Lovaas et al., 1973, gegeben sind) transparenter zu machen oder - besser noch - das Zeichensystem nicht mit groben Verhaltensklassen wie „soziales Verhalten", sondern mit *Indikatoren* zu definieren, so daß in einem zweiten Schritt durch Zusammenfassung der Ergebnisse für die Indikatoren auf die übergeordnete Klasse geschlossen werden kann: Wie gesagt, definierten Lovaas et al. (1973) fünf Zeichen, von denen jeweils zwei bzw. drei Zeichen später zusammengefaßt wurden. *Alles, was wir nicht explizit festlegen, überlassen wir dem Vorverständnis und der Intuition des Beobachters!* Es bleibt noch einmal darauf hinzuweisen, daß der expliziten Festlegung Grenzen gesetzt sind, wie wir an der Diskussion über die Wahl der richtigen Beschreibungssprache gesehen haben (Stichwort: Aktone vs. Aktionen bei Barker und Wright; vgl. Abschnitt 2.1 und Abschnitt 2.3).

Wir können nun den im vorigen Kapitel (Abschnitt 3.2.3) zurückgestellten Diskussionspunkt der „Fehlerquelle Beobachtungssystem" nachholen. Ein wichtiger und viel diskutierter Punkt ist in diesem Zusammenhang vor allem die *Komplexität* des Beobachtungssystems, d.h. die Zahl der verwendeten Klassen.[24] Faustregel: Je komplexer die Beobachtung (in diesem Sinne), desto unreliabler (Kazdin, 1977; Cone & Foster, 1982; Kent & Foster, 1977; Beck, 1987). Aber selbst wenn man dem hierbei zugrundeliegenden Komplexitätsverständnis folgt, erscheint die Faustregel zu plump. Ein wichtiger Faktor wird hier die *Struktur* des Beobachtungssystems sein. Eine Vielzahl unverbundener Zeichen (z.B.: Gesten mit dem Fuß, mit der Hand, mit den Armen, mit den Beinen, mit der Oberkörperhaltung, ...) ist sicher schwieriger zu verwenden und daher fehleranfälliger als eine (u.U. größere) Zahl hierarchisch und systematisch gut gegliederter Zeichen (z.B.: *1. Ebene („Körper")*: Kopf, Fuß, Hand, ...; *2. Ebene („Kopf")*: Auge, Nase, Mund, ...; *3. Ebene („Auge")*: Lider, Brauen, Augapfel, ...; etc.). Es ist jedenfalls fraglich, ob es Sinn macht, hier konkrete, d.h. zahlenmäßig fixierte Empfehlungen zu geben (z.B. nennt Fieguth, 1977a, eine Zahl von 31 Klassen). Selbstverständlich muß die Theorie Vorrang haben: Wenn aus ihr stringent 32 Klassen abgeleitet werden können, sollten derartige Empfehlungen niemanden verunsichern oder gar abhalten (sollte es in der Psychologie gelegentlich derart elaborierte Theorien geben, ist das ohnehin reichlich Anlaß zur Freude). Kurz: Eine große Klassenzahl ist, für sich allein betrachtet, sicher noch kein (Kunst-) Fehler.

---

[24] Man könnte sich auch andere Definitionen denken: Wieviel muß pro Beobachtungseinheit integriert werden? Ein Zeichensystem mit zahlreichen, aber jeweils sehr einfachen Verhaltensklassen (etwa zur Gestik; ob sich die Hand, der Fuß, das Bein etc. bewegt, ist nicht schwer zu unterscheiden) könnte in diesem Sinne weniger komplex sein als ein System mit relativ wenigen, aber hoch abstrakten Verhaltensklassen (z.B. sozial unterstützendes vs. abweisendes Verhalten, o.ä.; hier ist jedesmal die schwierige - „komplexe" - Entscheidung zu fällen, ob das beobachtete Verhalten ein Fall von ... ist.)

4.2 Die Segmentierung des Geschehens: formale versus semantische Einheiten

Das erste Problem, das sich uns bei der expliziten Festlegung eines Zeichensystems stellt, ist die Gliederung des Geschehens in sinnvolle Einheiten: Dem Beobachter muß zu jedem Zeitpunkt klar sein, wie umfassend der Verhaltensausschnitt ist, den er hinsichtlich vorgegebener Kategorien einordnen soll.

> Auch Barker und Wright (Abschnitt 2.1) hatten natürlich dieses Problem. Genaugenommen trat es bei ihnen gleich zweimal auf: Explizit wurde es bei den Anweisungen zur „Episodierung"; eher implizit blieb es bei der Wahl der angemessenen Beobachtungssprache, denn je „gröber" meine Sprache das Geschehen beschreibt, desto „gröber" ist natürlich auch seine Zergliederung (vgl. „er ging zur Schule" mit „er ging über die Straße, bog nach rechts in die Allee, ..." mit „er schaute nach links, nach rechts und noch einmal nach links, trat dann zügig vom Bordstein ...").

Welche Arten der Segmentierung bieten sich an? Ein Blick in die Literatur zeigt, daß die Möglichkeiten der Einheitenbildung sehr verschieden diskutiert werden (vgl. Faßnacht, 1995; Fieguth, 1977a; Kalbermatten & von Cranach, 1981; Manns et al., 1987); dies scheint zum Teil an verschiedenen theoretischen Ausrichtungen der Autoren zu liegen. Prinzipiell steht man hier vor der Alternative, entweder nach rein *formalen* Kriterien, die sich unabhängig von der inhaltlichen Fragestellung fassen lassen, oder nach *semantischen*, d.h. psychologisch-inhaltlichen Gesichtspunkten zu segmentieren.

(1) Formale Einheitenbildung

Es gibt vor allen Dingen *eine* Einheit, die sich in der geforderten Weise unabhängig von der inhaltlichen Fragestellung festlegen läßt: das Zeitintervall. In diesem Fall legen wir einfach ein gleichmäßiges Zeitraster über die gesamte Beobachtungsspanne und weisen die Beobachter an, das Verhalten in jedem Zeitintervall mit den vereinbarten Zeichen zu klassifizieren. Natürlich spielen inhaltliche Annahmen über das zu Beobachtende bei der Festlegung der Ausdehnung des Einheitsintervalls eine Rolle. Das Segmentieren in Zeitintervalle wollen wir hier „Zeittakt-Methode" nennen. In ihrer Übersicht über „Beobachtungsverfahren in der Verhaltensdiagnostik" stellen Manns et al. (1987) 26 Beobachtungssysteme unterschiedlicher inhaltlicher Bereiche vor; 19 dieser Verfahren arbeiten nach dem Zeittaktverfahren mit Zeitintervallen von 1 bis 30 Sekunden. Warum wird dieses Vorgehen so häufig gewählt? Allgemein kann man sagen, daß sich jede Notierung eines Beobachters zeitlich genau lokalisieren läßt, ohne daß er zusätzlich zu seiner Zeichenwahl eine Uhrzeit notieren müßte. Dies ist ein wichtiger Punkt für Untersuchungen zur Zuverlässigkeit eines Zeichensystems, da wir dafür wissen müssen, ob *dasselbe* Verhalten von zwei oder mehr Beobachtern identisch codiert wurde. Wir werden später ausführlich auf dieses Thema eingehen (Abschnitt 4.5). Zum anderen ergibt sich durch das Zeittakt-Verfahren (bei geeigneter kleiner Wahl des Intervalls) eine gute Schätzung nicht nur der

Häufigkeit einer Verhaltensweise, sondern auch der mittleren Dauer und der relativen Aufteilung des Gesamtgeschehens in die den Zeichen zugeordneten Verhaltensweisen.

Das praktische Vorgehen ist denkbar einfach: Man benutzt einen visuellen oder akustischen Taktgeber und codiert bei jedem „Aufblinken" oder „Piepsen" das im Intervall gezeigte Verhalten. Bei vermittelter Beobachtung (etwa der Auswertung eines Videobandes) kann das Signal ganz offen gegeben werden, bei direkter Beobachtung wird man ein akustisches Signal auf einen Ohrhörer geben.

Es lassen sich jedoch noch andere formale Einheiten denken. Betrachten wir etwa folgende Studie von Isen und Levin (1972, Experiment 2).

> Unter welchen Umständen sind Menschen bereit, anderen Personen Hilfe zu leisten? Dieses Thema wird eingehend in der Sozialpsychologie untersucht (zum Überblick etwa Bierhoff, 1984). Isen und Levin (1972) widmeten sich insbesondere dem Einfluß der „Stimmung" auf die Hilfsbereitschaft. In einem kleinen Experiment hinterlegten die Autoren Geldstücke im Rückgabeschacht einer Telefonzelle. Versuchspersonen, die das Geld fanden, sollten dadurch in eine (leicht) bessere Stimmung versetzt werden als Personen einer Kontrollgruppe, für die keine Münze bereitgelegt wurde. Sobald die Versuchsperson die Telefonzelle verließ, lief ihr ein Mitarbeiter der Experimentatoren über den Weg und ließ „unabsichtlich" einen Ordner mit Papieren fallen. Beobachtet wurde, ob die Versuchsperson Hilfe leistete. Tatsächlich war ein deutlicher Effekt festzustellen: Während kaum eine Person der Kontrollgruppe half, bemühten sich fast alle Personen, die das „Geldgeschenk" erhalten hatten, die Papiere einzusammeln.

In dieser Studie ist schlicht die Person die Beobachtungseinheit: Genau wie beim einzelnen Zeitintervall versucht der Beobachter, das gezeigte Verhalten *für jede Person* dem Zeichensystem zuzuordnen; das heißt in diesem Fall: Hat diese Person geholfen oder nicht?

Aber es gibt noch andere Möglichkeiten. Vor allem sind zwei Arten der Einheitenbildung zu nennen, die aber hinsichtlich ihres Anwendungsbereiches deutlich hinter der „Zeittakt-" und der „Person-als-Einheit"-Methode zurückstehen. Zum einen lassen sich Bewegungen als Veränderungen in Raum und Zeit anführen. Diese Art der Segmentierung ist jedoch offensichtlich nicht so universell anwendbar wie die „Zeittakt-Methode": Während letztere in ihrer Ausdehnung frei den Bedürfnissen der Fragestellung angepaßt werden kann, ist jene auf eine „mikroskopische" Ebene fixiert. Außerdem wissen wir ja, daß für viele Fragestellungen, bei denen die Zeichen eine Klassifizierung von Handlungen ermöglichen sollen, die Ebene der Bewegungsmuster uninteressant ist (vgl. Kap. 2; s. aber als Beispiel die strukturelle Kategorienbildung bei Kalbermatten & von Cranach, 1981). Zum anderen kann man bei allen Fragestellungen im Bereich verbaler Kommunikation auf syntaktische Formen zurückgreifen: Man klassifiziert dann etwa jeden Satz oder jede Silbe (Sprachstörungen!). Als Beispiel mag man sich die Kontrolle der Häufigkeit des Stotterns vorstellen: Einheit wäre die Silbe, für die festgestellt wird, ob sie korrekt ausgesprochen, tonisch (das „Nicht-heraus-bringen-können" von Wörtern) oder klonisch (das Wiederholen von Einzellauten oder Silben) gestottert wird (zur Terminologie vgl. Strunk, 1989). Aber auch hier wird häufig die inhaltlich bedeutungsvolle Äußerung als Einheit genommen.

(2) Semantische oder psychologische Einheitenbildung

In der Regel wird diese Art der Einheitenbildung unter dem Stichwort „natürliche Einheiten" diskutiert. Gemeint ist die im Zusammenhang der Verbalsysteme eingeführte „Handlung" als Basis der Segmentierung.[25] Jede Handlungs-Einheit kann dann wiederum in das Beobachtungssystem eingeordnet werden.

> In den Arbeiten der Forschergruppe um von Cranach (von Cranach, Kalbermatten, Indermühle & Gugler, 1980; Kalbermatten & von Cranach, 1981) kann man dieses Vorgehen nachvollziehen: Die Autoren bemühten sich, ein Geschehen, welches sich als quasi natürliche Gestalt aus dem Gesamtablauf ausgrenzen läßt („Zielebene"), in seine funktionalen Bestandteile zu zerlegen („funktionale Ebene") und wiederum deren strukturelle Merkmale festzuhalten („strukturelle Ebene"). Beobachtet wurden Kindergartenkinder, deren Spiel mit Videokameras aufgezeichnet wurde. Die Arbeitsgruppe wählte als Geschehenstypus den Konflikt zweier Kinder um ein Objekt. So stellte sich als erste Frage, ob Beobachter, denen die Aufgabe der Beschreibung des Geschehens und damit der Auftrag zu einer Segmentierung gegeben wurde, genau diese Konfliktsituationen als eine Einheit herausgriffen. Dies wurde mit hoher Übereinstimmung getan.
>
> Darüber hinaus wurden diese Konfliktsituationen in ihre funktionalen Bestandteile zerlegt. Die Situation läßt sich durch die Ziele der beiden beteiligten Kinder charakterisieren: eines möchte ein Objekt „erobern", das andere es verteidigen. Funktionale Bestandteile sind im Sinne von Kalbermatten und von Cranach die Einzelschritte, die der Erreichung des Zieles untergeordnet sind. Dazu gehören Bewegungsmuster wie „Annäherung an das Objekt", aber auch untergeordnete Handlungen wie „Drohen" oder eine „Rechtfertigung" aussprechen. Der Katalog der funktionalen Klassen bei Kalbermatten und von Cranach ist recht umfangreich (22 nicht-verbale und 10 verbale Klassen) und soll hier nicht wiedergegeben werden. Die Autoren betonen, „daß für die Kodierung auf dieser Ebene ein kontextuelles Alltagswissen unumgänglich ist. Der Beobachter muß gewisse soziale Konventionen kennen. Er muß ein 'Kämpfen gegen den Partner' (z.B. Stoßen) von einem freundschaftlichen 'Auf-die-Schulter-Klopfen' unterscheiden können" (Kalbermatten & von Cranach, 1981, S. 106f.). In diesem Sinne liegt auch auf dieser Ebene eine „semantische Segmentierung" des Geschehens vor. Erst wenn die funktionalen Einheiten in ihrer Struktur untersucht werden, wenn also die Einheit „Drohen" in ihre Bewegungsmuster zerlegt wird, haben wir eine formale Segmentierung vor uns. Kalbermatten und von Cranach bezeichnen ihr Vorgehen dementsprechend als „hierarchisch aufgebautes Beobachtungssystem".

Das schon erwähnte System von Lovaas et al. (1973) fällt ebenfalls in die Klasse der Systeme mit „semantischer Segmentierung". Wir werden im Abschnitt 5.3 ausführlich ein weiteres Beobachtungssystem vorstellen, welches mit „semantischer Segmentierung" im Bereich von Gruppenprozessen arbeitet: die Interaktionsprozeßanalyse von Bales (1950a).

Wann wählt man eine solche Segmentierung, und welche Vor- und Nachteile ergeben sich aus diesem Vorgehen? Je nach theoretischer Ausrichtung wird man etwa das „Zeittaktverfahren" als relativ künstlich und dem Untersuchungsgegenstand unangemessen betrachten. Gerade wenn – wie bei von Cranach und Mitarbeitern – die bedeutungsvolle Strukturierung des

---

[25] Zumindest können wir diese Behauptung für den Humanbereich aufstellen. Bei der Tierbeobachtung würde man von teleologischen Kategorien sprechen: Der Affe *flieht*, der Löwe *lauert* etc.

Geschehens Forschungsthema ist, wird man die Aufmerksamkeit des Beobachters voll auf diese Aufgabe konzentrieren und ihm nicht noch die Zeittaktung zumuten wollen. Allerdings muß man kritisch anmerken, daß dem Beobachter eigentlich zwei Aufgaben angetragen werden: Zum einen soll er das Geschehen sinnvoll einteilen, zum anderen muß er diese Einheit im Sinne des Systems codieren. Bei beiden Aufgaben können sich Unzuverlässigkeiten einstellen. Gerade wegen der Fehleranfälligkeit der Beobachtung ist dieses Problem nicht zu leicht zu nehmen. Und hierbei ergibt sich noch ein weiteres Problem: Der Hauptnachweis der Zuverlässigkeit eines Systems wird durch die Berechnung der Übereinstimmung mehrerer Beobachter gegeben. Es wird geprüft, ob ein und dasselbe Geschehen von mehreren Beobachtern übereinstimmend eingeordnet wurde. Diese Berechnung setzt aber voraus, daß die Zeit mitkodiert wird, da wir nur in diesem Fall bestimmen können, wie mehrere Beobachter *zu einem Zeitpunkt* geurteilt haben. Letztendlich muß auf den Protokollbögen genau notiert sein, von wann bis wann welches Zeichenverhalten gezeigt wurde. Welche Hilfsmittel setzt dies voraus? Bei unvermittelter Beobachtung gibt es zwei - letztlich beide nicht befriedigende - Lösungen: (1) der Beobachter notiert die Zeichen entlang einer auf dem Protokollbogen abgedruckten Zeitachse. Dieses Vorgehen setzt voraus, daß der Beobachter stets „mit einem Auge" die Uhr kontrolliert, was natürlich seiner Aufmerksamkeit für das Geschehen abträglich ist. (2) Der Beobachter benutzt ein Gerät, welches automatisch den Zeitpunkt einer Kodierung festhält. Früher war dies der sogenannte „Ereignisschreiber" (vgl. z.B. von Cranach & Frenz, 1969; Faßnacht, 1995, S. 128ff.). In diesem Gerät läuft mit konstanter Geschwindigkeit ein Papierstreifen unter einigen Schreibstiften. Jedem Stift ist ein Zeichen des Beobachtungssystems zugeordnet. Er läßt sich durch eine Taste bewegen, so daß der Beobachter die Aufgabe hat, immer solange die dem Zeichen zugeordnete Taste zu drücken, wie das entsprechende Verhalten gezeigt wird. Das „Auslenk"-Muster der Schreibstiftspuren läßt sich danach direkt in Anfangs- und Endzeitpunkt eines Geschehens übersetzen. Der Nachteil dieser Geräte ist ihre Unhandlichkeit. Heute wird man eher Computer in Taschenrechnergröße so programmieren, daß der genaue Zeitpunkt bestimmter Tastendrucke gespeichert wird (vgl. z.B. Bakeman & Gottman, 1986; solche Verfahren wurden für die Selbstbeobachtung eingesetzt von Pawlik & Buse, 1982, und Perrez & Reicherts, 1989). Bei vermittelter Beobachtung ergibt sich folgende Lösung: Das Originalgeschehen wird mit der Videokamera aufgezeichnet; eine Uhr ist stets eingeblendet. Da der Beobachter das Band nach Belieben anhalten kann, ist eine zusätzliche Mitkodierung der Zeit zuzumuten. Wegen der gestiegenen Verfügbarkeit der technischen Apparaturen bietet sich eine solche vermittelte Beobachtung zur Reproduzier- und besseren Auswertbarkeit an.

4.3 Die Aufgabe des Beobachters: Sortier- versus Detektorverfahren

Die Einteilung nach den Segmentierungsarten ist nicht die einzige, die im Zusammenhang der Einheitenbildung zu diskutieren ist. Es gibt noch eine weitere Entscheidungsalternative, die weitgehend unabhängig von der ersten ist. In den bisherigen Äußerungen und den Beispielen deutete sich an, daß das praktische Vorgehen bei der Anwendung der Beobachtungssysteme offenbar so ist, daß von dem Geschehen, welches sich vor dem Beobachter abspielt, zunächst immer wieder eine Einheit isoliert wird, welche dann mit Hilfe des Klassifikationssystems einsortiert wird. Der Beobachter arbeitet sozusagen als „2-Takter": Einheit isolieren - einsortieren (Abb. 14). Wir wollen dieses Vorgehen „Sortierverfahren" nennen.

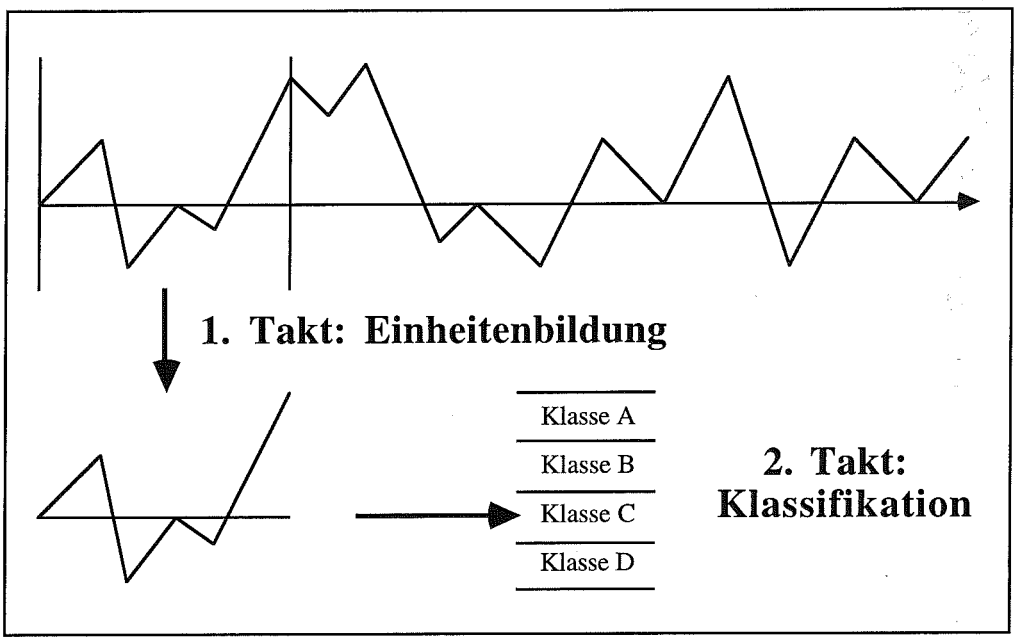


Abbildung 14: Beobachtung im Sortierverfahren

Dieses Vorgehen finden wir beim „Zeittakt"-Verfahren, aber auch bei von Cranachs handlungstheoretisch motivierter semantischer Segmentierung. Wie war es aber bei Lovaas et al. (1973)? Dort sollte der Beobachter offenbar nur reagieren, wenn ganz bestimmte Verhaltensweisen, die durch die Zeichen definiert wurden, gezeigt wurden. Tatsächlich ist es häufig so, daß uns nur sehr wenige Einheiten interessieren; die meisten der isolierten Segmente würden dementsprechend unter einer Rest-Klasse abgelegt. In diesen Fällen bietet sich ein anderes Vorgehen an: Wir versuchen als Beobachter, bestimmte Ereignisse im Geschehen zu entdecken, d.h. wir arbeiten als Detektor (Abb. 15).

Anders formuliert: Im Fall 1 (Sortierverfahren) suchen wir eine Klasse für eine segmentierte Einheit, im Fall 2 (Detektorverfahren) suchen wir eine passende Einheit für die jeweilige Klasse.

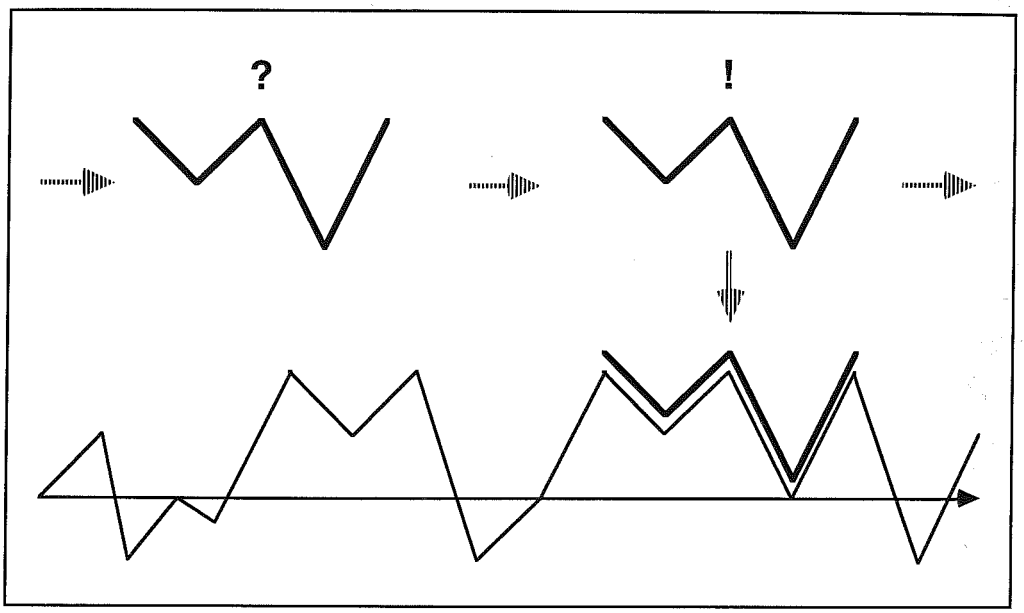


Abbildung 15: Beobachtung im Detektorverfahren

Wir hatten eben gesagt, daß die Unterscheidungen zwischen Sortier- und Detektorverfahren einerseits und zwischen formaler und semantischer Einheitendefinition andererseits unabhängig voneinander seien; also bleibt noch zu zeigen, daß der Beobachter als Detektor sowohl nach formalen als auch nach semantischen Mustern suchen kann. Lovaas et al. (1973) stehen für letzteres; ersteres liegt vor, wenn wir etwa im Zusammenhang einer ethologischen Studie das „Aufblicken" von Personen beim Essen registrieren (Barash, 1972; diese Studie wird im Kapitel 4.5 als Beispiel dienen und dort näher geschildert). Wir können die beiden Unterscheidungen in einer Übersicht zusammenstellen (s. Abb. 16). Wenn wir die gebräuchlichsten Beobachtungssysteme in diesem Vier-Felder-Schema unterzubringen hätten, würde sich allerdings kaum eine Gleichverteilung auf die vier Zellen ergeben. Die weitaus meisten Systeme würden entweder in der (Teil-) Zelle der „Zeittakt-Verfahren" oder unter der Rubrik „Mustererkennen nach semantischen Kriterien" landen.

Es ist wichtig, auf einen etwas problematischen Punkt dieser Einteilung hinzuweisen: Die Unabhängigkeit dieser beiden Dimensionen scheint nicht in allen Fällen stringent durchzuhalten zu sein. Der „erste Takt" des „2-Takters", die Segmentierung in Einheiten, ist im Fall einer semantischen Segmentierung in gewissem Sinne auch eine „Mustererkennung". Das

Beispiel der Handlungsbeobachtung macht das vielleicht deutlich. Um eine Handlung als *eine* Handlung zu erkennen (d.h. von der vorigen und folgenden Handlung abzugrenzen), muß ich sie als *diese* Handlung identifizieren. Das aber heißt nichts anderes, als im Sinne des „Mustererkenners" einen Fall einer bestimmten Klasse (hier: die Handlung vom Typ x) zu erkennen. Damit ist die Einordnung der segmentierten Ereignisse in die jeweilige Klasse - eigentlich der „zweite Takt" - schon vorweggenommen.

		Segmentierung des Geschehens	
		formal	semantisch
Aufgabe des Beobachters	Sortierverfahren	Segmentierung nach Zeittakten, Versuchspersonen, syntaktischen Spracheinheiten etc.	Segmentierung nach Handlungen, teleologischen Kategorien
	Detektorverfahren	Entdecken von Sprachstörungen, Bewegungsmustern etc.	Entdecken von Handlungsmustern, Interaktionsmustern etc.

Abbildung 16: Zwei Dimensionen der Bildung von Beobachtungseinheiten

Trotzdem sind beide Dimensionen sinnvoll: Die Einteilung Sortier- versus Detektorverfahren zielt vor allem ganz pragmatisch auf die Aufgabe des Beobachters ab. Während er im letzteren Fall das Geschehen vor sich ablaufen läßt, um immer „Alarm zu schlagen", wenn er ein Zeichen-Verhalten entdeckt, ist er im ersten Fall ständig gefordert, das Geschehen einzuteilen und zu klassifizieren. Dabei wird die Einteilung auch bei „semantischer Segmentierung" häufig nach „gröberen" Mustererkennungsprozessen ablaufen, als sie für das Klassifizieren im zweiten Takt nötig sind (ein Beispiel hierfür ist das Beobachtungsschema von Bales, 1950a; siehe Abschnitt 5.3).

Um Kriterien zu entwickeln, die uns im konkreten Fall helfen, zwischen den Varianten zu entscheiden, müssen wir uns genau überlegen, welche Informationen wir brauchen, welchen technischen Aufwand wir treiben wollen, welche meßtheoretischen Anforderungen wir an unser System stellen etc. Natürlich hängt die Wahl in erster Linie von unserer inhaltlichen Fragestellung ab. Bei unseren Überlegungen müssen wir uns aber immer von der Frage leiten lassen, wie wir nachweisen können, daß die Ergebnisse unserer Untersuchung zuverlässig sind. Neben inhaltlicher Bedeutsamkeit ist methodische Strenge immer geboten. Im Kapitel 3 haben wir die Vielfalt von Beobachtungsfehlern kennengelernt. Es entstand möglicherweise der Eindruck, daß Beobachter sehr unzuverlässige Meßinstrumente sind. Gerade aus diesem

Grunde ist das wichtigste Mittel zum Nachweis der Güte der Untersuchung der Einsatz mehrerer Beobachter, deren unabhängig voneinander gewonnene Ergebnisse dann auf Übereinstimmung geprüft werden können. Sowohl die Bestimmung der Übereinstimmung als auch die Höhe derselben hängen stark von der Segmentierungsmethode ab. Wählt man etwa die Rasterung in Zeittakte, ist der einzig problematische Punkt die Wahl des Zeichens; überläßt man dagegen schon die Segmentierung den Beobachtern, indem man sie auffordert, das Geschehen nach sinnvollen Handlungseinheiten einzuteilen, können Unterschiede gleich zweifacher Art auftreten: Es ist nämlich weder sicher, ob gleiche Segmentierungspunkte, noch ob gleiche Zeichen gewählt werden.

In den folgenden Abschnitten wollen wir diese Probleme nacheinander behandeln. Zunächst wenden wir uns den Problemen der Einheitenbildung zu, die aus der inhaltlichen Fragestellung entstehen (Abschnitt 4.4). Die Stichworte „Operationalisierung" und „Validität" werden dabei im Zentrum der Überlegungen stehen. Zum Abschluß wird speziell das Problem der Beobachterübereinstimmung angesprochen (Abschnitt 4.5).

4.4 Festlegung der Zeichen: „Operationalisierung"

Wir hatten im ersten Kapitel am Ende unserer kleinen historischen Skizze (Abschnitt 1.3) die theoretische Vernetzung nicht beobachtbarer Begriffe (wie Erwartung, Absicht, Motivation, geistige Leistungsfähigkeit) mit beobachtbaren Verhaltensweisen (die Reaktionen der Schüler, Lob und Tadel der Lehrer) sowie die Verzahnung verschiedener Methoden der Datengewinnung als Kennzeichen der „modernen" Psychologie herausgestellt. Hier ist jetzt der Ort, diese Beziehungen etwas genauer herauszuarbeiten.

Zunächst: Wie hat man sich das Verhältnis nicht beobachtbarer Begriffe zu beobachtbaren Verhaltensweisen vorzustellen? Nehmen wir ein Beispiel: Glennon und Weisz (1978; eine Kurzzusammenfassung findet sich in Manns et al., 1987) entwickelten die „Preschool Observation Scale of Anxiety" (POSA), ein Zeichensystem zur „Messung des Ausmaßes situationsbedingter Angst bei Kindern, die zu jung sind, um genaue Auskunft über ihren inneren Zustand zu geben" (Manns et al., 1987, S. 215). Das System umfaßt 30 Verhaltensklassen, die nach Durchsicht der Literatur und dem Expertenurteil von drei Psychologen, die im klinischen Bereich mit Kindern arbeiten, als die wichtigsten beobachtbaren Verhaltenszeichen von Angst ausgewählt worden waren. Zu der Zeichenmenge gehören etwa „Schreien", „Weinen", „Zitternde Lippen" etc. Die Daten werden im „Zeittakt"-Verfahren (30-Sekunden-Intervalle) erhoben. Der Wert der „situationsbedingten Angst" (typische Situation: das Kind in einer Anforderungssituation mit einem Erwachsenen) für ein Kind ergibt sich als Gesamtsumme der registrierten Zeichen. Ist „Angst" nun genau das, was die POSA mißt? Haben die Verhaltenszeichen definitorischen Charakter für das Konstrukt „Angst"? Nicht ohne weiteres. Offenbar haben wir schon vorher einen recht komplexen Begriff von Angst, der zunächst bei den subjektiven Qualitäten

ansetzt: „Angst" heißt, eine Bedrohung (und sei sie auch nur vorgestellt) wahrzunehmen, einen affektiv negativen Zustand zu empfinden, bestimmte Reaktionsmuster zu zeigen. „Angst" besteht aus mehreren Komponenten, die zum Teil wesentliche Bestandteile (Angst ist in aller Regel Angst vor etwas; insofern können wir uns Angst ohne Bedrohungskognition schwer vorstellen), zum Teil kontingente Begleiterscheinungen darstellen. So würden wir etwa dem Erwachsenen, der Angst äußert, in der Regel diesen Zustand auch unterstellen, selbst wenn bestimmte charakteristische Verhaltensmanifestationen fehlen. Andererseits würden wir gerade im Kindesalter genau diesen Verhaltenskomponenten besondere Aufmerksamkeit widmen und gerade auch entgegen der Selbstauskunft (der „tapfere" Sechsjährige: „I-Ich hab' doch k-keine Angst im D-Dunkeln") den Zustand annehmen.

Wir haben damit zwei Probleme angesprochen, die - eng miteinander verknüpft - gerade die Psychologie besonders „plagen": das Problem der Validität von Meßverfahren beim Fehlen eines Kriteriums und die Frage, warum wir unsere Psychologie nicht auf „harten" Beobachtungsfakten im Sinne sogenannter operationaler Definitionen aufbauen können: „Angst" *ist* das, was der Test xy mißt. Bleiben wir zunächst bei dieser letzten Frage.

4.4.1 Das Problem operationaler Definitionen

Die Schwierigkeit des Verhältnisses von Verhaltensindikatoren und psychologischen Konstrukten zeigt sich in besonderer Weise bei den sogenannten Dispositionsprädikaten: intelligent, aufrichtig, leistungsmotiviert etc. Gemeinsam ist diesen Begriffen, daß sie offenbar eine bestimmte „Fähigkeit oder Neigung ... [bezeichnen], unter geeigneten Umständen in bestimmter Weise zu reagieren" (Stegmüller, 1974, S. 214). So sollte etwa ein intelligenter Mensch besonders viele Aufgaben in einem Intelligenztest lösen. Warum dann nicht gleich die Disposition „Intelligenz" über das Abschneiden in einem solchen Test definieren? Das hätte etwa folgende Form: Für alle Personen gilt: Eine Person hat die Disposition „intelligent" per definitionem genau dann, wenn sie im Test T einen Wert von über y erreicht (vgl. allgemein zur operationalen Definition Brandtstädter, 1986; Herrmann, 1973; Stegmüller, 1974). Tatsächlich gibt es einige Probleme bei diesem Vorgehen. Zum einen wirkt eine konkrete Aufgabenliste einigermaßen beliebig, um definitorisch das von uns durch theoretische Überlegungen oder sprachliche Analysen Gemeinte auszudrücken. Zum anderen gibt es eine formale Schwierigkeit, auf die der Wissenschaftstheoretiker Carnap hingewiesen hat (wir folgen hier der Darstellung von Brandtstädter, 1986, sowie Stegmüller, 1974). Im Kern der operationalen Definition steht eine logische Implikation: „Wenn: Person in Testsituation T, dann: Person hat Wert über y". Genau dann, wenn diese Implikation wahr ist, ordnen wir der Person das Merkmal „intelligent" zu. Das Problem besteht darin, daß die logische Implikation (also der Satz als Ganzes) immer (auch) dann wahr ist, wenn der Vordersatz (der „wenn..."-Teil) falsch ist (vgl. z.B. Seiffert, 1973). Das heißt: nehmen wir die operationale Definition ernst, müssen wir die Folgerung ziehen, daß wir alle Perso-

nen, die noch nie einen Intelligenztest mitgemacht haben (bei denen also der „wenn"-Teil falsch ist), „intelligent" nennen müssen. Das mag etwas „haarspalterisch" klingen, und sicherlich wird mancher Leser die Ursache des Problems nicht so sehr bei der operationalen Definition, sondern bei deren Formalisierung sehen. Allerdings sollte man nicht vergessen, daß der Vorteil der operationalen Definition gerade darin lag, daß eine klare, formalisierbare Lösung für das Problem der Dispositionsbegriffe gegeben werden sollte. Zweifelt man die Formalisierung an, verliert die Lösung an Bedeutung.

Carnap suchte eine andere, angemessenere Lösung zur Konzeptualisierung der Dispositionsbegriffe: Er stellte die Bestandteile der operationalen Definition so um, daß eine Aussage über die Disposition nur noch dann gemacht werden kann, wenn die Person den Testbedingungen ausgesetzt wird. Er formulierte sogenannte „Reduktionssätze":

1. „Wenn die Person der Testsituation ausgesetzt wird, dann hat sie das Merkmal z genau dann, wenn sie das Verhalten y zeigt." (bilateraler Reduktionssatz)
2. „Wenn die Person der Testsituation ausgesetzt wird, dann wird sie, falls sie das Merkmal z hat, das Verhalten y zeigen." (notwendiger Reduktionssatz)
3. „Wenn die Person der Testsituation ausgesetzt wird, dann besitzt sie, falls sie das Verhalten y zeigt, das Merkmal z." (hinreichender Reduktionssatz)

Die Reduktionssätze sind keine Definitionen, denn für Definitionen gilt: „Jeder Satz, in dem das definierte Symbol vorkommt, muß aufgrund der Definition in einen anderen Satz korrekt übersetzbar sein, in dem es nicht mehr vorkommt."[26] (Stegmüller, 1974, S. 227) Eine solche Ersetzbarkeit ist bei Reduktionssätzen nicht mehr möglich, da das Dispositionsprädikat mitten im Satz steht. Ein weiterer Unterschied gegenüber der operationalen Definition ergibt sich dadurch, daß wir im Falle der Reduktionssätze nur noch dann eine Aussage über die Disposition machen können, wenn die Testsituation gegeben ist: Wir spezifizieren die Bedeutung des Dispositionsbegriffes also nur zum Teil über einen Reduktionssatz. Tatsächlich spricht nach Carnap nichts dagegen (aber viel dafür!), die Bedeutung eines Dispositionsbegriffs über ein ganzes System von Reduktionssätzen zu erfassen. Wir können verschiedene Verhaltensindikatoren unter verschiedenen Testsituationen in Reduktionssätzen mit der Disposition in Beziehung setzen. Dabei werden diese Sätze teils die Form hinreichender, notwendiger und bilateraler Reduktionen annehmen. Das mag für unser Beispiel der Intelligenzmessung folgende Form annehmen: „Wenn die Person den Intelligenztest T bearbeitet, dann ist sie 'intelligent', falls sie einen Wert größer y erhält"; als notwendiger Reduktionssatz wurde vorgeschlagen: „Wenn die Person in die Situation gebracht wird, ein komplexes, dynamisches System zu steuern (etwa den Bürgermeister einer Kleinstadt zu spielen), dann wird sie dies, falls sie 'intelligent' ist, erfolgreich tun." Das Interessante an den Reduktionssätzen ist nun, daß sich zwei solcher Sätze so verbinden lassen, daß eine empirische Hypothese entsteht: Die Feststellungen „Die Person hat den Test T bearbeitet und einen Wert größer y erhalten" und „Die Person hat das komplexe System

26 Für die operationale Intelligenz-Definition heißt das: Der Satz „Peter ist intelligent" läßt sich in den Satz übersetzen „Peter hat im Test T einen Wert von über y erreicht".

nicht erfolgreich gesteuert" sind *bei Geltung der Reduktionssätze* nicht zusammen möglich. Ob sie allerdings *tatsächlich* gemeinsam zutreffen, kann empirisch getestet werden und führt möglicherweise zu einer Revision des Reduktionssatzsystems und damit zu einer Veränderung der Konzeptualisierung des Dispositionsbegriffes. (Tatsächlich trifft das im Beispielfall zu: Der Reduktionssatz über die Steuerung des komplexen Systems läßt sich so nicht beibehalten, vgl. zu diesem Beispiel Jäger, 1986.)

Allerdings wendet Brandtstädter (1986) mit Recht ein, daß „man bestimmte Reduktionssätze, die zentrale Bedeutungsgehalte des fraglichen Konzeptes wiedergeben, nicht preisgeben wollen [wird]. Insofern erscheinen Listen von Reduktionssätzen weder als rein definitorische oder analytische noch als rein empirisch-hypothetische Satzsysteme; vielmehr erscheint in ihnen Begriffliches und Empirisches miteinander verzahnt" (S. 203). Wenn unsere Person zwar einerseits im Test T einen Wert von über y erzielt hat (wir sie also daraufhin intelligent nennen würden), aber zu einem anderen Zeitpunkt eine sehr einfache Denksportaufgabe nicht löst (wie wir es von intelligenten Menschen erwarten würden), werden wir weder sofort sagen, daß diese Denksportaufgabe nichts mit unserem Begriff von Intelligenz zu tun hat, noch werden wir der Person das Merkmal gleich wieder absprechen. Vielmehr nutzen wir einfache Hilfsannahmen (sie war bei dem Versuch der Lösung der Denkaufgabe gerade auf etwas anderes konzentriert, sie hatte schlecht geschlafen etc.), um uns das hypothesendiskrepante Ergebnis zu erklären. Offenbar läßt sich die Reduktionssatzidee in ihrer einfachen Form nicht halten, der zufolge sich allein aus einem Reduktionssatzpaar eine zwingende empirische Hypothese ableiten läßt. Das wäre – wie auch Carnap sah – keine angemessene Konzeptualisierung angesichts des rationalen Vorgehens des Wissenschaftlers, der nicht bei jedem empirischen Widerspruch zu seinen begrifflichen Annahmen diese sofort verwirft (s. Abschnitt 1.4.2). Also muß die Verbindung von Testsituation, Verhalten und Disposition so rekonstruiert werden, daß es sich dabei um theoretische Annahmen handelt, deren Prüfung von der Geltung geeigneter Randbedingungen abhängt. Wir schaffen uns ein theoretisches Netz, in das Dispositionsbegriffe als theoretische (nicht beobachtbare) Begriffe eingelagert sind und welches außerdem Aussagen über Beobachtbares enthält. Diese Überlegungen führen uns gleich zu der zweiten Frage, die wir in diesem Unterabschnitt ansprechen wollten: der Frage nach der Validität.

Zuvor muß allerdings noch darauf hingewiesen werden, daß auch heute noch von „Operationalisierungen" in der Psychologie gesprochen wird. Damit ist jedoch nicht mehr der Anspruch verbunden, das betreffende Konstrukt zu *definieren*; vielmehr sucht man den theoretischen Begriff möglichst treffend mit Meßverfahren in Beziehung zu setzen. Dabei ist die Idee der Reduktionssätze recht hilfreich, da auf diese Art eine Analyse des angezielten Konstruktes angegangen werden kann.

4.4.2 Das Problem der Validität

Im Kapitel 3.2.1 sind wir auf die Gütekriterien der Beobachtung eingegangen; wir hatten dort allerdings die Validität der Beobachtung nur recht kurz eingeführt. Tatsächlich ist die Validität das wichtigste Kriterium; wir wollen deshalb den Punkt an dieser Stelle noch einmal aufnehmen. Wird wirklich das gemessen, was gemessen werden sollte? Mit dieser Frage hatten wir oben die Validität eingeführt. Wie kann man sie nun aber beantworten? Als einfachste Lösung für dieses Problem bietet sich der Vergleich der Messung mit einem unabhängigen Standard, einem Kriterium an; wir sprechen dann von der sogenannten Kriteriumsvalidität.

(1) Kriteriumsvalidität

Allgemein wird Kriteriumsvalidität definiert als „der Grad der Genauigkeit, mit dem von den Ergebnissen eines Tests direkt auf ein Kriteriumsverhalten geschlossen werden kann" (Jäger, 1986, S. 275). Denken wir etwa an das Beispiel von Lovaas et al. (1973; vgl. Abschnitt 4.1): Ein Kriterium für die Validität des Zeichensystems wäre etwa, daß wir aufgrund der erhaltenen Werte unter Autismus leidende Kinder von anderen unterscheiden können.

Wir können einen zweiten Fall konstruieren: Nehmen wir an, daß wir ein System zur Erfassung von Aufmerksamkeit im Unterricht erarbeitet haben (vgl. Abschnitt 5.2; dort wird ein solches System unter einem anderen Gesichtspunkt vorgestellt). Eine Beobachtungskategorie unseres Systems sei: „Schüler schaut in der Klasse herum." Dahinter steckt die partielle Rückführung des Begriffes „aufmerksam" (bzw. „unaufmerksam") auf den Satz: „Wenn der Schüler in der Klasse herumschaut, dann ist er nicht aufmerksam." Wir haben hier also wieder die Struktur eines (hinreichenden) Reduktionssatzes vor uns.

Was passiert, wenn ein Schüler unaufmerksam ist? Offenbar wird er den Lernstoff dieser Stunde nur sehr ungenügend aufnehmen. Der aufmerksame Schüler dagegen wird hoffentlich - abseits aller Verständnisprobleme - nachher wissen, was in der Stunde vom Lehrer geboten wurde. Aus diesen Überlegungen können wir folgenden Satz bilden: „Wenn ein Schüler während der Stunde unaufmerksam war, wird er in einem Wissenstest über den Lernstoff der Stunde ein schlechtes Ergebnis erzielen." Hier haben wir also einen notwendigen Reduktionssatz vor uns. Verknüpfen wir die beiden Aussagen miteinander, erhalten wir die Formulierung „Wenn der Schüler während der Stunde herumgeschaut hat, dann wird er in einem Wissenstest über den Lernstoff der Stunde ein schlechtes Ergebnis erzielen." Genaugenommen muß man diese Aussage als Wahrscheinlichkeitsaussage formulieren, da der Schüler natürlich vorgelernt haben mag, gerade einige wichtige Konzepte, die während der Stunde gelehrt wurden, mitbekam usw. Das beeinträchtigt aber nicht das Grundprinzip.

Wir würden also als Kriterium für unser Beobachtungssystem zur Messung von Aufmerksamkeit im Unterricht den anschließenden Wissenstest vorgeben. Die Kopplung zwischen Beobachtung und Kriterium ist bei diesem Beispiel allerdings komplexer als bei dem System von Lovaas und Mitarbeitern (1973): Die beiden Messungen (a) durch das Beobachtungssystem

und (b) durch den Wissenstest werden über die nicht beobachtbaren Begriffe „Aufmerksamkeit" und „Wissen" und deren theoretische Verknüpfung zueinander in Beziehung gesetzt. Wir sprechen in einem solchen Fall von Konstruktvalidierung.

(2) Konstruktvalidität

Der Begriff „Konstruktvalidität" geht auf eine Arbeitsgruppe des amerikanischen Psychologen-Verbandes (APA) aus den fünfziger Jahren zurück, die Richtlinien für Gütekriterien von Tests erarbeitete. Heute wird meist eine Veröffentlichung von zwei Mitgliedern dieses Komitees, Lee J. Cronbach und Paul E. Meehl, als frühe Quelle zu diesem Thema zitiert (Cronbach & Meehl, 1955). Was damals insbesondere für psychologische Tests erarbeitet wurde, gilt ebenso für jedes andere Meßinstrument in der Psychologie, also auch für Beobachtungssysteme.

Ein wichtiges Ergebnis der Überlegungen von Cronbach und Meehl war, daß „die Untersuchung zur Konstruktvalidität eines Tests ... sich nicht wesentlich von den allgemeinen wissenschaftlichen Verfahren zur Entwicklung und Bestätigung von Theorien" unterscheidet (S. 300; Übersetzung G/W). (Konstrukt-)Validitätsbestimmung ist Theorieüberprüfung, und sei es auch nur eine simple Theorie wie in dem Aufmerksamkeitsbeispiel. Was heißt das im Detail? Cronbach und Meehl haben die Grundprinzipien der Konstruktvalidierung in ihrem Artikel sehr kompakt zusammengefaßt (1955, S. 290f.); wir werden ihrer Darstellung folgen.

1. Um zu klären, was ein wissenschaftlicher Begriff *ist*, muß man die Gesetze aufstellen, in denen der Begriff auftritt. Das System ineinandergreifender Gesetze, also eine Theorie, soll *nomologisches Netzwerk* heißen.
2. Die Gesetze im nomologischen Netzwerk können (a) meßbare Größen zueinander, (b) theoretische Konstrukte zu meßbaren Größen oder (c) theoretische Konstrukte zueinander in Beziehung setzen. Die Gesetze können dabei statistischer oder deterministischer Natur sein; d.h. es sind sowohl Wahrscheinlichkeitsaussagen, Aussagen über Erwartungswerte, aber auch Aussagen der Art „*Immer* wenn x, dann y" zugelassen.
3. Eine notwendige Bedingung dafür, daß wir ein Konstrukt als wissenschaftlich betrachten, ist, daß dieses Konstrukt einen Platz in einem nomologischen Netzwerk hat, in dem zumindest einige Gesetze meßbare Größen enthalten. Dabei muß es nicht so sein, daß das Konstrukt direkt in einer Beziehung zu einer meßbaren Größe steht: Es darf durchaus so sein, daß es nur über Zwischenschritte, durch die Vermittlung anderer theoretischer Konstrukte zu Meßgrößen in Beziehung steht. Das Konstrukt wird nicht auf Beobachtbares „reduziert"; im Zusammenhang mit anderen Konstrukten kann es aber dazu dienen, Meßbares vorherzusagen.
4. Mehr über ein theoretisches Konstrukt zu erfahren, heißt nach diesem Ansatz, das nomologische Netz zu erweitern und die Komponenten deutlicher herauszuarbeiten. Am Anfang wird dieses Netz noch sehr einfach sein, die Konstrukte und meßbaren Größen werden nur über wenige Beziehungen verbunden sein.

5. Die Weiterentwicklung des Netzes durch die Hinzufügung weiterer Konstrukte oder Gesetze ist dann gerechtfertigt, wenn dies durch die Empirie nahegelegt wird oder wenn durch die Veränderung Vorhersagen vereinfacht werden.

Wenn Daten mit der Theorie im Widerspruch stehen, muß das nomologische Netz verändert werden. Allerdings gibt es dabei „Spielraum": Das nomologische Netz kann an verschiedenen Stellen verändert werden, um Einklang zwischen Theorie und empirischen Daten herzustellen.

(3) Inhaltsvalidität

Ein „nomologisches Netz", welches wir jedoch immer zur Beurteilung von Meßverfahren heranziehen, ist unser (Alltags- oder Fach-)Hintergrundwissen. Als Glennon und Weisz (1978; s.o.) ihr System zur Messung der Angst bei Kindern entwickelten, bezogen sie sich genau auf dieses Hintergrundwissen, nutzten Expertenurteile, zogen Literatur zu Rate. Wir sprechen von „Augenschein-Validität" (face-validity), wenn Meßverfahren, also auch Beobachtungskategorien, *offensichtlich* für das stehen, was sie messen sollen. „Augenscheinvalidität" ist damit ein Aspekt der Inhaltsvalidität. Der zweite Aspekt dieser dritten Validitätsart läßt sich ebenso an dem Beispiel des Systems von Glennon und Weisz (1978) verdeutlichen: Die Autoren hatten darauf zu achten, daß das ganze Spektrum der Erscheinungsweisen von ängstlichem Verhalten in ihrer Zeichenliste abgedeckt war (vgl. zum Begriff der Inhaltsvalidität Mees, 1977b).

Exkurs: Begriffliche Voraussetzung empirischer Untersuchungen

An diesem Punkt drängt sich eine interessante Frage auf. Die theoretisch oder vom gesunden Menschenverstand erwarteten Zusammenhänge, könnte man einwenden, stehen doch aber möglicherweise ihrerseits zur Disposition. Woher nehmen wir die Berechtigung, uns auf derartige Zusammenhangsvermutungen, d.h. auf Theorien zu verlassen, obwohl ihnen empirische Daten widersprechen? Hatten wir nicht oben (Abschnitt 1.4.2; Exkurs) gerade gesehen, daß dies der entscheidende Nutzen empirischer Daten ist: die *Widerlegung* von Vermutungen und Theorien? Eine allgemeine Antwort auf diese Frage ist sicher schwierig, aber es ist wichtig festzuhalten, daß einige dieser Zusammenhänge tatsächlich nicht zur Disposition stehen, sondern „a priori" (d.h. vor aller Erfahrung) festliegen (wir haben diesen Punkt bereits im Zusammenhang mit der Diskussion der Reduktionssätze angesprochen). Wie ist das gemeint? Ein berühmtes Beispiel zur Verdeutlichung.

Die Frage, wieviel Prozent aller Junggesellen tatsächlich unverheiratet sind, braucht nicht empirisch untersucht zu werden. Junggesellen sind dadurch definiert, daß sie unverheiratet sind; der Fall eines verheirateten Junggesellen ist a priori ausgeschlossen. Wenn man einen solchen Fall empirisch „fände", *muß* dieser Fall durch eine fehlerhafte Messung (entweder von „Junggeselle" oder von „verheiratet") zustande gekommen sein.

Vielleicht noch ein psychologisches Beispiel zur Erläuterung: Die Tatsache, daß jemand, der *dankbar* ist, die Hilfeleistung, für die er dankbar ist, auf denjenigen zurückführt, dem er dankbar ist, kann nicht mit empirischen Mitteln bezweifelt werden. Der *Begriff* der Dankbarkeit impliziert, daß man *jemandem* dankbar ist, d.h. daß es überhaupt jemanden geben muß, dem man die entsprechende Handlung zuschreibt.

Zu diesem Punkt müßte man aber selbstverständlich nicht nur einen knappen Exkurs, sondern eigentlich ein eigenes Buch schreiben. Als Einführung in die wichtige Frage, was denn in der Psychologie überhaupt empirisch zur Disposition steht, empfehlen sich die

Arbeiten von Smedslund (1978, 1984) und Brandtstädter (1982, 1984; vgl. auch die Aufsätze in dem 1987 von ihm herausgegebenen Sammelband).

Nach diesen Überlegungen über die *Validität* von Beobachtungssystemen wenden wir uns nun nochmals der *Reliabilität* (vgl. Abschnitt 3.2.1) zu, und zwar einem bei Beobachtungssystemen zentralen Aspekt, der Untersuchung der Beobachterübereinstimmung.

4.5 Beobachterübereinstimmung

Auf die Fehleranfälligkeit von Beobachtungsdaten ist im letzten Kapitel ausführlich hingewiesen worden. Dabei hatten wir festgehalten, daß es Ziel von Beobachtertrainings ist, die Fehlerrate zu *senken*. Sie zu *identifizieren*, in ihrer Höhe zu bestimmen, ist dagegen Aufgabe von Untersuchungen zur Beobachterübereinstimmung. Um die Güte eines Beobachtungssystems bzw. einer speziellen Anwendung desselben zu beurteilen, soll ein „Index der Beobachterübereinstimmung" definiert werden. Wir wollen die Entwicklung eines solchen Index an einem konkreten Beispiel herleiten.

Wie weit lassen sich beim Menschen noch Reste instinktgesteuerten Verhaltens nachweisen? Wie sehr zeigt auch der Mensch noch unreflektierte Mechanismen zum Schutz bei der Nahrungsaufnahme, wie man sie in der Tierwelt beobachten kann? Diesen Fragen widmete Barash (1972) eine kleine ethologische Beobachtungsstudie. Er beobachtete die Besucher der „Morris Hall Snack Bar", „strategisch getarnt hinter einer Kaffeetasse" (S. 577; Übersetzung G/W). Suchen Einzelpersonen häufiger die „geschützten" Wandtische auf als Gruppen? Tatsächlich fand Barash genau dieses Ergebnis. Darüber hinaus wies er nach, daß Einzelpersonen, die doch an einem freistehenden Tisch Platz nahmen, häufiger von ihrem Teller aufblickten als die Einzelpersonen an den Wandtischen; für die Gruppen fand sich dieser Unterschied nicht. Barash deutete dieses Verhalten im Sinne von Schutzmechanismen, die sich in ehemals bedrohlichen Umwelten als überlebenswichtig herausgebildet haben mögen und heute lediglich „Verhaltensanachronismen" darstellen: „the snack-bar security syndrome".

Nehmen wir nun für die Einführung der Beobachterübereinstimmung das einfache Zeichensystem von Barash (1972), welches nur die Kategorie des „Aufblickens" enthält. Besteht hier die Notwendigkeit, einen Zuverlässigkeitsnachweis zu führen? Wir denken schon: Auch bei dieser – auf den ersten Blick klaren – Verhaltenskategorie kann es zu Fehleinschätzungen kommen. Beispielsweise ist dem Beobachter immer bewußt, ob er eine Einzelperson oder jemanden aus einer Gruppe beobachtet: Er ist sozusagen nicht „blind" gegenüber Versuchsaufbau und Hypothesen (vgl. die Diskussion der Erwartungseffekte im Abschnitt 3.2.3). Grund genug also, zwei Beobachter einzusetzen. Nehmen wir an, diese beiden Beobachter arbeiten bei jeder Versuchsperson nach der Zeittakt-Methode. Für jedes 5-Sekunden-Intervall geben die Beobachter an, ob sie ein „Aufblicken" erkannt haben. Sehen wir uns dazu einen fiktiven Ausschnitt aus dem Beobachtungsprotokoll für eine Versuchsperson an (Abb. 17).

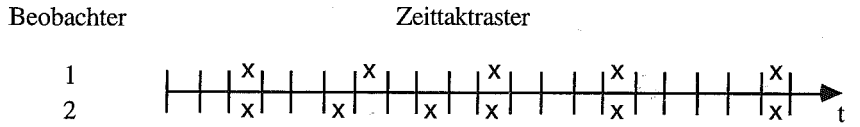


Abbildung 17: Fiktive Beobachtungsdaten

Wie wir sehen, hat in unserem Beispiel Beobachter 1 fünfmal „Aufblicken" beobachtet, während 2 sogar sechsmal das Auftreten dieses Verhaltens wahrgenommen hat. In der Originaluntersuchung von Barash (1972) wurden teilweise drei Beobachter eingesetzt. Tatsächlich fanden sich dort auch leicht unterschiedliche Summen. Diese leichte Diskrepanz in der Summe macht aber noch nicht deutlich, daß die Zuverlässigkeit der Beobachtung eventuell noch mehr zu wünschen übrig läßt: Bezogen auf unser fiktives Beispiel sind in der ersten Hälfte der Zeitachse gleich drei Fälle von Nicht-Übereinstimmung festzustellen. (Da Barash darauf verzichtete, ein Zeittakt-Verfahren zu benutzen, und seine Beobachter als Detektoren im Sinne von Abschnitt 4.3 *ohne* Zeitnotierung arbeiteten, hätte er diese Fälle von Nicht-Übereinstimmung nicht feststellen können.)

Um solche „Rohdaten", wie sie in der Abbildung 17 wiedergegeben sind, besser vergleichen zu können, benötigen wir ein methodisches Hilfsmittel: die Kreuz- oder Kontingenztabelle.

Die Zeilen des Beobachtungsprotokolls aus Abbildung 17 können in folgender Form aufgezeichnet werden, um die Übereinstimmung deutlich zu machen: Da es vier Möglichkeiten für die Ergebnisse eines Zeitintervalles gibt - nämlich: Beide Beobachter haben das „Aufblicken" entdeckt; beide haben nichts dergleichen gesehen; der erste hat das Verhalten notiert, der zweite nicht, und, schließlich, der erste hat nichts bemerkt, der zweite dagegen schon - teilen wir ein Quadrat in vier Einzelzellen und ordnen jeder dieser Zellen ein mögliches Ergebnis zu.

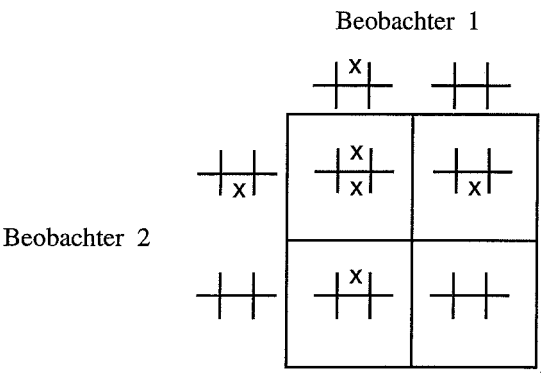


Abbildung 18: Die Kreuztabelle

Die linke obere Zelle steht also für das mögliche Ergebnis, daß beide Beobachter das „Aufblicken" gesehen haben, die untere rechte dafür, daß beide nichts bemerkt haben.
Für jedes tatsächliche Ergebnis der 20 Zeitintervalle notieren wir einen Strich in der zugehörigen Zelle; im Ergebnis sieht das dann so aus (der Einfachheit halber wurden als Spalten- und Zeilenüberschriften '1' für „Kategorie gewählt" und '0' für „... nicht gewählt" benutzt):

		B 1		
		1	0	
B 2	1	4	2	6
	0	1	13	14
		5	15	

Abbildung 19: Die Kreuztabelle für das Beispiel in Abbildung 17

In dieser Kreuztabelle sind zusätzlich zu den Zahlenwerten für die Häufigkeiten noch die „Randhäufigkeiten" aufgezeichnet, also daß der Beobachter 1 fünfmal „Aufblicken" gesehen hat und fünfzehnmal nicht, Beobachter 2 dagegen sechsmal das Verhalten vermerkt hat und vierzehnmal nicht. Dieser Tabelle können die Informationen in folgender Weise entnommen werden: Wenn man sich die Zeile 1 - also die Zeile „Beobachter 2 hat Verhaltenskategorie gewählt" - ansieht, so kann man ablesen, daß der Beobachter 2 sechsmal die Kategorie gewählt hat, in vier Fällen mit Beobachter 1 darin übereinstimmt, in zwei Fällen jedoch nicht. Spalte 1 der Tabelle liefert die entsprechenden Aussagen für „Beobachter 1 hat die Kategorie gewählt".

Wir können nun in einem ersten Schritt unseren „Index der Übereinstimmung" als „Prozentsatz der Übereinstimmung" (%Ü) definieren. Dies scheint zunächst die plausibelste Lösung zu sein. Wir nehmen dazu einfach die Werte aus den Übereinstimmungszellen (also der linken oberen und der rechten unteren), addieren sie und relativieren sie auf die Gesamtzahl der Intervalle:

$$(4 + 13)/20 = 85\ \%Ü,$$

ein recht hoher Wert. Eine einfache und zunächst überzeugende Lösung. Trotzdem gibt es gute Gründe, das Thema Beobachterübereinstimmung nicht schon mit %Ü als abgehandelt zu betrachten. Das Maß %Ü ist nämlich mit prinzipiellen Mängeln behaftet (vgl. dazu Asendorpf & Wallbott, 1979; Bakeman & Gottman, 1986; Feger, 1983).

Wenn man sich die Randhäufigkeiten anschaut, so stellt man - wie schon gesagt - fest, daß die Kategorie „Aufblicken" nicht in gleichem Maße von den beiden Beobachtern gewählt

wurde. Sicherlich - die Differenz beträgt nur eins; man kann sich aber leicht Fälle vorstellen, in denen die Tendenz, eine Kategorie zu wählen, noch stärker differiert. Gedanken über solche Differenzen hat man sich in einem anderen Zweig der Psychologie, der Wahrnehmungspsychologie, im Zusammenhang der Forschung über Wahrnehmungsschwellen gemacht. Der daraus entstandene Ansatz ist unter dem Namen „Signalentdeckungstheorie" (SDT, signal detection theory) bekannt geworden, deren Skizzierung der folgende Exkurs dient.

Exkurs: Die Signalentdeckungstheorie
Während in der älteren Wahrnehmungspsychologie nach absoluten Wahrnehmungsschwellen gesucht wurde (also etwa die Reizintensität bestimmt wurde, die eine Versuchsperson nur noch in 50% der Durchgänge veranlaßt, mit „Ja, ich habe den Reiz wahrgenommen" zu antworten), merkte man später, daß diese absolute Schwelle offenbar so nicht existiert: Die Erwartungen der Versuchspersonen darüber, ob ihnen ein Reiz dargeboten wurde oder nicht, prägen sehr stark die Antworttendenz.

Man begann, das Experiment zu variieren: Nun wurde tatsächlich nicht mehr in jedem Durchgang, sondern nur noch mit einer (der VP bekannten) Wahrscheinlichkeit ein Reiz präsentiert. Es gibt nun für jeden Durchgang vier Ergebnismöglichkeiten, wie sie in der Abbildung 20 veranschaulicht sind.

		Antwort: ja	Antwort: nein
Reiz	an	korr. Entdeckg.	falsche Zurückweisg.
	aus	falscher Alarm	korr. Zurückweisg.

Abbildung 20: Ergebnistafel bei Reizentdeckungsexperimenten

Wenn man nun - bei gegebener Reizstärke - die Erwartung der Versuchspersonen manipuliert, wird man jeweils andere Ergebnistafeln erhalten - ein Indiz dafür, daß Reizentdekkungsleistung und Reaktionstendenz unterschiedliche Komponenten darstellen.

Die Erwartung der Versuchspersonen wird in einem solchen Versuch durch mindestens zwei Quellen gespeist: Zum einen wird die subjektive Wahrscheinlichkeit des Reizauftretens eingehen, zum anderen die „Kosten", die mit „falschem Alarm" bzw. „falscher Zurückweisung" verbunden sind. Man manipuliert diese beiden Quellen in einem solchen Experiment dadurch, daß einerseits die objektive Wahrscheinlichkeit verändert und den Vpn mitgeteilt wird und andererseits per Instruktion der „falsche Alarm" als weniger erwünscht dargestellt wird als die „falsche Zurückweisung" (oder umgekehrt). Daß die im Experiment solchermaßen manipulierten Größen „subjektive Wahrscheinlichkeit" und „Kosten" auch in „Nicht-Labor"-Situationen Bedeutung haben, kann man sich leicht am Beispiel der Radarbeobachtung klarmachen (vgl. einführend zur Signalentdeckungstheorie Velden, 1982).

Wie man sieht, hängt die Bereitschaft zur Wahl der „Ja"-Antwort von sehr vielen Dingen ab, offensichtlich nicht ausschließlich von der Stärke des Reizes. Was hat das nun mit den Überlegungen zur Beobachterübereinstimmung zu tun? Nun, man kann die Grundidee der Signalentdeckungstheorie nicht nur auf Wahrnehmungsexperimente mit schwachen Reizen beziehen, sondern auch auf die unterschiedliche Bereitschaft, eine Beobachtungskategorie zu wählen. Um die Parallelen ganz deutlich zu machen: Die Wahl einer Kategorie hängt, außer vom tatsächlichen Sachverhalt, auch von Erwartungen über die Wahrscheinlichkeit des Auftretens und von Kostengesichtspunkten ab und nicht nur von der Eindeutigkeit und Schärfe der Festlegungen.

Wir können diese Überlegungen auf die Problematik des %Ü-Maßes beziehen. Denken wir uns folgendes Ergebnis einer Übereinstimmungsuntersuchung. (Die beiden nachfolgenden Zahlenbeispiele sind - leicht modifiziert - Light (1971) entnommen, der ebenfalls auf die Problematik des %Ü aufmerksam macht.)

		B 1		
		1	0	
B 2	1	20	0	20
	0	55	25	80
		75	25	

Abbildung 21: Zahlenbeispiel 1

Offensichtlich haben die beiden Beobachter in diesem Fall eine stark unterschiedliche Tendenz, die Kategorie zu wählen: Während Beobachter 1 dies in 75 von 100 Fällen tut, hält sich Beobachter 2 mit 20 Wahlen sehr zurück. Natürlich kann man zunächst ein Beobachtungssystem stark anzweifeln, welches eine so unterschiedliche Randverteilung (20 zu 80 vs. 75 zu 25) erbringt. Bevor wir aber diese Differenz voll und ganz dem Beobachtungssystem anlasten, sollte man die Tabelle etwas genauer anschauen: Zeile 1 besagt nämlich, daß in den 20 Fällen, in denen Beobachter 2 die Kategorie gewählt hat, er dies übereinstimmend mit Beobachter 1 getan hat. Man könnte also behaupten, daß eine Konsistenz in der Benutzung gegeben ist, trotz der unterschiedlichen Bereitschaft, die Kategorie zu wählen. Es ist sogar so, daß - gegeben die unterschiedlichen Randverteilungen - ein Maximum an %Ü erzielt wird. Diesen Sachverhalt kann man sich leicht daran klarmachen, daß man eine leere Kreuztabelle aufmalt, zunächst nur die Randverteilungen mit 20 zu 80 vs. 75 zu 25 festlegt und dann versucht, die Zellenwerte so zu

bestimmen, daß einmal ein Maximum und im anderen Fall ein Minimum an %Ü entsteht; Abbildung 22 verdeutlicht dies.

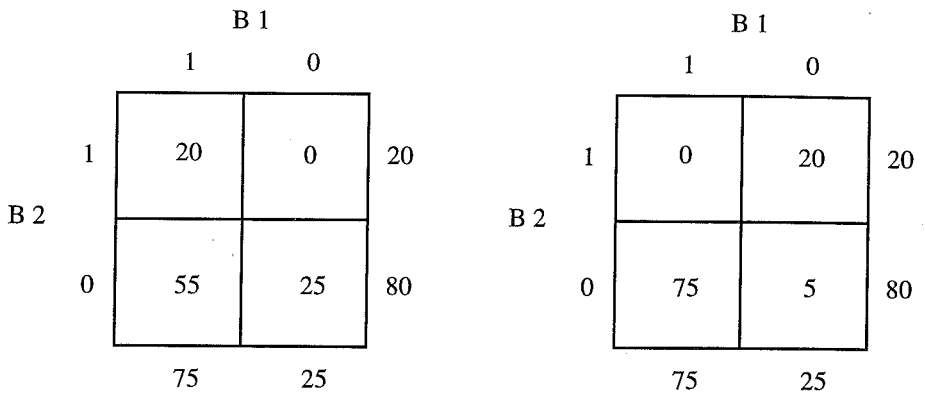


Abbildung 22: Verschiedene %Ü bei gleicher Randverteilung (Bsp. 1)

Betrachten wir nun aber ein zweites Zahlenbeispiel (Abb. 23).

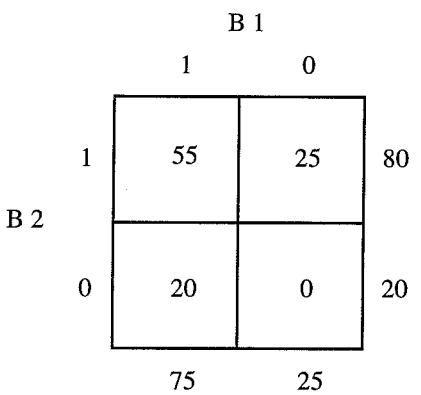


Abbildung 23: Zahlenbeispiel 2

Im Gegensatz zu den 45 %Ü des ersten Beispiels haben wir hier sogar 55 %Ü und auch noch fast identische Randverteilungen! Eine bessere Übereinstimmung als im ersten Beispiel, so könnte man zunächst sagen. Wenn man aber hier dem Vorschlag folgt, die Randverteilungen an eine leere Kreuztabelle zu schreiben und diese Randhäufigkeiten so zu verteilen, daß einmal ein

Maximum, ein zweites Mal ein Minimum an %Ü entsteht, so stellt man fest, daß im zweiten Zahlenbeispiel nur ein Minimum an %Ü - gegeben die Randverteilungen - verwirklicht wurde; Abbildung 24 verdeutlicht dies.

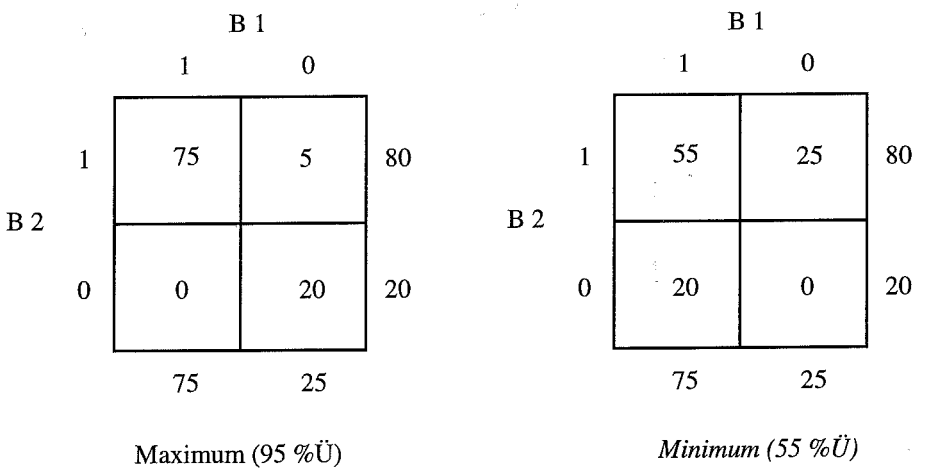


Abbildung 24: Verschiedene %Ü bei gleicher Randverteilung (Bsp. 2)

Egal, ob wir uns die Kategorie-Wahlen des ersten oder zweiten Beobachters anschauen – immer ergeben sich einige Fälle von Nicht-Übereinstimmung; es scheint also an der Konsistenz der Nutzung der Kategorien zu mangeln. Damit haben wir zwei Fehlermöglichkeiten herausgearbeitet, die es zu unterscheiden gilt:

1. „Fehler der unterschiedlichen Bereitschaft"; es kann eine unterschiedliche Neigung der Beobachter geben, eine Kategorie zu wählen, so daß sich unterschiedliche Randverteilungen ergeben.
2. „Fehler der mangelnden Konsistenz"; die Beobachter benutzen die Kategorie nicht übereinstimmend.

Wichtig ist nun, daß dies zwei konzeptuell sehr unterschiedliche Fehler sind; häufig ist sogar in Frage zu stellen, ob wir die Unterschiedlichkeit der Randverteilungen überhaupt als Fehler rechnen und uns nicht lieber auf die Konsistenz der Benutzung konzentrieren wollen. Dies ist wieder auf Fragestellung und Einzelfall zu beziehen und zu begründen. Beispielsweise könnte es in der Trainingsphase eines Beobachters sinnvoll sein, vor allem darauf zu achten, daß er eine Kategorie, *wenn* er sie benutzt, auch richtig benutzt, und dabei zunächst unberücksichtigt zu lassen, ob er sie in *allen* Fällen benutzt, in denen der Trainer sie benutzt. Das %Ü-Maß differenziert auf jeden Fall nicht zwischen diesen beiden Fehlern!

Die Randverteilungen interessieren aber noch aus einem weiteren Grunde. Man stelle sich vor, unsere beiden Beobachter beobachten tatsächlich überhaupt nicht, sondern beide würfeln jeder für sich ihr „Beobachter-Ergebnis" aus; der eine notiert immer dann, wenn er eine eins

würfelt, einen Fall von „Aufblicken" auf seinem Bogen, er notiert nichts bei allen anderen Ergebnissen; der andere dagegen notiert nichts, wenn er eine eins würfelt, aber „Aufblicken" in den restlichen Fällen. Wenn sie dies unendlich oft - ein Gedankenexperiment! - tun würden, so sollten sich folgende *relative* Randhäufigkeiten bzw. Wahrscheinlichkeiten ergeben (Abb. 25; zur Notation: die „0" vor dem Dezimalpunkt wird immer weggelassen).

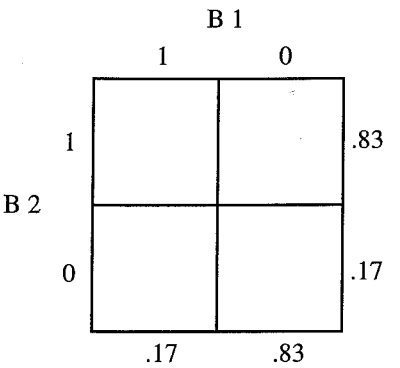


Abbildung 25: Relative Randhäufigkeiten im „Würfelexperiment"[27]

Was erwarten wir nun bei diesem Experiment für die Zellhäufigkeiten? Aus der Wahrscheinlichkeitstheorie wissen wir, daß wir für die Wahrscheinlichkeit des Zusammentreffens zweier unabhängiger Ereignisse nur die Wahrscheinlichkeiten zu multiplizieren brauchen (vgl. etwa Bortz, 1993, Kap. 2), hier also:

p(B1 wählt 1 und B2 wählt 1) = p(B1 wählt 1) * p(B2 wählt 1)[28]
= .83 * .17 = .14

Für Die Randverteilungen in Abbildung 25 ergeben sich dann die in Abbildung 26 dargestellten Zellenwahrscheinlichkeiten.

Da uns in einer echten Beobachterübereinstimmungs-Untersuchung interessiert, inwieweit das tatsächliche Ergebnis von einem zufälligen abweicht, bestimmen wir auch hier die Erwartungswerte für die Zellen, um diese dann mit den tatsächlich erhaltenen zu vergleichen. Um diese Erwartungswerte zu bestimmen, brauchen wir aber die Randverteilungen! Man kann zwei Wege beschreiten:

1. Man nimmt die tatsächlichen relativen Randhäufigkeiten als Schätzung für die Wahrscheinlichkeit, d.h. man gesteht den „Fehler der unterschiedlichen Bereitschaft" zu.
2. Um die objektiven Wahrscheinlichkeiten des Auftretens zu schätzen, kann man die Annahme machen, daß beide Beobachter einen gleich großen „Fehler der unterschiedlichen Bereit-

---

27 Die Werte sind so zu lesen: Mit einer Wahrscheinlichkeit von .83 (=5/6) bzw. in 83% aller Fälle wählt Beobachter 2 eine '1', mit einer Wahrscheinlichkeit von .17 (=1/6) eine '0'.

28 So zu lesen: „Die Wahrscheinlichkeit für das Ereignis „B1 wählt 1 und B2 wählt 1" ist gleich dem Produkt der Wahrscheinlichkeit für das Ereignis „B1 wählt 1" und für das Ereignis „B2 wählt 1".

schaft“ - nur in verschiedener Richtung - machen und die Mittelwerte der relativen Randhäufigkeiten als Schätzung nehmen. Diese Annahme ist aber natürlich alles andere als selbstverständlich und muß im Einzelfall begründet werden.

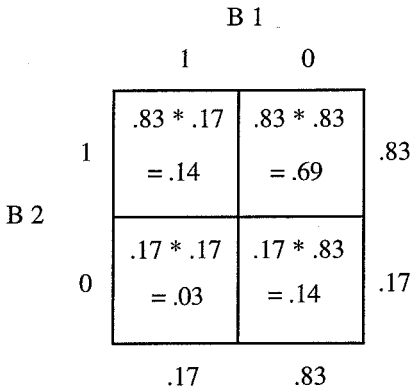


Abbildung 26: Relative erwartete Zellhäufigkeiten im „Würfelexperiment“

Diese Überlegungen seien jetzt noch einmal an den beiden Beispiel-Kreuztabellen demonstriert. Wir notieren sie diesmal mit *relativen* Häufigkeiten (also absolute Häufigkeiten geteilt durch die Anzahl der Fälle; Abb. 27).

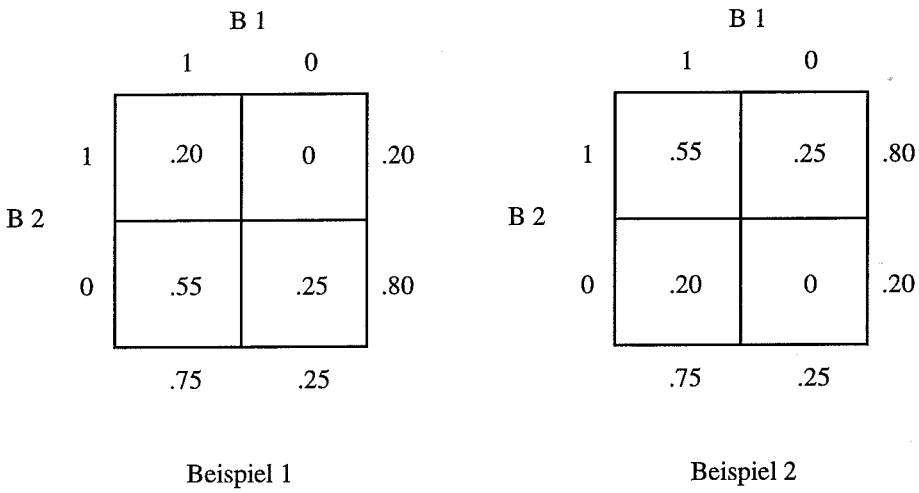


Abbildung 27: Tatsächliche relative Häufigkeiten für die beiden Zahlenbeispiele

Geht man nach Prinzip 1 vor (d.h. die tatsächlichen relativen Häufigkeiten werden als Schätzer für die Wahrscheinlichkeiten genommen), erhalten wir folgende Erwartungswerte für die Zellen:

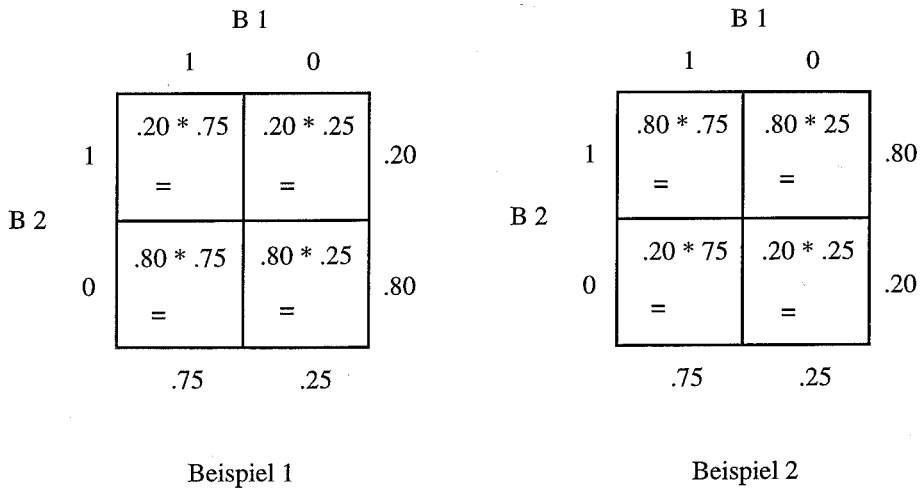


Abbildung 28: Relative Erwartungshäufigkeiten nach Prinzip 1 für beide Zahlenbeispiele

Wie man sieht, ist die Abweichung der tatsächlichen von der erwarteten Häufigkeit in beiden Fällen nicht besonders hoch, wir kommen darauf zurück!

Verfährt man nach Prinzip 2, müssen wir zunächst die Erwartungen für die Randhäufigkeiten durch Mittelung bestimmen, um im zweiten Schritt die Zellenerwartungen zu berechnen: (Abb. 29).

Während sich für das erste Zahlenbeispiel die Erwartungswerte drastisch verändern, zeigt sich beim zweiten Beispiel nur eine geringfügige Diskrepanz; dieses Ergebnis war natürlich zu erwarten, da im zweiten Beispiel die Randhäufigkeiten für die beiden Beobachter fast gleich waren und daher die Entscheidung, ob eine Mittelung angebracht ist oder nicht, wenig ins Gewicht fällt.

B 2 \ B 1	1	0	
1	.475*.475 = .226	.475*.525 = .249	(.20+.75)/2 = .475
0	.525*.475 = .249	.525*.525 = .276	(.80+.25)/2 = .525
	(.75+.20)/2 = .475	(.25+.80)/2 = .525	

Beispiel 1

B 2 \ B 1	1	0	
1	.775*.775 = .601	.775*.225 = .174	(.80+.75)/2 = .775
0	.225*.775 = .174	.225*.225 = .051	(.20+.25)/2 = .225
	(.75+.80)/2 = .775	(.25+.20)/2 = .225	

Beispiel 2

Abbildung 29: Relative Erwartungshäufigkeiten für unsere Zahlenbeispiele nach Prinzip 2

Um ein Gütemaß für die Beobachterübereinstimmung zu bekommen, müssen nun die tatsächlichen auf die erwarteten Häufigkeiten relativiert werden.

Um die folgenden Formeln besser einführen zu können, sei folgende Notation vereinbart:

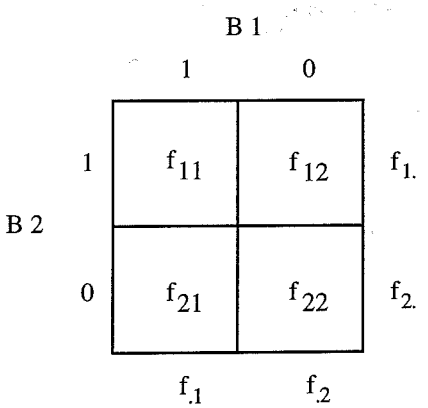


Abbildung 30: Einführung eines Notationssystems

wobei hier alle Platzhalter („f") für *relative* Häufigkeiten stehen. Außerdem wollen wir von „Beobachtungstabelle" sprechen, wenn von den tatsächlich erhaltenen relativen Häufigkeiten die Rede ist, von „Erwartungstabelle" dann, wenn wir auf die erwarteten Zellhäufigkeiten Bezug nehmen.

Man kann nun ein Gütemaß G nach folgender Formel[29] bestimmen:

$$G = \frac{P_o - P_e}{1 - P_e}$$

mit

P_o: tatsächliche („observed") Übereinstimmung
(= $f_{11} + f_{22}$ aus der „Beobachtungstabelle")

P_e: erwartete („expected") Übereinstimmung
(= $f_{11} + f_{22}$ aus der „Erwartungstabelle")

Wie kommt man zu dieser Formel? Diese Art der Verrechnung schafft zwei sinnvolle „Eckpunkte" eines Übereinstimmungsmaßes: Der Koeffizient nimmt den Wert „0" an, wenn die erwartete gleich der tatsächlichen Übereinstimmung ist (der Zähler wird „0"); er wird „1", wenn die tatsächliche Übereinstimmung maximal ist, d.h. wenn alle Fälle in den Zellen f_{11} und f_{22} liegen (P_O=1 und damit: Zähler gleich Nenner). Alle Werte zwischen „0" und „1" bedeuten dann graduelle Abstufungen der Übereinstimmung. Werte unter „1" zeigen an, daß die Beobachter systematisch nicht übereinstimmend geurteilt haben. Im einfachsten Fall liegt hier ein Mißverständnis vor: Der eine Beobachter hat immer dann einen Strich gemacht, wenn er laut Vereinbarung keinen machen sollte.

29 G ist für den Fall nicht definiert, daß Pe = 1 ist. Dies ist dann der Fall, wenn beide Beobachter völlig übereinstimmend keinerlei Variabilität des beobachteten Verhaltens wahrnehmen („Beobachten Sie im fünf-Minuten-Takt, ob dieser Stuhl niest!").

Je nachdem, ob wir f_{11} und f_{22} der Erwartungstabelle nach Prinzip 1 oder 2 berechnen, erhalten wir den von Cohen (1960) vorgeschlagenen Koeffizienten κ („Kappa") bzw. den von Scott (1955) entwickelten Koeffizienten π („Pi") (vgl. auch Bortz & Döring, 1995, S. 253f.; Frick & Semmel, 1978; Feger, 1983; Zwick, 1988):

κ (Cohen, 1960):

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

mit

P_o: tatsächliche Ü
$= f_{11} + f_{22}$ (f_{11} und f_{22} aus der „Beobachtungstabelle")

P_e: erwartete Ü
$= f_{11} + f_{22} = f_{1.} * f_{.1} + f_{2.} * f_{.2}$ (f_{11} und f_{22} aus der „Erwartungstabelle" nach κ)

π (Scott, 1955):

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

mit

P_o: tatsächliche Ü
$= f_{11} + f_{22}$ (f_{11} und f_{22} aus der „Beobachtungstabelle")

P_e: erwartete Ü
$= f_{11} + f_{22} = \left(\frac{f_{1.} + f_{.1}}{2}\right)^2 + \left(\frac{f_{2.} + f_{.2}}{2}\right)^2$ (f_{11} und f_{22} aus der „Erwartungstabelle"nach π)

Um zu sehen, wie diese beiden Koeffizienten „funktionieren", wollen wir sie nun auf unsere beiden Beispielfälle anwenden.

		Beispiel 1	Beispiel 2
a)	%Ü	$= 45$	$= 55$
b)	π	$= \frac{.45-.50}{1-.50} = -.10$	$= \frac{.55-.65}{1-.65} = -.29$
c)	κ	$= \frac{.45-.35}{1-.35} = .15$	$= \frac{.55-.65}{1-.65} = -.29$

Wie man sieht, ist in Beispiel 2 - gleichgültig, ob man π oder κ „zu Rate zieht" - die schlechtere Übereinstimmung gegeben; der negative Wert weist sogar darauf hin, daß die tatsächliche Übereinstimmung geringer als die Zufallserwartung ist! Unser zunächst so plausibel scheinendes Maß %Ü wird dieser Tatsache nicht gerecht.

Beispiel 1 muß dagegen differenziert beurteilt werden: Wiegt der „Fehler der unterschiedlichen Bereitschaft" genauso schwer wie der „Fehler mangelnder Konsistenz", nehmen wir π als Gütemaß. Da π negativ ist, wissen wir, daß auch hier die tatsächliche Übereinstimmung geringer als die zufällig erwartete ausfällt. Werten wir dagegen die unterschiedliche Tendenz nicht so stark, sondern interessieren uns vorwiegend für den „Fehler mangelnder Konsistenz", erhalten wir mit dem κ-Koeffizienten von .15 ein zwar äußerst schlechtes, aber zumindest positives Übereinstimmungsmaß. Wie wir sehen, nimmt κ also nicht etwa den Maximalwert von 1 an, nur weil wir - bei gegebener Randverteilung - ein Maximum an %Ü haben. Das wäre sicher auch nicht wünschenswert, da die Übereinstimmung ja weit vom Optimum entfernt ist. Aber κ ist größer als π - eine Beziehung, die immer gilt, wenn die Randverteilungen unterschiedlich sind.

Wir müssen jedoch noch einen Schritt weitergehen, um den so erhaltenen Koeffizienten zu beurteilen: In dem Gedankenexperiment haben wir die beiden Beobachter *unendlich* oft würfeln lassen; erst wenn wir so verfahren, erhalten wir die relativen Häufigkeiten, wie sie oben angegeben wurden. Was passiert aber, wenn wir - entsprechend den hundert Zeitintervallen unseres Übereinstimmungsversuches - jeden Beobachter genau hundertmal würfeln lassen? Selbstverständlich könnten dabei die verschiedensten Kombinationen und damit verschiedensten Kreuztabellen „erwürfelt" werden. Das bedeutet dann natürlich, daß wir auch bei tatsächlicher „Erwürfelung" unserer Beobachtungsergebnisse π- bzw. κ-Koeffizienten erhalten, die von Null verschieden sind. Glücklicherweise kann man mit Hilfe der Wahrscheinlichkeitstheorie ermitteln, mit welcher Wahrscheinlichkeit ein bestimmter Koeffizient zu erwarten wäre, gesetzt den Fall, die Beobachter hätten nur „gewürfelt"! Die Strategie, die man dabei verfolgt, ist diese:

1. Man berechnet in einem Übereinstimmungsversuch einen κ- (bzw. π-) Koeffizienten.
2. Man macht zunächst die Annahme, dieses Ergebnis sei zufällig zustande gekommen.
3. Man berechnet, wie wahrscheinlich ein solches (oder besseres) Ergebnis unter dieser Annahme ist.
4. Sollte das Ergebnis sehr unwahrscheinlich unter dieser Annahme sein, verwirft man diese Annahme und geht davon aus, daß die Beobachter nicht „gewürfelt" haben (also dasselbe beobachtet und dieses Beobachtete außerdem gemäß einer vereinbarten Zuordnungsregel klassifiziert haben). Den unscharfen Begriff des „sehr Unwahrscheinlichen" muß man natürlich präzisieren: Üblicherweise gelten Ereignisse als „sehr unwahrscheinlich", wenn die Wahrscheinlichkeit geringer als 5% ($p < .05$) bzw. als 1% ($p < .01$) ist.

Speziell zur Beurteilung unserer Koeffizienten π und κ geht man so vor: Die Zufallsannahme bedeutet hier, daß π/κ den Wert Null annimmt, wenn die Anzahl der Würfe gegen unendlich geht. Das heißt aber auch, daß wenn wir immer wieder Stichproben von 100 Würfen ziehen, die π/κ-Werte dieser Stichproben sich um den Mittelwert Null gruppieren. Wir brauchen nun lediglich noch Angaben über die Form dieser Verteilung. Glücklicherweise entspricht diese Verteilung bei großen Stichproben (Anzahl der Beobachtungseinheiten >100; vgl. Cohen, 1968) der sogenannten „Normalverteilung", einer glockenartigen Kurve, die zur Beschreibung vieler Zufallsprozesse dient. Um die Breite dieser Verteilung (und damit die Wahrscheinlichkeit einzelner Werte) zu erfassen, wird die „Standardabweichung" berechnet. Dieses Maß reicht aus, um bei einer Normalverteilung die Streuung der Werte um ihren Mittelwert zu beschreiben. Fleiss (1973; vgl. auch Fleiss, Cohen & Everitt, 1969; Hubert, 1977; Light, 1971) gibt die Standardabweichungsformel für κ [30] an, die für unsere einfache Kreuztabelle so aussieht:

$$SD_\kappa = \sqrt{\frac{1}{n(1-P_e)^2} * (P_e + P_e^2 - f_{1.} * f_{.1} * (f_{1.} + f_{.1}) - f_{2.} * f_{.2} * (f_{2.} + f_{.2}))}$$

Diese Formel kann an dieser Stelle nicht ausführlich hergeleitet oder begründet werden; wesentlich ist, daß die erwartete Übereinstimmung (P_e) und die Anzahl der Beobachtungen (n) eingeht.

Der etwas unübersichtliche rechte Teil der Formel mit den relativen Randhäufigkeiten erweist sich bei näherem Hinsehen als verwandt zur Formel zur Berechnung von P_e: wieder wird die erwartete Häufigkeit für die Übereinstimmungszellen über das Produkt der Randhäufigkeiten berechnet, diesmal allerdings gewichtet mit der Summe derselben.

Die Standardabweichung bei unserem Beispiel 1 beträgt demnach:

$$SD_\kappa = \sqrt{\frac{1}{100(1-0.35)^2} * (0.35 + 0.35^2 - 0.15 * 0.95 - 0.20 * 1.05)} = 0.053$$

Was fangen wir jetzt mit diesem Wert an? Wir hatten gesagt, daß bei einer Normalverteilung bei gegebenem Mittelwert (nach unserer Annahme Null) und Standardabweichung genau die Verteilung der einzelnen Werte bestimmt ist; d.h. wir wissen zum Beispiel, wie häufig Werte, die zwischen dem Mittelwert und einer Standardabweichung liegen, „erwürfelt" werden (genau 34.13% aller Werte). Wir wissen auch, ab welchem Wert wir nur noch 5% (1%) aller Fälle zu erwarten haben: Alle Werte, die mehr als 1.645 (2.327) Standardabweichungen vom Mittelwert entfernt liegen, machen genau diesen Prozentsatz aus. Unser Kappawert im Beispiel 1 von .15 ist (.15/.053=) 2.83 Standardabweichungen vom (angenommenen) Mittelwert entfernt; ein

30 Als entsprechende Formel für π schlägt Zwick (1988) die Formel für Huberts „matching model" (1977, p. 292) vor, die der hier angegebenen Formel für κ gleicht; lediglich n wird durch (n-1) ersetzt.

solcher (oder niedrigerer) Wert ist unter der Zufallsannahme also viel seltener als 1% zu erwarten (da 2.83 > 2.327). Wir verwerfen also die Zufallsannahme.

An diesem Beispiel sieht man allerdings auch, daß der Nachweis einer überzufälligen Übereinstimmung nur die Minimalbedingung für ein zuverlässiges Beobachtungssystem darstellt. Man kann es so formulieren: Scheitert der inferenzstatistische Test, d.h. hebt sich die empirische Übereinstimmung nicht von der per Zufall erwarteten ab, sagen die Ergebnisse der Beobachtung überhaupt nichts mehr aus. Wir können aus unseren Bemühungen lediglich einen Schluß ziehen: Dieses Beobachtungssystem funktioniert (noch) nicht. Andererseits kann man den durch Beobachtung gewonnenen Werten um so mehr trauen, je näher der Übereinstimmungskoeffizient an 1 herankommt. Insbesondere dann, wenn wir die Beobachtungswerte zu anderen Variablen in Beziehung setzen wollen, werden wir immer um so besser beurteilen können, ob ein Zusammenhang besteht, je weniger die Beobachtungswerte (und natürlich auch die anderen Variablen) durch Zufallsschwankungen „belastet" sind (vgl. dazu die Überlegungen zur „Konstruktvalidität" im Abschnitt 4.4.2). Bakeman und Gottman (1986, S. 82) geben aufgrund ihrer Erfahrung als Faustregel an, daß κ-Werte unter .70 mit etwas Skepsis betrachtet werden sollten, und zitieren andererseits Fleiss (1983), der κ-Werte zwischen .40 und .60 als annehmbar, zwischen .60 und .75 als gut und darüber als ausgezeichnet bezeichnet. Frick und Semmel (1978) halten in typischen Anwendungsfällen κ-Werte ab .75 für akzeptabel.

Diese Logik der Absicherung gegen Zufallsergebnisse, die hier nur in ihrer einfachsten Form angedeutet wurde, liegt im übrigen der gesamten „schließenden Statistik" (Inferenzstatistik) zugrunde und dient der Beurteilung aller möglichen statistischen Zusammenhänge. So konnte sich etwa Barash (1972), um den Bogen zum Anfang dieses Unterkapitels zu schlagen, nicht damit zufriedengeben, daß bei Einzelpersonen ein Häufigkeitsunterschied im „Aufblicken" festzustellen ist, je nachdem, wo sie in der Snack-Bar Platz nahmen. Er mußte auch zeigen, daß ein solcher Unterschied nur mit einer geringen Wahrscheinlichkeit per Zufall zu erwarten ist. Der an den Problemen der Inferenz-Statistik interessierte Leser kann sich in einschlägigen Lehrbüchern (etwa Bortz, 1993) über die hiermit zusammenhängenden Problematiken eingehend informieren. Hager (1987) diskutiert die Probleme der Inferenzstatistik im Zusammenhang der Versuchsplanung, des Teilgebietes der psychologischen Methodenlehre, in welchem Techniken erarbeitet werden, um psychologische Untersuchungen möglichst aussagekräftig zu gestalten.

Wir sollten uns überdies vor Augen halten (wir hatten das im Abschnitt 3.2 bereits angedeutet), daß die Beobachterübereinstimmung nicht nur lediglich eine *Schätzung* der Reliabilität der Beobachtung darstellt, sondern vor allem keine Gewähr für Genauigkeit und Objektivität dieser Beobachtung liefert. Johnson und Bolstad (1973) weisen darauf hin, daß es irreführend sei, Beobachterübereinstimmung einerseits und Genauigkeit andererseits häufig beide als Synonym (bzw. Kriterium) für Reliabilität und insofern gleichbedeutend zu benutzen. Der Einwand ist berechtigt. Es ist wichtig, sich klarzumachen, daß die Beobachterübereinstimmung von der Genauigkeit unabhängig sein kann (vgl. hierzu auch Foster & Cone, 1980), z.B. dann, wenn alle Beobachter denselben Fehler machen. Dies ist in praktischen Zusammenhängen zum

Beispiel der Fall, wenn den Beobachtern im gemeinsamen Beobachtertraining ein bestimmter Fehler (versehentlich) beigebracht wurde (Manns et al., 1987, S. 48). Nun kann es natürlich vorkommen (faktisch wird dieser Fall nicht selten sein), daß der Eichbeobachter zugleich der Untersuchungsleiter ist, der die Beobachter trainiert (d.h. auch: beeinflußt) hat und daher denselben Fehler macht wie die an ihm geeichten Beobachter. Wenn Genauigkeit *definiert* ist als die Übereinstimmung mit diesem konkreten Eichbeobachter, wären in diesem Fall dann tatsächlich die Übereinstimmung (der Beobachter untereinander) und die Genauigkeit (Übereinstimmung der Beobachter mit dem Eichbeobachter) praktisch gleich hoch (obwohl alle Beobachter Fehler machen). Die Beobachterübereinstimmung kann schließlich unverändert hoch bleiben, während sich die Beobachtungsleistung faktisch verändert (fehlerbelasteter wird): Dies ist das Phänomen des sogenannten „consensual observer-drift" (Abschnitt 3.2.3). Hier hilft vor allem der wiederholte Einsatz von unabhängigen Kalibrierungsbeobachtern bzw. allgemeiner: der wiederholte Vergleich mit einem Standard.

Beck (1987, S. 73f.) weist darauf hin, daß die Übereinstimmung im Ergebnis der Beobachtung auch deshalb nicht unbedingt ein Beweis für die gleiche Wahrnehmung sei, weil es denkbar ist, daß die verglichenen Beobachter zwar unterschiedliche, aber einander ausgleichende Fehler gemacht haben. Etwa könne in einem Fall die „Tendenz zur Milde" ein Urteil in der einen Richtung, im anderen die „zentrale Tendenz" das Urteil in der anderen Richtung beeinflußt haben, so daß zwar beide Beobachter zu äußerlich gleichen Ergebnissen kommen, dabei aber („ursprünglich") weit auseinanderliegende Wahrnehmungen gemacht haben. Trainingsmethoden, die nur auf eine möglichst hohe Beobachterübereinstimmung abzielten, verharrten daher „an der Oberfläche eines höchst komplexen Prozesses" (Beck, 1987, S. 73f.). Man müsse vielmehr, folgert Beck, jeden Fehler einzeln und unabhängig kontrollieren. Die Kontrolle der Übereinstimmung in verschiedenen Fällen wird jedoch in aller Regel dann ausreichen, wenn kein Anlaß zu der Vermutung besteht, daß diese verschiedenen Fälle in relevanter Hinsicht gleich waren. Der Einwand, es sei immer noch nicht auszuschließen, daß verschiedene Fehler die eigentlich verschiedenen Wahrnehmungen zufällig derart verzerrt haben, daß gleiche Ergebnisse gleiche Wahrnehmungen vortäuschen, beschreibt dann mehr eine logische als eine naheliegende oder wahrscheinliche Möglichkeit. Es bleibt jedoch festzuhalten, daß

1. die Genauigkeit einer bestimmten Beobachtungsmethode durch den Vergleich verschiedener Beobachter mit derselben Methode nicht sicher ermittelt werden kann, und daß
2. die Beobachterübereinstimmung (der Beobachter untereinander wie die Übereinstimmung mit einem Eichbeobachter) keine hinreichende Bedingung für Fehlerfreiheit der Beobachtung ist.[31]

31 Das Phänomen des „observer-drift" wurde bereits erwähnt, man braucht aber einfach nur an kulturelle Selbstverständlichkeiten oder Vorurteile zu denken, die alle Personen einer Kultur unhinterfragt teilen, um zu verstehen, daß es sich hier durchaus nicht um einen künstlichen oder kleinlichen Einwand handelt.

Wir beenden damit zunächst das Thema Beobachterübereinstimmung. Es ist allerdings offensichtlich, daß wir bis hierher nur ein Übereinstimmungsmaß für den einfachsten Fall - eine Verhaltensklasse und zwei Beobachter - definiert haben. Was machen wir bei mehreren Klassen? Was machen wir in dem Fall einer semantischen Segmentierung (vgl. Kap. 4.2), bei der schon die Einteilung in Beobachtungseinheiten problematisch sein kann? Wie geht man bei anderen als Zeichensystemen vor? Kurzum, während dieses Unterkapitel dazu diente, das prinzipielle Vorgehen zur Schätzung der Beobachterübereinstimmung herauszuarbeiten, werden erst in einem späteren Abschnitt (Kap. 5.5) die Erweiterungen für möglichst viele Anwendungsfälle beschrieben. Zunächst soll aber das Thema „Beobachtungssysteme" vervollständigt werden. Hierzu muß noch einmal das Stichwort „Messung" aufgegriffen werden, da sich durch die Erörterung dieses Begriffs eine Ordnung für die Systeme ergibt.

Literaturempfehlungen

Zum Thema Zeichensysteme und Wahl der Einheit: Faßnacht (1995; ausführlich), von Cranach und Frenz (1969), Fieguth (1977a), Kalbermatten und von Cranach (1981). Das Thema „Operationalisierung" wird ausführlich bei Stegmüller (1974; recht schwierige Lektüre) und Herrmann (1973; besser zur Heranführung geeignet) diskutiert. Probleme der Validität werden am Beispiel der Intelligenz sehr gut bei Jäger (1986) diskutiert; Cronbach und Meehl (1955) ist der klassische Text zur Konstruktvalidität. Zum Einstieg in das Thema Beobachterübereinstimmung eignen sich Bortz und Döring (1995, S. 253f.; hier nur 'κ'); Faßnacht (1995, S. 206ff.); Feger (1983) oder Bakeman und Gottman (1986).

Kapitel 5

Beobachtung als Messung

Wir hatten schon im Einführungskapitel angedeutet, daß wir im Falle der Beobachtung nicht immer den strengen Begriff der Messung benutzen wollen. Dies gilt aus naheliegenden Gründen für die Verbalsysteme und sicherlich zunächst auch für einfache Zeichensysteme. Bevor wir uns den formal strengeren Skalensystemen zuwenden, sollten wir klären, was diesen strengen Begriff der Messung ausmacht und was wir unter einer „Skala" verstehen wollen.

5.1 Messung und Skalierung

Zunächst ist es wichtig, die Grundbegriffe einzuführen, um dann die mit diesen Definitionen verbundenen Probleme der *Repräsentation*, *Eindeutigkeit* und *Bedeutsamkeit* von Meßmodellen darzustellen. Wir stützen uns bei der Darstellung weitgehend auf Gigerenzer (1981), der eine gute Einführung in die Meßtheorie auf der Basis der klassischen Arbeiten etwa von Suppes und Zinnes (1963) gibt.

5.1.1 Homomorphe Abbildung, Messung und Skala

Im Abschnitt 1.4 hatten wir davon gesprochen, daß wir unter Messung nicht nur die Zuordnung von Symbolen (in der Regel Zahlen) zu empirischen Gegebenheiten verstehen wollen, sondern jeweils auch eine Angabe darüber verlangen, welche mathematischen Operationen wir mit den Symbolen durchführen dürfen, so daß deren Ergebnisse wieder auf den empirischen Bereich rückbeziehbar sind. Der erste Schritt, um diese mehr intuitiv verständlichen Bemerkungen zu präzisieren, sollte darin bestehen, den Begriff der „Zuordnung" enger zu fassen. Unter „Zuordnung" kann man vielerlei verstehen; schauen wir uns die Abbildung 31 an.

In beiden Teilabbildungen werden den empirischen Objekten (linke Menge) Symbole (rechte Menge) zugeordnet. Wie man sieht, werden aber in Abbildung 31a manchen Objekten nicht nur ein, sondern gleich zwei Symbole, andererseits einigen Objekten kein Symbol zugeordnet.

Die Abbildung 31a kann etwa als Veranschaulichung des oben eingeführten Zeichensystems von Lovaas et al. (1973) verstanden werden: Die „empirischen Objekte" sind die individuellen Verhaltensweisen pro Zeitabschnitt, die „Symbole" unsere Verhaltensklassen „Selbst-Stimula-

tion", „Echolalie" usw. (s. Kap. 4.1). Es kann sowohl der Fall eintreten, daß wir für eine Zeiteinheit keine Eintragung vornehmen, als auch der Fall, daß wir mehrere Zeichen notieren.

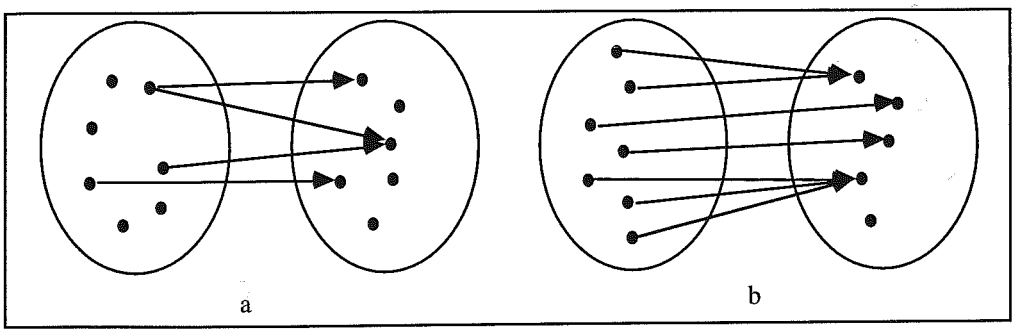


Abbildung 31: Relation und Abbildung

Schauen wir uns dagegen Abbildung 31b an, so stellen wir fest, daß jedem Objekt in der *linken* Menge genau *ein* Symbol in der rechten Menge zugeordnet wird. Während das Beispiel in Abbildung 31a unter den allgemeineren Begriff der Relation fällt, spricht man im Beispiel b von einer spezifischen Relation - der Abbildung oder Funktion. In der weiteren Diskussion der Skalensysteme soll immer von dem strengeren Abbildungsbegriff ausgegangen werden.

Der zweite Schritt zur präzisen Einführung der Messung soll an einem kleinen Gedankenbeispiel deutlich gemacht werden: Der Leser möge sich vorstellen, vor ihm sei ein Tisch mit Weinproben aufgebaut. Die Aufgabe besteht nun darin, für jedes Paar von Weinen anzugeben, welcher der beiden beteiligten „Tropfen" dem Tester besser schmeckt. Wird diese Aufgabe erfüllt, hat der Tester eine *Relation* auf der Menge der beteiligten Weine definiert: Aus der Menge aller Paare (a,b) hat er diejenigen ausgesucht, für die gilt: „Wein a schmeckt mir besser als Wein b". Ebenso können wir den Tester bitten, daß er uns angibt, welche „Paarlinge" für ihn ununterscheidbar sind: Er sucht aus allen Paaren (a,b) diejenigen aus, für die gilt „a und b kann ich nicht unterscheiden". Eine dritte Möglichkeit wäre, daß er für jeweils zwei Paare von Weinen angibt, welches Paar sich „ähnlicher" ist. Während er in den ersten beiden Fällen eine *zweistellige* Relation definiert hat, handelt es sich im zweiten Fall um eine *vierstellige* („a und b sind sich ähnlicher als c und d"). Wenn nun eine oder mehrere solcher Relationen auf einer Menge von empirischen Objekten definierbar sind, sprechen wir von einem *empirischen System* bzw. *empirischen Relativ*, formal:

$$E = <A,Q_1,...,Q_s>$$

Dabei ist A die Menge der Objekte (im Beispiel die Weine), während Q_1 bis Q_s die Relationen sind (im Beispiel wäre Q_1 die Präferenzrelation „a schmeckt mir besser als (oder gleich gut wie) b").

Ein solches Relativ können wir ohne weiteres auch im Bereich der Zahlen definieren. Ziehen wir etwa willkürlich zwei Zahlen aus der Menge der natürlichen Zahlen (dabei kann die gleiche Zahl zweimal gezogen werden), so läßt sich immer eine Relation der Art „a ist größer als (bzw. gleich groß wie) b" definieren. Analog zum Begriff des empirischen Relativs läßt sich ein *numerisches System* bzw. *Relativ* definieren:

$N = <R,P_1,...,P_s>$ [32]

Man sagt nun, daß zwei Systeme vom gleichen *Typ* sind, wenn sie eine gleiche Anzahl von Relationen umfassen und die Stelligkeit der Relationen sich jeweils entspricht (also Q_1 von gleicher Stelligkeit ist wie P_1).

Es liegt nun nahe, die Menge der empirischen Objekte auf die Menge der Zahlen abzubilden. Machen wir uns an dem „Wein"-Beispiel klar, was dies bedeutet: Wir ordnen *jedem* Wein *genau* eine Zahl zu (dabei dürfen wir jedoch verschiedenen Weinen die gleiche Zahl zuordnen). Natürlich gibt es noch unendlich viele Möglichkeiten, dieses zu tun. Schön wäre es natürlich, wenn bei allen Paaren (a,b), die die Relation „a schmeckt mir besser als (oder gleich gut wie) b" definieren, die zugeordneten Zahlen f(a) und f(b) die numerische Relation $f(a) \geq f(b)$ erfüllen. Gilt allgemein, daß die den empirischen Objekten zugeordneten Zahlen immer (den Relationen des empirischen Relativs) entsprechende Relationen im numerischen Relativ definieren, sprechen wir von einer *homomorphen Abbildung* bzw. einem *Homomorphismus*. Und damit sind wir bei dem Begriff der Messung angekommen: Eine homomorphe Abbildung eines empirischen Systems in ein numerisches System heißt *Messung*. Der Begriff der *Skala* meint nun das Gesamt von empirischem und numerischem Relativ sowie den Homomorphismus (vgl. zu diesen Definitionen Gigerenzer, 1981).

5.1.2 Das Repräsentationsproblem

Durchdenkt man einen solchermaßen definierten Meßbegriff, stößt man auf ein fundamentales Problem, welches mit dem Begriff der Relation zusammenhängt: Wir haben oben die Relationen sowohl des empirischen als auch des numerischen Relativs immer mit uns vertrauten Beschreibungen wie „schmeckt mir besser", „ist größer als" versehen; zu der formalen Definition der Relation gehört aber nichts dergleichen: Jede *beliebige* Teilmenge aller Paare von Mengenelementen definiert z.B. eine zweistellige Relation.

Zu welchen Problemen dies führt, kann man sich an dem Wein-Beispiel klarmachen: Angenommen, es gilt, drei Weine a, b und c zu testen und eine Präferenzrelation zu erstellen. Der Tester notiert nun seine Präferenzen in der Form (a,b) mit der Bedeutung „a schmeckt mir besser als b": (a,b), (b,c) und (c,a). Im Klartext: Wein a schmeckt ihm besser als Wein b, Wein

---

32 Mit R ist hier immer die Menge der reellen Zahlen gemeint.

b besser als Wein c, aber Wein c zieht er Wein a vor! Das empirische Relativ läßt sich also so schreiben:

E = < (a,b,c) , ((a,b),(b,c),(c,a)) >

Um von Messung nach unserer obigen Definition zu sprechen, brauchen wir nun ein numerisches Relativ vom gleichen Typ, so daß wir eine homomorphe Abbildung der Menge der Weine in die Menge der Zahlen definieren können. Wie wäre es mit N = < R, ((1,2),(2,3),(3,1)) > und den Zuordnungen a → 1, b → 2 und c → 3? Unsere Anforderungen an eine Messung sind erfüllt, wie man sich überzeugen kann: Es wurde eine Zahlenmenge R und eine Relation P (= ((1,2), (2,3), (3,1))) auf dieser Zahlenmenge definiert; die Abbildung erfüllt die Kriterien einer homomorphen Abbildung. Fazit scheint zu sein: Die Bedingungen, die wir für eine Messung fordern, sind offensichtlich in jedem Fall ganz simpel herzustellen!

Das Problem besteht nun darin, daß die angegebene numerische Relation vollkommen willkürlich gewählt ist; keine der uns vertrauten zweistelligen Relationen (wie „=", „>", „<") paßt zu ihr, vor allem nicht die von uns insgeheim angestrebte „Größer"-Relation. Kurz gesagt: In diesem Fall haben wir keinerlei Gewinn durch die Abbildung auf Zahlen! Hätte unser Tester dagegen den Wein a dem Wein c vorgezogen (Entsprechung in der numerischen Relation: (1,3) statt (3,1)), wäre diese Ordnungsrelation verwirklicht gewesen. Man muß also einschränkend sagen, daß wir von Messung nur dann sprechen wollen, wenn die Relationen im numerischen Relativ „einfach" bzw. „vertraut" sind; Gigerenzer (1981) spricht hier von einem „pragmatischen Grundprinzip" (S. 47). Bei jeder Messung wird versucht, „ob ein empirisches System auf ein *geeignet gewähltes* numerisches System homomorph abgebildet werden kann" (S. 47, Hervorhebung im Original); dieses Problem wird als das sogenannte „Repräsentationsproblem" der Messung bezeichnet. In unserem Beispiel würden wir etwa die Frage stellen, ob die Präferenzen des Testers durch das numerische System <R,'>'> abgebildet werden können.

Wir können nun fragen, welche Bedingungen die Präferenzurteile des Testers erfüllen müssen, damit eine homomorphe Abbildung in das numerische System <R,'>'> möglich ist. Es kann gezeigt werden, daß es in einem solchen Fall genügt, wenn die Bedingungen der *Konnexität* und *Transitivität* im empirischen Relativ erfüllt sind. Dabei meint die Eigenschaft der *Konnexität*, daß die empirische Relation vollständig ist; d.h. im Beispiel: Jedes Paar von Weinen läßt sich hinsichtlich der Präferenz ordnen. *Transitivität* ist dadurch definiert, daß die Präferenz „übertragbar" ist. Im Beispiel: Wenn Wein a Wein b und Wein b Wein c vorgezogen wird, dann muß Wein a Wein c vorgezogen werden. Erfüllen also die Urteile des Testers die Anforderungen der Konnexität und Transitivität, gibt es in jedem Fall eine homomorphe Abbildung in <R,'>'>, und die natürliche Ordnung der Zahlen spiegelt eine Präferenzhierarchie des Testers wider.

5.1.3 Das Eindeutigkeitsproblem

Das nächste Problem, mit dem man sich im Zusammenhang der Messung beschäftigen muß, ist das sogenannte „Eindeutigkeitsproblem". Wieder kann man sich das Problem an unserem kleinen Beispiel deutlich machen: Nehmen wir an, die Präferenzurteile des Weintesters genügen den Bedingungen der Konnexität und Transitivität; es ist also ein Homomorphismus möglich. Oben haben wir die Abbildung

$a \rightarrow 1$
$b \rightarrow 2$
$c \rightarrow 3$

gewählt. Es hält uns natürlich nichts davon ab, beliebige andere Zahlen zu wählen, solange die Kriterien einer homomorphen Abbildung erfüllt sind. Wir können noch einen Schritt weitergehen: Abbildungen sind nicht nur für die Zuordnung von Zahlen zu empirischen Objekten gedacht; ebenso können wir Abbildungen von Zahlen auf Zahlen definieren. Bleiben wir bei der obigen Zuordnung, können wir eine Funktion definieren, die den Zahlen 1, 2, 3 wiederum andere Zahlen zuordnet:

$1 \rightarrow 11$
$2 \rightarrow 12$
$3 \rightarrow 13$

Die Frage ist nun, ob nach dieser *Transformation* weiterhin ein Homomorphismus bzgl. des empirischen Systems bestehen blieb. Im Beispiel ist dies klar mit „ja" zu beantworten. Eine andere Funktion wäre aber:

$1 \rightarrow 0$
$2 \rightarrow 0$
$3 \rightarrow 1$

In diesem Fall würde die neue Zuordnung keinen Homomorphismus bzgl. E ergeben, da nicht gilt: $0 > 0$. Das Eindeutigkeitsproblem ist also dadurch gekennzeichnet, daß wir für jeden Homomorphismus angeben müssen, welche Transformationen hinsichtlich einer gegebenen numerischen Zuordnung f erlaubt sind, d.h. den Homomorphismus erhalten (vgl. Gigerenzer, 1981).

5.1.4 Das Bedeutsamkeitsproblem

Letztendlich ist unser Ziel, mathematische Operationen auf der Menge der zugeordneten Zahlen anzuwenden: Es gilt, Meßwerte zu aggregieren, statistische Parameter zu berechnen und vor allem Meßwerte entsprechend unseren hypothetischen Vorstellungen in Beziehung zu setzen. Wir hatten oben darauf hingewiesen, daß bei der Anwendung der Operationen jedesmal die Frage zu stellen ist, ob das Ergebnis noch sinnvoll auf das empirische Relativ rückzubeziehen ist.

Diese Aussage kann jetzt etwas genauer gefaßt werden: Wir müssen uns jedesmal überlegen, ob das Ergebnis einer Operation über der Menge der für den Homomorphismus zulässigen Transformationen *invariant* bleibt.

Was heißt das? Nehmen wir noch einmal das Beispiel des Weintestens: Wir bitten einige Tester, eine Auswahl von acht Weinen zu beurteilen. Wir gehen davon aus, die Präferenzwahlen seien konnex und transitiv. Wir können also eine homomorphe Abbildung in das numerische Relativ N <R,'>'> definieren. Um den Testern jeden Paarvergleich zu ersparen, bitten wir sie, die acht Weine in eine Rangordnung zu bringen; wir ordnen den Rangplätzen dann die Zahlen 1 bis 8 zu. Neun Tester haben ihre Urteile abgegeben; wir erhalten also nun für jeden Wein neun Werte und können prinzipiell die Maße der zentralen Tendenz *Medianwert* bzw. *arithmetisches Mittel*_bilden (der Median ist der mittlere Wert, wenn eine Meßreihe geordnet vorliegt; das arithmetische Mittel ist schlicht der alltagssprachliche „Durchschnitt": also alle Werte aufaddiert und durch die Anzahl der Werte geteilt; vgl. Bortz, 1993). Der „Rolzenheimer Bübchenberg" sei mit 6,6,5,8,4,4,2,5,5 eingestuft worden; Median und arithmetisches Mittel ergeben beide den Wert 5. Die zulässigen Transformationen sind bei der Präferenzrelation allgemein durch die Klasse der streng monoton steigenden Funktionen gegeben (wenn für zwei Werte a, b gilt: a > b, dann sollte auch für die transformierten Werte a', b' gelten: a' > b'). Eine solche Funktion ist etwa die quadratische Funktion $f(x) := x^2$; wenden wir sie auf unsere Meßwerte an und berechnen erneut Median und arithmetisches Mittel, so erhalten wir: Median = 25 und arithmetisches Mittel = 27.44. *Invarianz* des Kennwertes bedeutet nun, daß das Ergebnis einer Anwendung der Transformation auf den alten Kennwert identisch mit demjenigen Parameter ist, der aufgrund der transformierten Kennwerte zustande kam. In dem Beispiel gilt dies für den Median, aber nicht für das arithmetische Mittel; letzteres darf also nicht interpretiert werden! Die Frage der Interpretierbarkeit von Operationsergebnissen definiert das Bedeutsamkeitsproblem.

5.1.5 Die Skalentypen

Glücklicherweise gibt es nur wenige einfache und vertraute Relationen und dazu korrespondierende Klassen von Funktionen, so daß mathematische Operationen danach eingeteilt werden können, für welche Skalentypen sie geeignet sind (die nachfolgenden Skalenbegriffe gehen auf Stevens, 1951, zurück). Von einer *Nominalskala* sprechen wir, wenn wir lediglich Äquivalenzklassen bilden. Ein typisches Beispiel sind biologische Klassifikationen: Dieses Tier gehört zu den „Fischen", jenes zu den „Säugetieren". Eine homomorphe Abbildung erhalten wir, wenn wir als numerische Relation die '='-Relation wählen. Im Sinne des Eindeutigkeitsproblems sind alle eineindeutigen Transformationen zugelassen, d.h. alle Funktionen, die jedem einzelnen Wert wieder genau einen einzelnen Wert zuordnen. Es ist beliebig, welche Werte man den biologischen Klassen zuordnet, solange nicht „Fisch" und „Säugetier" dieselbe Zahl erhalten.

Im Sinne des Bedeutsamkeitsproblems bleiben etwa Häufigkeitsaussagen und der Modus (d.h. die am stärksten besetzte Kategorie) erhalten.

Die *Ordinalskala* repräsentiert eine Rangfolge der empirischen Objekte, etwa die von uns als Beispiel gewählte Präferenzrelation des Weintesters. Erlaubt sind hierbei alle monoton steigenden Funktionen, die die Rangfolge der Zahlenwerte des numerischen Relativs beibehält. Dabei ist es unerheblich, ob der Abstand zwischen zwei benachbarten Werten erhalten bleibt, da dieses Merkmal nicht durch die empirischen Relationen „gedeckt" wird. Dementsprechend bleibt hier - wie in unserem Beispiel im Abschnitt 5.1.4 näher ausgeführt - der Median, aber nicht das arithmetische Mittel erhalten. Viele psychologische Phänomene lassen sich durch solche Ordnungsrelationen abbilden.

Die *Intervallskala* repräsentiert gleiche Abstände zwischen in einer Ordnungsrelation benachbarten Objekten. In den Begriffen unserer Skalierungsbemühungen im Abschnitt 5.1 brauchen wir für eine Intervallskala mehrstellige Relationen im empirischen Relativ. Als Beispiel kann man etwa die Temperaturmessung nennen: Der physikalische Unterschied zwischen 10° und 20° ist genauso groß wie der Unterschied zwischen 60° und 70°. Alle linearen Funktionen ($y = ax + b$) stellen zulässige Transformationen dar, so daß Statistiken wie arithmetisches Mittel nach Transformationen erhalten bleiben. Hieraus ergibt sich auch die große Beliebtheit der Intervallskala in der psychologischen Forschung: Verfahren der Datenauswertung, vor allem inferenzstatistische Methoden, besitzen Gültigkeit für ein spezifisches Datenniveau. Die vielseitigsten Methoden sind dabei sicherlich für das Intervallniveau definiert, obwohl auch auf Nominal- oder Ordinalniveau Verfahren ständig weiterentwickelt werden. Es ist viel darüber geschrieben worden, ob bei sozialwissenschaftlichen Fragestellungen Intervallqualität der Daten angenommen werden kann. Die Einschätzung von Bortz (1993, Kap. 1.1) ist sicher treffend, daß typische Meßoperationen in der Psychologie häufig Datenqualitäten liefern, die zwischen Ordinal- und Intervallniveau anzusiedeln sind. Das heißt, daß wir für weite Bereiche der Skala Intervallqualität annehmen können, in Grenzbereichen jedoch lediglich Ordinalniveau anzutreffen ist. Wir befinden uns in einem Dilemma: Verwenden wir nun Verfahren, die für Ordinaldaten konstruiert wurden, begehen wir möglicherweise den Fehler, nicht die Methode zu nutzen, die uns am besten darüber aufklärt, ob ein bestimmter Zusammenhang gegeben ist oder nicht. Rechnen wir einen Test für Intervalldaten, müssen wir uns vorwerfen, daß unsere Daten nicht diese Qualität besitzen. Man kann in diesem Zusammenhang die Äußerung von Elashoff und Snow (1971/1972) zitieren, die dafür plädieren, stets zu überprüfen, ob es alternative Zugangswege zur Datenauswertung gibt. Der Forscher bekommt ein sichereres Gefühl bei seinen Ergebnissen, wenn verschiedene Methoden, für die man in einem konkreten Fall argumentieren kann, zu konvergenten Ergebnissen führen.

Schließlich soll noch die *Verhältnisskala* erwähnt werden. Bei ihr besitzt der Nullpunkt absolute Bedeutung: Eine Länge von Null Metern, eine Häufigkeit von Null Einheiten sind gute Illustrationsbeispiele. Bei einer Verhältnisskala sind nur Streckungs- oder Stauchungstransformationen erlaubt ($y = ax$), so daß hier erstmals die Rede von "a ist doppelt so groß (lang, häu-

fig ...) wie b" Sinn macht. Diese Qualitäten werden allerdings in der Psychologie selten erreicht, so daß dieser Skalentyp nicht so bedeutsam ist.

Was hat dies alles mit Beobachtung zu tun? Zunächst ist unser etwas „hemdsärmelig" eingeführter Begriff des Zeichensystems daraufhin zu überprüfen, warum wir in diesem Fall nicht von Messung sprechen können. Daraufhin ist das Kategoriensystem einzuführen, für das in der Regel Nominalniveau beansprucht wird (Abschnitt 5.2, ein bekanntes Beispiel dafür stellen wir in Abschnitt 5.3 vor). Schließlich ist es unumgänglich, noch eine deutlich andere Variante von Beobachtungssystemen einzuführen, die sogenannte Einschätz- oder Ratingskala (Abschnitt 5.4). Eine solche Ratingskala ist der prototypische Fall psychologischer Messung, bei dem man nicht weiß: ist's Ordinal- oder doch Intervallniveau?

5.2 Einführung der Kategoriensysteme

Die Datenerhebung mittels eines Zeichensystems entspricht nicht dem Begriff der Messung, weil erlaubt ist, daß (a) einem empirischen Objekt (einer Einheit) mehrere Zeichen und (b) manchen Einheiten keine Zeichen zugeordnet werden dürfen. Der Sprung von einem Zeichen- zu einem Kategoriensystem ist nun denkbar einfach: Wir definieren unsere Klassen lediglich so, daß die Liste erstens vollständig ist (*jeder* Einheit wird eine Klassen zugeordnet) und zweitens die Klassen exklusiv formuliert sind (jeder Einheit wird nur eine Klassen zugeteilt). Damit entsprechen diese Kategoriensysteme exakt der soeben eingeführten *Nominalskala* und fallen damit - im Gegensatz zu den Zeichensystemen - unter den strengen Begriff der „Beobachtung als Messung".

Für die Zeichensysteme können wir in unserer durch den Meßbegriff erweiterten Terminologie nun sagen, daß sie sich aus mehreren, in einem nicht näher explizierten Verhältnis zueinander stehenden Kategoriensystemen zusammensetzen: Jedes Zeichen kann als ein kleines Kategoriensystem mit den beiden Kategorien 'Zeichenverhalten wird beobachtet' und 'Zeichenverhalten wird nicht beobachtet' aufgefaßt werden. Diese Art der Betrachtung hat für auswertungstechnische Zusammenhänge und die Bestimmung der Beobachterübereinstimmung Bedeutung, wie wir in Abschnitt 5.5 noch sehen werden. Es besteht eine gewisse Tendenz, Kategoriensysteme anstelle von Zeichensystemen zu konstruieren. Dabei spielen Fragen der Auswertbarkeit, aber vor allem der Eindeutigkeit eine große Rolle. Von den 26 Beobachtungsverfahren, die Manns et al. (1987) zitieren, werden von den Autoren 25 als Kategoriensysteme eingestuft.

Wie stehen Kategoriensysteme zu unserer Einteilung nach Segmentierungsformen einerseits und der praktischen Vorgehensweise (Sortier- versus Detektorverfahren) andererseits? Die Frage, welche Segmentierungsform bei einem konkreten Beobachtungssystem vorliegt, ist unabhängig von der Klassifizierung als Zeichen- oder Kategoriensystem: Formale und semantische Segmentierung lassen sich mit beiden Arten von Beobachtungssystemen vereinbaren. Etwas anders liegt der Fall bei der Einteilung hinsichtlich des praktischen Vorgehens. Streng genom-

men kann ein Detektor-System nur ein Zeichensystem sein, da nicht jedwede Einheit einem Zeichen zugeordnet wird. Sind die Zeichen jedoch einander ausschließend definiert und fällt somit der wichtigste Unterschied zwischen Zeichen- und Kategoriensystem in dem konkreten Fall weg, kann durch nachträgliches Einführen der Restkategorie für auswertungstechnische Zusammenhänge auch ein Detektor-Vorgehen mit einem Kategoriensystem einhergehen. Insofern ist die Unterscheidung Zeichen- versus Kategoriensystem als unabhängig von den beiden Klassifikationsgesichtspunkten anzusehen.

Wie steht der Begriff des Kategoriensystems zu den Skalenniveaus? Wir hatten oben das Niveau der Nominalskala zugeordnet. In fast allen Fällen entspricht dies genau der Konstruktion und Formulierung des Systems. Es ist jedoch nicht ausgeschlossen, daß Kategorien definiert werden, die hinsichtlich eines Konstruktes eine Rangreihe bilden. Wir denken etwa an Verhaltensindikatoren solcher Konstrukte wie „Aggressivität", „Angst" etc. Wichtig ist, daß sich die Ordnung der Kategorien aus theoretischen Erwägungen ergibt.

Ehrhardt, Findeisen, Marinello und Reinartz-Wenzel (1981) konstruierten ein Kategoriensystem zur Beobachtung von Aufmerksamkeit im Unterricht. Tatsächlich „operationalisierten" (vgl. Abschnitt 4.4) die Autoren das Aufmerksamkeitsverhalten über drei *Zeichen*, die hier kurz wiedergegeben werden sollen (Erhardt et al., 1981, S. 284):

1. Blickrichtung; blickt zum Unterrichtsmittelpunkt (vs. blickt woandershin);
2. Körperhaltung und Körperausdruck; ausgerichtet auf Unterrichtsmittelpunkt und angespannt (vs. abgewandt, erschlafft);
3. Tätigkeit; übt die für die Aufgabe notwendige Tätigkeit aus (vs. tut nebenher etwas anderes).

Es handelt sich also zunächst um ein Zeichensystem, da die Zeichen sich nicht gegenseitig ausschließen. Die Daten wurden im Zeittaktverfahren (10-Sekunden-Intervalle) erhoben. Allerdings - eine Besonderheit dieser Studie - sollten die Beobachter nicht notieren, welche der drei Zeichen in jedem Intervall erfüllt waren, sondern lediglich, *wieviele* dieser Verhaltensweisen simultan zu sehen waren. Korrekte Notierungen waren also die Zahlen 0, 1, 2 oder 3. Da sich diese Angaben ausschließen und keine weitere Angabe möglich ist, handelt es sich nun um ein Kategoriensystem, dessen Kategorien eine Rangreihe bilden. Im Abschnitt 5.5 wird die Bedeutung dieser Überlegungen für die Bestimmung der Beobachterübereinstimmung deutlich werden.

Im folgenden Abschnitt wollen wir mit der sogenannten Interaktionsprozeßanalyse von Bales (1950a) eines der bekanntesten Kategoriensysteme vorstellen. Wir werden dieses Beispiel etwas ausführlicher diskutieren, weil hier sehr schön zu sehen ist, wie ein handhabbares Beobachtungssystem theoriebezogen konstruiert wurde.

5.3 Das bekannteste Beispiel: Die Interaktionsprozeßanalyse nach Bales

Die Interaktionsprozeßanalyse (IPA) von Robert F. Bales (1950a) stellt wohl das bekannteste Beispiel für ein Kategoriensystem unter den Beobachtungsmethoden dar, das in fast allen Arbeiten, die einen Überblick zum Thema Verhaltensbeobachtung geben, dargestellt (z.B. von Cranach & Frenz, 1969; Faßnacht, 1995; Grümer, 1974; Koeck & Strube, 1977; Manz, 1974; Martin & Wawrinowski, 1991) oder zumindest erwähnt wird (z.B. Graumann, 1966; Hasemann, 1983; Schaller, 1980). Es handelt sich um ein Verfahren zur Erfassung sozialen und emotionalen Verhaltens von Individuen in Kleingruppen. Untersucht werden Problemlösungsversuche, Rollen und Statusstrukturen sowie die Veränderung dieser Variablen über die Zeit (Bales, 1968). Ausgangspunkt für die Entwicklung des Beobachtungssystems war die Annahme, daß sich in jeder Interaktion Strukturen finden lassen, die - unabhängig von spezifischen situativen Randbedingungen - typisch für soziale Beziehungen in Gruppen sind. Die theoretisch bedeutsamen Merkmale und Kennzeichen der Interaktion im Gruppenprozeß sind dementsprechend unabhängig vom Thema oder von der Aufgabe einer Gruppe zu erfassen. In diesem Sinne handelt es sich bei der IPA um ein unspezifisches Beobachtungsverfahren mit einem breiten Anwendungsbereich, was die Beliebtheit der Anwendung dieses Beobachtungssystems erklärt. Für die Entwicklung der Kleingruppenforschung war die IPA von großer Bedeutung, da es sich hier um den ersten umfassenden Ansatz zur direkten Beobachtung von Gruppeninteraktionen handelt (Scharpf, 1988). Die folgende Darstellung hat *nicht* zum Ziel, den aktuellen Stand der Forschung zur Interaktion in Kleingruppen wiederzugeben. Vielmehr soll in groben Zügen die theoriebezogene Entwicklung eines Kategoriensystems und dessen Handhabung dargestellt werden. Zu diesem Zweck bietet sich die IPA von Bales als ein klassisches Beispiel an.

(1) Der theoretische Ansatz

Die IPA baut auf dem funktionalistischen Ansatz in der Soziologie auf (Merkens & Seiler, 1978). Die Grundannahme besagt, daß soziale Systeme generell die Tendenz haben, ein dynamisches Gleichgewicht zu erreichen. Dazu müssen ständig zwei einander entgegengesetzte Arten von Anpassungsleistungen erbracht werden: einerseits Anpassung des Systems nach außen, andererseits Integration nach innen.

Die Anpassung eines Systems an Umweltbedingungen erfordert die Bewältigung verschiedener Aufgaben. In der Gruppe führt dies unter anderem zu Arbeitsteilung und in der Folge zu Statusunterschieden. Die funktionelle Aufgabenorientiertheit gefährdet damit den Zusammenhalt, so daß eine Anpassungsleistung in Richtung einer Integration nach innen notwendig wird. Sie wiederum zielt auf die Gleichheit der Gruppenmitglieder ab und gefährdet somit die Aufgabenorientiertheit. Das System strebt eine Balance zwischen optimaler Anpassung an Umweltbedingungen und optimaler innerer Integration an. Dieser Zustand wird „Equilibrium" genannt.

Ein Interaktionsprozeß wird immer dann in Gang gesetzt, wenn der Gleichgewichtszustand vorübergehend aufgehoben wird. Störungen des Gleichgewichts kommen zustande durch Aktionen, also neue Ideen, Meinungen und Vorschläge. Darauf folgen Reaktionen in Form von Feedback, bis die Störung hinreichend behoben ist.

Neben diesen Anpassungsleistungen geht Bales davon aus, daß in jedem Interaktionssystem Kommunikationsprobleme auftreten. Dabei werden sechs bipolare Dimensionen unterschieden, die den positiven, förderlichen bzw. den negativen, behindernden Umgang im jeweiligen Kommunikationsbereich zeigen. In jeder Gruppe vollziehen sich Prozesse der Orientierung (Wie ist die Lage?), Bewertung (Welche Einstellung zur Situation?) und Kontrolle (Was ist zu tun?). Auf diese Prozesse folgen immer Prozesse der Entscheidung, Spannungsbewältigung und Integration.

(2) Das Kategoriensystem

Die genannten Anpassungsleistungen und Kommunikationsprobleme bildeten die Grundlage für die Entwicklung des Kategoriensystems. Nach mehreren Revisionen mit Kategorienzahlen zwischen 5 und 85(!) entstand ein Beobachtungsschema, das als Kompromiß zwischen der Forderung nach theoretischer Differenzierung einerseits und der Praktikabilität in der Durchführung andererseits angesehen werden kann. Das endgültige Kategoriensystem besteht aus 12 Kategorien (Abb. 32), die nach verschiedenen Gesichtspunkten gruppiert und in mehrfacher Hinsicht aufeinander bezogen werden. Die Hälfte der Kategorien (1-3 und 10-12) wird dem sozial-emotionalen Bereich zugeordnet, bei dem es um die Integration der Gruppe nach innen geht. Die Kategorien 4-9 repräsentieren den Aufgabenbereich, der der Anpassung nach außen zuzuordnen ist. Von innen nach außen sind jeweils zwei Kategorien paarweise aufeinander bezogen. Der gemeinsame Oberbegriff für jedes Kategorienpaar bezeichnet das dazugehörige Kommunikationsproblem (a-f). Jedes dieser Paare wird von einer emotional positiven und einer emotional negativen Kategorie (sozial-emotionaler Bereich) bzw. einer Frage- und einer Antwortkategorie (Bereich Aufgaben) gebildet.

Es wird davon ausgegangen, daß jede Handlung genau einer der 12 Kategorien zuzuordnen ist, d. h. die Kategorien sind erschöpfend und exklusiv. Jede Kategorie wird über eine Reihe von einzelnen beobachtbaren Verhaltensweisen beschrieben. Besonders in den Kategorien des sozial-emotionalen Bereichs kommen häufiger auch nonverbale Verhaltensweisen (Gestik, Mimik und Betonung) als Kategorieninhalte vor. Eine vollständige inhaltliche Darstellung der einzelnen Kategorien findet sich bei Bales (1950a).

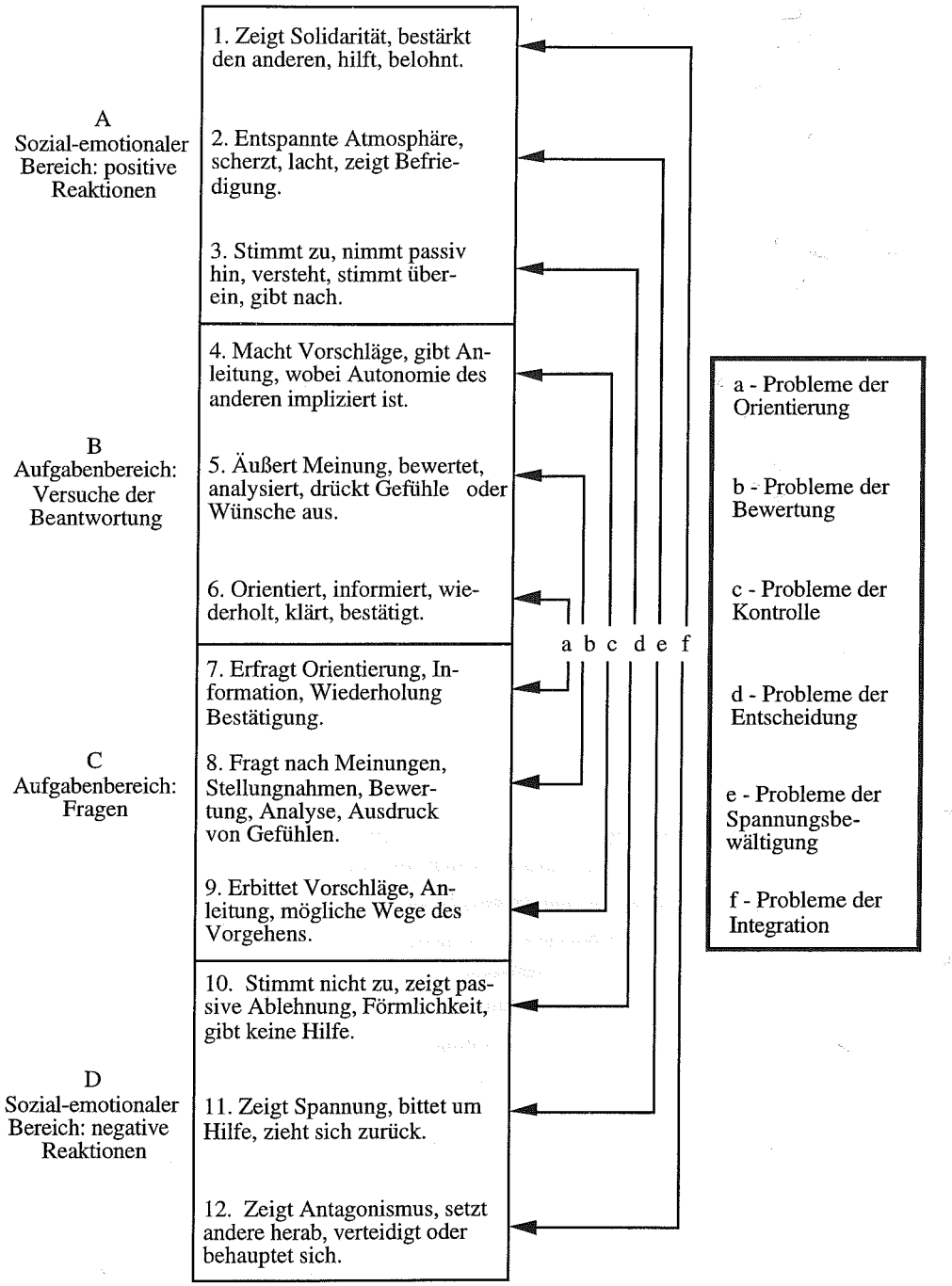


Abb. 32: Das Kategoriensystem nach Bales

(3) Die Registrierung

Wichtig ist, daß die Kategorien nicht dazu dienen, den *Inhalt*, sondern die *Art* der Interaktion wiederzugeben. Entsprechend sind die Kategorien denn auch als relativ abstrakte Verhaltensklassen definiert. Dies führt zum Problem der Wahl einer Beobachtungseinheit und damit zur Frage, wieviel Verhalten „auf einmal" kodiert werden soll (vgl. Absatz 4.2). Bales (1968, S. 158) nennt als Kriterium für den Übergang von einer Beobachtungseinheit zur nächsten „einen *Bedeutungswechsel* innerhalb eines Systems von Symbolen, die der Mitteilung dienen" (Hervorhebung G/W). Wie kann dieses Kriterium in einer aktuellen Beobachtungssituation vom Beobachter umgesetzt werden?

Selbstverständlich beginnt eine neue Beobachtungseinheit dann, wenn die handelnde Person wechselt, d.h. wenn in einer Gruppe ein anderes Gruppenmitglied das Wort ergreift. Viel häufiger wird es jedoch vorkommen, daß eine Person mehrere Äußerungen mit unterschiedlicher Bedeutung nacheinander macht. So werden beispielsweise bei folgender Äußerung: „Ich finde dieses Beispiel sehr gelungen. Ich würde gerne wissen, wie du es findest. Solltest du es nicht gut finden, dann ..." drei verschiedene Aspekte mitgeteilt. Entsprechend sind auch drei verschiedene Einheiten (1. Satz - Kategorie 5, 2. Satz - Kategorie 8, 3. Satz - Kategorie 12) einzutragen.

In der Regel handelt es sich bei der Beobachtungseinheit um einen einfachen Satz. Daraus ergibt sich, daß die Anzahl der Eintragungen der Dauer der Sprechzeit etwa proportional ist. Bales (1968) berichtet von durchschnittlich etwa 10 bis 20 Eintragungen pro Minute. Dies mag als Hinweis dafür angesehen werden, wie schwierig es für den Beobachter ist, die z.T. sehr schnell wechselnden Beobachtungseinheiten zu registrieren und auf dem Beobachtungsbogen zu vermerken.

Wenn wir das IPA-System auf unsere im letzten Kapitel entwickelten Unterscheidungen beziehen (Abschnitt 4.2 und 4.3), so handelt es sich um ein Sortier-Verfahren mit semantischer Segmentierung. Der Beobachter ist ständig aufgefordert, das Geschehen auf *Bedeutungswechsel* zu überprüfen und dann einer Kategorie zuzuordnen.

Mit jedem Protokolleintrag werden drei verschiedene Aspekte gleichzeitig kodiert:

1. eine qualitative Einstufung der Handlung durch die Zuordnung zu einer Kategorie,
2. die Benennung der Person, die die Handlung initiiert (originator), und
3. die Benennung der Person, an die die Handlung gerichtet ist (target).

Vor der Kodierung werden die Gruppenmitglieder durchnumeriert und sodann jede Eintragung in der Reihenfolge Kategorie – initiierende Person – Zielperson vorgenommen. Sofern es für die Auswertung von Interesse ist, kann als vierter Aspekt die Stellung einer Handlung im zeitlichen Verlauf registriert werden. Dies geschieht mit Hilfe eines Ereignisschreibers (vgl. Abschnitt 4.2), bei Bales und Gerbrands (1948) „Interaction-Recorder" genannt. Erst die Verwendung des Ereignisschreibers ermöglicht die Berechnung der Beobachterübereinstimmung, da bei jeder Kodierung eindeutig ist, auf welches Zeitintervall sie sich bezieht (vgl. Abschnitt

4.5). Mit der Einführung von Videogeräten dürfte dieses Gerät allerdings zunehmend weniger Verwendung finden.

Zur Vermeidung von Kodierungsschwierigkeiten hat Bales (1950a, S. 91f) für den Beobachter zwei Regeln formuliert, die bei der Anwendung des Kategoriensystems zu beachten sind.

1. Betrachte jeden Akt als eine Antwort auf den unmittelbar vorhergegangenen oder in seiner Wirkung auf den unmittelbar folgenden.

Diese Regel soll den Beobachter von Interpretationen abhalten, die nicht unmittelbar aus dem Kontext ersichtlich bzw. zu erschließen sind.

2. Bevorzuge die Kategorie, die weiter von der Mitte entfernt liegt. Ordne den Akt in eine Kategorie, die näher am Anfang oder am Ende der Liste steht.

Damit wird versucht, einer möglichen Tendenz von Beobachtern entgegenzusteuern, Verhalten vorzugsweise den mittleren, aufgabenorientierten Kategorien zuzuordnen und die sozialemotionalen Kategorien insbesondere bei negativen, sozial unerwünschten Verhaltensweisen wegen ihres affektiven Gehalts eher zu meiden.

(4) Die Auswertung

Trolldenier (1985) faßt die Auswertungsmöglichkeiten und Darstellungsformen für die mit dem Kategoriensystem gewonnenen Daten in fünf Gruppen zusammen. In *Interaktions-Profilen* wird die Häufigkeitsverteilung der Kategorien für einzelne Personen oder für eine ganze Gruppe erfaßt.

Tabelle 3: Durchschnittliche Häufigkeit von Kategorienbesetzungen in Prozent (Angaben aus Bales 1950b, S. 262)

Kategorien-Nr.	Kurzbezeichnung	Häufigkeit (%)
1	Zeigt Solidarität	1.0
2	Entspannte Atmosphäre	7.3
3	Stimmt zu	12.2
4	Macht Vorschläge	5.2
5	Äußert Meinung	30.0
6	Orientiert	21.2
7	Erfragt Orientierung	5.4
8	Fragt nach Meinung	3.5
9	Erbittet Vorschläge	0.8
10	Stimmt nicht zu	6.6
11	Zeigt Spannung	4.4
12	Zeigt Antagonismus	2.4

Aus der Beobachtung verschiedener Diskussionen in Seminaren, Familien, Ehepaaren, Kommitees, Spielgruppen sowie anderer formeller und informeller Gruppen kommt Bales (1950b) zu dem in Tabelle 3 dargestellten Durchschnittsprofil in bezug auf die Besetzung der Katego-

rien. Eine zweite Darstellungsform betrifft die *„Interaktionshäufigkeit in bezug auf Richtungen"*, mit der die Aktivität einzelner Gruppenmitglieder quantitativ erfaßt werden kann. Dazu eignet sich eine sogenannte Wer-zu-wem-Matrix, im Original (Bales 1950a, S. 89) „scoring who-to-whom" genannt. Die einzelnen Kategorien bleiben hierbei unberücksichtigt. Die *Kombination von Kategorienbesetzung und Interaktionsrichtung* ist die häufigste Art der Darstellung. In graphischen Darstellungen wird die Auftretenshäufigkeit einer oder mehrerer Kategorien über die Beobachtungszeit hinweg beschrieben. Die Kodierung wird in diesem Fall durch die Verwendung des bereits erwähnten Ereignisschreibers ermöglicht. Das beschriebene Vorgehen dient der Untersuchung von *Verläufen in Zeitabschnitten*. Schließlich geht es bei der Untersuchung von *Kategoriensequenzen* um die Beantwortung von Fragen derart: „Welche Kategorie kommt vor welcher und wie oft?" oder „Welche Kategorie folgt welcher und wie oft?" Hier versucht man also, zu Erkenntnissen hinsichtlich der Bedingtheit in der Abfolge von Kategorien zu kommen und dabei auftretende typische Muster zu identifizieren.

(5) Kritische Bemerkungen

Die Kritik an der Interaktionsprozeßanalyse richtet sich vor allem gegen den Universalitätsanspruch des Balesschen Kategoriensystems (vgl. Manz, 1974; Grümer, 1974). Die Tatsache, daß es sich um ein unspezifisches Verfahren zur Analyse von Interaktionsprozessen handelt, läßt vor allem dort Anwendungsgrenzen erwarten, wo spezifische Fragestellungen eine „feinere" Differenzierung von Verhalten erfordern, als dies mit den relativ abstrakten Verhaltensklassen des Kategoriensystems von Bales möglich ist (z.B. bei der Beobachtung und Analyse von Mutter-Kind-Interaktionen; vgl. Innerhofer, 1977; Innerhofer & Peterander, 1981). Aus diesem Grunde warnt Manz (1974, S. 49) vor dem falschen Eindruck, das Balessche Kategoriensystem sei „losgelöst von seinem theoretischen Hintergrund das ideale und universell brauchbare Beobachtungsinstrument zum Studium der Interaktionsprozesse".

Bales stützt seine Ergebnisse im wesentlichen auf Laboruntersuchungen: Es wurden ad hoc Diskussionsgruppen ohne explizite Rollen gebildet, die sich in einem festgesetzten Zeitrahmen mit einem vorgegebenen Problem zu beschäftigen hatten, dessen Lösung für sie ohne praktische Konsequenz war. Diese Bedingungen sind sicherlich nicht so ohne weiteres auf die Realität übertragbar. Einen weiteren Ansatzpunkt für Kritik bietet das bereits angesprochene Problem der Wahl von Beobachtungseinheiten. Kodierungsprobleme aufgrund des Abstraktionsgrads einzelner Kategorien dürften ergänzt werden durch „Fehler zu Lasten des Beobachters" (vgl. Abschnitt 3.2.3). Insgesamt soll mit diesen kritischen Bemerkungen der Wert des Kategoriensystems von Bales keineswegs geschmälert werden. Bis in die jüngste Vergangenheit hinein erfreut sich die IPA großer Beliebtheit. Trolldenier (1985) gibt einen umfassenden Überblick über die Erprobungen des Beobachtungsschemas durch Bales selbst sowie über eine Vielzahl von Modifikationen.

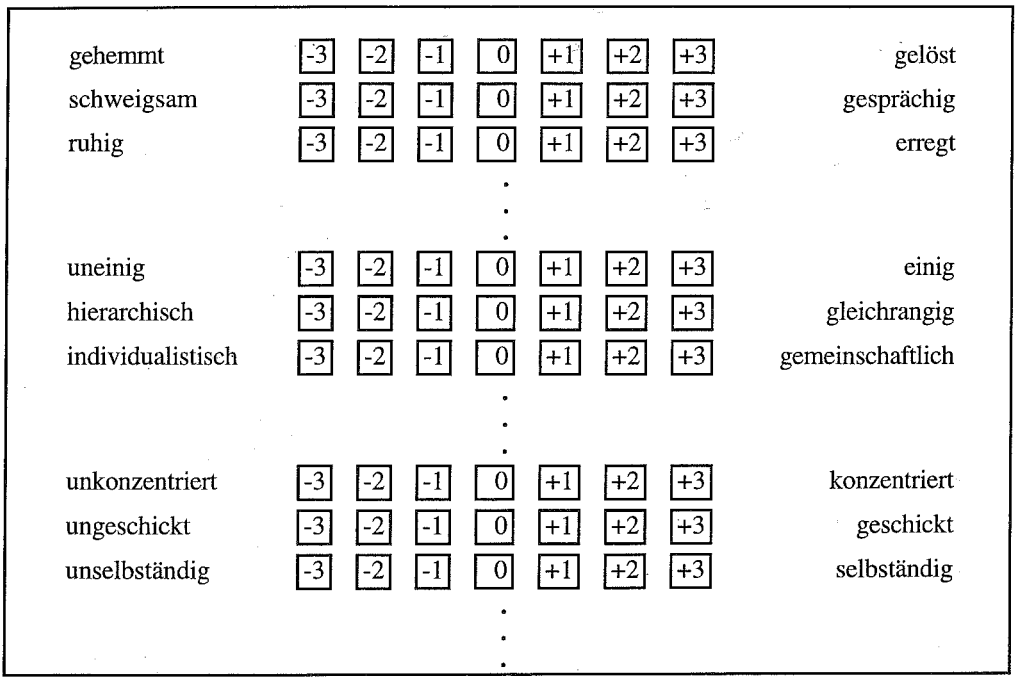
Die vorangegangene Darstellung läßt die spätere Einbettung und Revision des Balesschen Beobachtungsschemas in ein mehrdimensionales System zur Analyse von Gruppen, genannt

SYMLOG, („*Sy*stem for the *M*ultiple *L*evel *O*bservation of *G*roups"; vgl. Bales & Cohen, 1982) unberücksichtigt. Mit diesem Ansatz, der nicht mehr einfach als Modifikation der IPA angesehen werden kann, verbindet sich ein umfangreiches Methodeninventar zur Erforschung der verschiedenen Verhaltensebenen in Kleingruppen. Dabei werden verbale und nonverbale Verhaltensweisen, Einstellungen und Eindrücke der Gruppenmitglieder über die Gruppensituation als auch deren Persönlichkeitseigenschaften mit Hilfe von Interaktionssignierung via Beobachtung als auch anhand von Ratings erfaßt. Im Gegensatz zur IPA wird nicht nur die Art der Interaktion, sondern auch der Inhalt kodiert. Grundlage des SYMLOG-Systems ist die "systematische mehrstufige Feldtheorie" (Bales, 1982), eine integrative Rahmentheorie zur Gruppendynamik, die Elemente aus verschiedenen Bereichen der Persönlichkeits- und Sozialpsychologie enthält. Eine umfassende Darstellung verschiedener Anwendungsbereiche von SYMLOG findet sich in Polley, Hare und Stone (1988). Für den deutschen Sprachraum sei auf die Arbeiten der Saarbrücker Forschungsgruppe zum SYMLOG-Ansatz verwiesen (vgl. die Literaturhinweise bei Kohler, 1986).

Das IPA-System von Bales erfordert recht komplexe Wertungs- und Interpretationsleistungen durch die Beobachter. Wir haben uns damit schon recht weit sowohl von verbalen Beschreibungen als auch von Beobachtungssystemen entfernt, in denen klare, offensichtliche Verhaltenskategorien gegeben waren, wie bei Barash (1972; vgl. Abschnitt 4.5) oder Bandura (1966; vgl. Abschnitt 1.1). Wir gehen aber jetzt mit der Einführung der Rating- oder Einschätzskalen noch einen Schritt weiter. Beginnen wollen wir diese Einführung mit einem Beispiel, welches ein wenig an die Untersuchung von Bandura erinnert.

5.4 Einführung der Ratingskalen

Charlton, Liebelt, Sültz und Tausch (1974) interessierten sich in einer Erkundungsstudie für die unmittelbaren Auswirkungen unterschiedlicher Verhaltensmodelle auf das Gruppenverhalten von Kindern. Die Autoren zeigten u.a. je 20 Kleingruppen einen Film mit aggressivem (aber bestraftem) bzw. kooperierendem Modellverhalten; die Kleingruppen umfaßten jeweils drei Kinder. Danach wurde den Gruppen eine kleine Aufgabe („Münzen sortieren") gegeben, die Kooperation erforderte. Um möglichst viele Aspekte des Gruppenverhaltens zu berücksichtigen, entschieden sich Charlton et al. nicht dafür, ein Zeichen- oder Kategoriensystem einfacher Verhaltensweisen zu erstellen, sondern vertrauten der Fähigkeit der Beobachter, die „Gruppenatmosphäre" während der zehnminütigen Aufgabenphase global einzuschätzen. Sie erstellten eine Liste von Adjektivgegensatzpaaren, wie sie in Abbildung 33 wiedergegeben ist. Durch jedes Gegensatzpaar wurde eine Dimension aufgespannt, so daß etwa das Gruppenverhalten von „sehr ruhig" über mehrere Abstufungen bis zu „sehr erregt" eingestuft werden konnte.



Abbilkdung 33: Adjektivpaare zur Beurteilung der Kindergruppen in der Untersuchung von Charlton et al. (1974, S. 167f.)

Bevor wir diese Art der Messung - die Einschätz- oder Rating-Skala - eingehender diskutieren, seien noch kurz die Ergebnisse der Studie von Charlton et al. referiert. Es zeigte sich an den Ergebnissen, daß durch die Adjektive im wesentlichen drei unabhängige Aspekte des Gruppenverhaltens erfaßt wurden; dabei unterschieden sich die Gruppen, die die unterschiedlichen Filme gesehen hatten, nicht in dem Aspekt „kooperierendes vs. rivalisierendes Gruppenverhalten". Auch in dem Aspekt „systematische vs. unsystematische Gruppenarbeit" zeigte sich kein Unterschied. Lediglich in einem dritten Aspekt wirkten sich die Filmszenen verschieden aus: Die Gruppen, die aggressive Filmszenen beobachtet hatten, wirkten eher ernst-gehemmt, die Gruppen, die kooperiende Verhaltensmodelle gesehen hatten, wirkten heiterer-gelöster.

Ähnlich wie bei Charlton et al. können wir fast beliebig Ratingskalen entwickeln. Dabei spielt das Merkmal der Bipolarität nur eine untergeordnete Rolle. Wir könnten etwa daran denken, statt des umfangreichen Kategorienkatalogs von Glennon und Weisz (1978; vgl. Abschnitt 4.4) zur Erfassung der Angst von Vorschulkindern eine Ratingskala zu konstruieren, die einem Beobachter direkt ein summarisches Gesamturteil abverlangt:

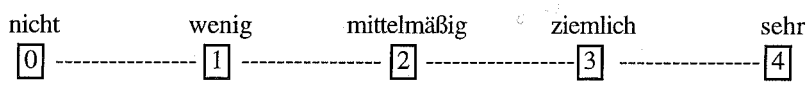


Abbildung 34: Ratingskala zur Erfassung von Angst

Was ist nun das Besondere an derartigen Einschätzskalen? Was trennt sie von Zeichen- und Kategoriensystemen? Rein formal betrachtet, verlassen wir den Bereich der Nominalsysteme und messen nun auf dem Niveau von Ordinal- oder – zumindest häufig intendiert – Intervallniveau (vgl. dazu unsere Erörterungen in Kap. 5.1.5). Vom Klassifikationsaspekt bei den Zeichen- und Kategoriensystemen gehen wir über zur Erfassung von Häufigkeits- und Intensitätsmerkmalen der beobachteten Einheiten. Mit dieser formalen Abgrenzung sind einige andere Merkmale verbunden, zwar nicht zwangsläufig, aber in der konkreten Forschungspraxis.

(1) Die zeitliche Dauer der Einheiten

Das Beispiel von Charlton et al. (1974; s.o.) ist sicherlich typisch für die Verwendung von Ratingskalen: Einheit ist hier das Verhalten der gesamten Gruppe in Eineinhalb-Minuten-Intervallen. Verglichen mit den eher kurzen Intervallen bei Zeichen- oder Kategoriensystemen mit Zeittaktung, die sich im übrigen in der Regel auf individuelles Verhalten beziehen (s.o.), wird hier eine äußerst ausgedehnte Einheit zur Beurteilung herangezogen. Der Beobachter fällt sein Urteil durch Integration der Beobachtungen von einzelnen Verhaltensweisen über die Zeit. In dieser Hinsicht liegt der Hauptvorteil des Ratingverfahrens in seiner Ökonomie: Man beachte, daß es für Charlton et al. (1974) möglich gewesen wäre, einzelne Verhaltensindikatoren für die Ratingdimensionen zu definieren, um diese im Zeittaktverfahren mit kurzen Intervallen festzustellen. Im zweiten Schritt hätten die Autoren von den Häufigkeiten einzelner Kategorien auf das Gruppenklima geschlossen; dieses Vorgehen wäre natürlich ungleich aufwendiger gewesen. Offenbar machten Charlton et al. die Annahme, daß die durch ein derartiges Indikatorvorgehen erhaltene Information kaum aussagekräftiger gewesen wäre als die durch das Ratingverfahren hergestellte. Dieser Ökonomieaspekt wird vor allem bei Ratingsystemen deutlich, die nicht einen Intensitätsaspekt, sondern die Häufigkeit bestimmter Verhaltensweisen einschätzen lassen (vgl. etwa die Diskussion bei Beck, 1987, S. 45f.). Kritisch anzumerken ist natürlich, daß die Beobachtung viel fehleranfälliger wird: Vor allem die oben (Abschnitt 3.2.2) eingeführte Problematik der „Fehler zu Lasten der Beobachter" stellt sich hier verschärft. Zu betonen ist noch einmal, daß die Integration über einen längeren Zeitraum nicht notwendigerweise ein Merkmal von Ratingverfahren ist. Faßnacht (1995) berichtet etwa über die Skala von Lee (1932) zur Erfassung „emotionaler Instabilität von Kindergartenkindern". Hier wird im 15-Sekunden-Takt eingeschätzt.

(2) Die Intransparenz des Urteils

Ein zweites Merkmal von Ratingskalen ist die Intransparenz des Urteilsvorganges: Es ist in den meisten Fällen nicht klar nachzuvollziehen, welche Integrationsfunktion durch die Rater realisiert wird. Nehmen wir etwa das Adjektivpaar „gehemmt - gelöst" aus der Studie von Charlton et al.: Während man sich sehr wohl vorstellen kann, daß Beurteiler hierbei zu übereinstimmenden Einschätzungen gelangen, ist jedoch nicht recht durchschaubar, welche Verhaltensindikatoren mit welcher Gewichtung in das Urteil einfließen. Sind es eher verbale Merkmale (schweigsames Erledigen der Aufgabe versus lautes Reden, Lachen der Kinder) oder Bewegungsmuster („steifes" Sitzen, wenige Bewegungen fast nur der Arme versus Bewegungen des ganzen Körpers), die das Urteil bestimmen? Wie gehen einzelne Verhaltensweisen in der Integration über die Zeit ein? Das „Heruntertransformieren" einer komplexen Einschätzung auf einfache Verhaltensweisen (Indikatoren) wird nach Langer und Schulz von Thun (1974) entweder daran scheitern, daß die Indikatoren dem komplexen Begriff nicht mehr gerecht werden oder ihrerseits noch eine Einschätzung voraussetzen (wie laut ist „lautes Reden"?). Darüber hinaus wird die Anzahl der Indikatoren nach Ansicht der Autoren zu groß, als daß die so erhaltenen Beobachtungssysteme noch handhabbar wären.

Bei Ratingskalen wird also in der Regel in besonderer, in dieser Form nicht ersetzbarer Weise der Meßapparat „Mensch" mit seinen komplexen Fähigkeiten genutzt. So schreiben auch Langer und Schulz von Thun (1974): „Ratingverfahren nutzen die Fähigkeit des menschlichen Gehirns zur *'Indikatorenverschmelzung'*, zur automatischen Integration einer Vielzahl von Einzelindikatoren, die sich in ihrer Wirkung in einer fast unentwirrbaren Weise verstärken, aufheben oder sonstwie in Wechselwirkung treten" (S. 20; Hervorhebung im Original).

Während Langer und Schulz von Thun (1974) somit aus der Intransparenz des Urteilsvorganges die Unverzichtbarkeit der Ratingskalen zur Erfassung komplexer Sachverhalte ableiten, sieht Beck (1987) hierin gerade den Nachteil; er schreibt: „Mit der Überantwortung der 'Indikatorenverschmelzung' an einen internen Verarbeitungsprozeß, der offenbar für unkontrollierbar gehalten wird, öffnen sich nämlich die Schleusen für eine Vielzahl von Unklarheiten, Irrtümern und Fehlern, die sich von der empirischen Basis bis in die Theorie hinein fortpflanzen" (S. 40). Wie im letzten Abschnitt schon betont, müssen die Fehlerquellen, die wir im Kapitel 3 vorgestellt haben, bei der Anwendung von Ratingsystemen besonders berücksichtigt werden.

Aus diesen Gründen sind gerade bei Ratingskalen Nachweise der Validität und Reliabilität in besonders starkem Maße gefordert. Intensives Beobachtertraining an Beispielepisoden ist ebenso nötig wie der Nachweis der Beobachterübereinstimmung. Dabei sollte gerade hier daran gedacht werden, daß nicht allein die Übereinstimmung zwischen gemeinsam geschulten Beobachtern als Tauglichkeitsargument gelten kann, da sich gerade bei diesen eine wiederum intransparente Konsensfindung während des Trainings einstellen mag. Die Übereinstimmung zwischen Beobachtern, die verschiedenen Trainingsgruppen (die sich natürlich auf das gleiche dokumentierte Trainingsmaterial stützen) entstammen, dürfte wesentlich aussagekräftiger sein. Nicht zuletzt wird sich die Tauglichkeit einer Ratingskala in ihrer Validität zeigen. Dabei spielen sowohl

Aspekte der Face-Validität als auch der Konstruktvalidität eine bedeutende Rolle: kann man anhand der Formulierung der Skala den Bezug zu theoretischen Begriffen erkennen? Verhält sich das über die Ratingskala gemessene Konstrukt im „Konzert" anderer Konstrukte so, wie es theoretisch postuliert wurde?

Es ist in diesem Zusammenhang wichtig, darauf hinzuweisen, daß in der Trainingssituation versucht werden sollte, die Intransparenz teilweise aufzuheben. Langer und Schulz von Thun (1974) haben Vorschläge erarbeitet, wie durch eingehende Merkmalsbeschreibung des einzuschätzenden Begriffes, durch Umschreibungen der einzelnen Skalenstufen und durch Ankerbeispiele (also in welchen Fällen unzweifelhaft ein bestimmter Wert vergeben werden muß) eine Homogenisierung des Urteilsvorganges erreicht werden kann.

Auch hier sollte man allerdings betonen, daß sich aufgrund dieses Kriteriums der „Intransparenz" Ratingskalen nicht von anderen Systemen *definitorisch* abgrenzen lassen: Einerseits lassen sich sowohl sehr wahrnehmungsnahe Ratingskalen konstruieren (Lautstärke von Äußerungen), andererseits ist die Intransparenz bei Systemen wie etwa dem Balesschen IPA (Abschnitt 5.3) ebenfalls nicht unbedingt gering.

(3) Erfassung von Dispositionskonstrukten

Ein weiterer Punkt, der bei der Diskussion von Ratingskalen eine Rolle spielt, ist die Frage danach, was typischerweise mit Ratingskalen erfaßt wird. So sehen Cairns und Green (1979) darin eine Hauptunterscheidung zwischen gängigen Zeichen- oder Kategoriensystemen und Ratingverfahren: Mit letzteren sollen in der Regel Dispositionskonstrukte direkt erfaßt werden, während erstere Verhaltensweisen, also potentielle Indikatoren von Dispositionen, messen.

Erinnert man sich an die Einführung der Problematik der Dispositionsbegriffe im Kapitel 4.4 und greift vor diesem Hintergrund das Thema der Intransparenz des Urteilsprozesses noch einmal auf, so löst sich ein in dem Zusammenhang implizit gebliebener Widerspruch zumindest partiell auf: Wir hatten bei der Diskussion der Intransparenz durchaus zugestanden, daß Rater zu übereinstimmenden Urteilen kommen mögen, ohne daß sichergestellt ist, daß die Integrations- und Gewichtungsfunktionen übereinstimmen. Die Übereinstimmung läßt sich darauf zurückführen, daß die zu ratenden Merkmale auf alltagssprachlichen Begriffen beruhen, deren Bedeutung offenbar intersubjektiv von den Mitgliedern einer Sprachgemeinschaft geteilt werden. Implizit sind uns allen die Reduktionssatzsysteme (und damit die Zuordnungsregeln) für Dispositionsbegriffe wie „ängstlich", „selbständig" und „friedfertig" bekannt. Die Begriffe sind in ein „Sprachnetz" eingehängt, welches auch sehr beobachtungsnahe Begriffe enthält.

Wiederum muß man festhalten, daß Ratingsysteme nicht dadurch definiert werden, daß sie eher zur Erfassung von Dispositionskonstrukten konstruiert werden. Gerade dann, wenn kurzzeitige Verhaltensweisen eingeschätzt werden sollen, ist dies nämlich nicht der Fall. Wenn wir aber oben schon festgehalten haben, daß Ratings sich typischerweise als globales Urteil auf größere Einheiten beziehen, wird der Charakter der Dispositionsmessung deutlicher.

Wir haben nun mit den Kategorien- und Zeichensystemen die Palette der wichtigsten Beobachtungsverfahren vervollständigt. Abschließend müssen aber für sie die Berechnungsmethoden der Beobachterübereinstimmung, deren Grundgedanke wir im Abschnitt 4.5 kennengelernt haben, etwas erweitert werden.

5.5 Beobachterübereinstimmung: Erweiterungen

Im Abschnitt 4.5 hatten wir das Problem der Beobachterübereinstimmung und die prinzipielle Lösung erläutert. Mit den dort entwickelten Koeffizienten π („Pi") und κ („Kappa") wurde auch die praktische Herangehensweise bei dem einfachen Fall einer Vier-Felder-Kontingenztafel verdeutlicht. Mittlerweile dürfte deutlich geworden sein, daß wir in der Praxis der Beobachtung eine Vielfalt von Beobachtungssituationen und Skalen antreffen. An dieser Stelle sollen für die häufigsten Fälle Beobachterübereinstimmungsmaße angegeben werden. Wir werden dabei sehen, daß sich sehr viele der Fälle als Verallgemeinerungen des einfachen Falles deuten lassen.

(1) Kategoriensystem mit mehr als zwei Kategorien

Dies ist sicherlich die einfachste Erweiterung: Tatsächlich sind π und κ für diesen allgemeineren Fall definiert. Zur Erinnerung: Die Grundformel sowohl für π als auch κ setzte die erwartete zur tatsächlich beobachteten Übereinstimmung in Beziehung:

$$\pi / \kappa = \frac{P_o - P_e}{1 - P_e}$$

Wie wir gesehen haben, unterscheiden sich nur die Wege, auf denen die Erwartungswerte berechnet werden. In der zugrundeliegenden Formel ist aber von der Anzahl der Zellen noch keine Rede. Die erwartete bzw. tatsächliche Übereinstimmung ergab sich durch die Summierung der relativen Häufigkeiten (erwartet bzw. tatsächlich) der Übereinstimmungszellen f_{11} und f_{22} (vgl. Kap. 4.5).

Die Erweiterung auf umfangreichere Systeme liegt nun auf der Hand: Im Fall weiterer Kategorien verfahren wir einfach ebenso; lediglich die Kontingenztafel wird größer (Abb. 35).

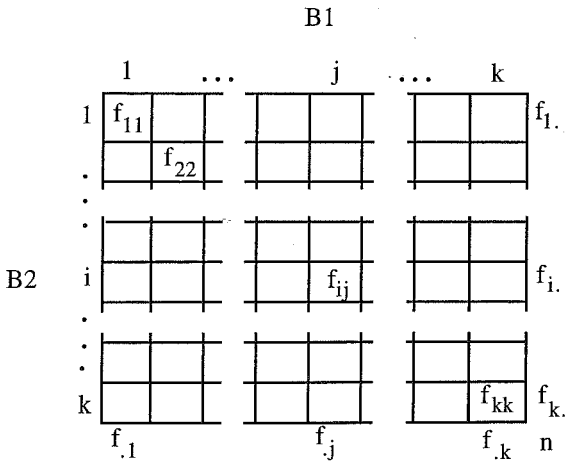


Abbildung 35: Kontingenztafel mit mehr als zwei Kategorien[33]

Die P_o- und P_e-Berechnungen ergeben sich ganz analog zu der einfachen Kreuztabelle. Um den Rechenaufwand zu vermindern (wie wir gleich sehen werden), sollen die Platzhalter („f") ab jetzt wieder für die *absoluten Häufigkeiten* stehen.

$$P_o = \frac{1}{n} * \sum_{i=1}^{k} f_{ii}$$

κ: $$P_e = \frac{1}{n^2} * \sum_{i=1}^{k} f_{i.} * f_{.i}$$

π: $$P_e = \frac{1}{n^2} * \sum_{i=1}^{k} \left(\frac{f_{i.} + f_{.i}}{2}\right)^2$$

(Zur Erläuterung: Der Ausdruck $\sum_{i=1}^{k} f_{ii}$ ist so zu lesen: $f_{11} + f_{22} + f_{33} + ... + f_{kk}$.)

Die Formeln stehen im Grunde nur für eine Aufsummierung der erwarteten bzw. tatsächlichen *relativen* Häufigkeiten. Da die relativen Häufigkeiten sich aber immer aus der Division der absoluten Häufigkeit durch die Gesamthäufigkeit n ergeben, kann man sich einige Rechnerei ersparen: Deshalb ist die immer wiederkehrende Division durch n in den Formeln vor das Summenzeichen gezogen worden. Die Formel für die Standardabweichung (vgl. Abschnitt 4.5) ist ebenfalls einfach zu verallgemeinern (vgl. Fleiss, 1973):

---

33 Um Mißverständnissen vorzubeugen: Die Zeilenbezeichnung i bzw. die Spaltenbezeichnung j bezeichnen nicht eine bestimmte Zeile bzw. Spalte, sondern sind *variable Bezeichnungen*, die jeweils für *irgendeine* Zeile bzw. Spalte stehen können, also z.B. auch für Zeile 1, Zeile k, Spalte 1 oder Spalte k.

$$SD_\kappa = \sqrt{\frac{1}{n(1-p_e)^2} * (P_e + P_e^2 - \frac{1}{n^3} \sum_{i=1}^{k} f_{i.} * f_{.i} * (f_{i.} + f_{.i}))}$$

(2) Zeichensystem mit mehr als zwei Kategorien

Liegt der Beobachtung ein Zeichensystem zugrunde, unterscheiden wir zwei Fälle:

1. Sprechen wir nur deshalb von einem Zeichensystem (im Gegensatz zu einem Kategoriensystem), weil wir nicht jede Beobachtungseinheit in eine Kategorie einordnen, im übrigen aber, wenn wir zuordnen, dies eindeutig tun, so ergibt sich die Übereinstimmung ganz einfach: Wir erweitern das System um eine Restkategorie und verfahren dann wie mit einem Kategoriensystem.
2. Liegt jedoch der Fall vor, daß wir für Beobachtungseinheiten mehrere Kategorien vergeben dürfen, müssen wir für jede der Kategorien eine Vier-Felder-Kontingenztafel mit den Randbezeichnungen „Kategorie gewählt" und „Kategorie nicht gewählt" zur Grundlage der Berechnung nehmen. Wir erhalten damit also für jede Kategorie des Zeichensystems ein Übereinstimmungsmaß. Um in diesem Fall die Güte des gesamten Systems prägnant zu verdeutlichen, kann man etwa den Mittelwert über alle Koeffizienten bilden.

(3) Beobachterübereinstimmung für mehr als zwei Beobachter

Bisher wurden die Übereinstimmungsmaße lediglich für den Fall zweier Beobachter eingeführt. Werden dagegen mehrere Beobachter eingesetzt, bietet sich als einfachste Lösung an, für alle möglichen Beobachterpaare die bisher vorgeschlagenen Koeffizienten zu berechnen und dann den Mittelwert zu bilden.

In der Literatur finden sich mehrere Vorschläge zu diesem Thema. Light (1971) und Fleiss (1971) haben unabhängig voneinander den Versuch unternommen, κ zu verallgemeinern. Eine Untersuchung des Ansatzes von Light zeigt jedoch sehr schnell, daß seine Formel äquivalent zu unserem Vorschlag der Mittelung aller möglichen κ-Werte ist (vgl. dazu auch Conger, 1980). Fleiss' Ausgangspunkt für eine Verallgemeinerung ist die Klassifizierung jeder Einheit durch eine zufällige Auswahl von Beobachtern (deren Anzahl jedoch stets gleich ist). Da in diesem Fall die Randwahrscheinlichkeiten eines einzelnen Beobachters nicht sinnvoll erhoben werden können, basiert seine Formel für die erwartete Übereinstimmung auf den relativen Wahlhäufigkeiten der Kategorien über alle Entscheidungen aller Beobachter hinweg. Insofern handelt es sich bei Fleiss' Index um einen generalisierten π–Koeffizient. Conger (1980) und Uebersax (1982) haben Fleiss' Vorschlag dahingehend modifiziert, daß die unterschiedlichen Randwahrscheinlichkeiten Berücksichtigung finden (vgl. zu diesem Problem außerdem Berry & Mielke, 1988).

Ein Problem dabei bleibt allerdings, wie "Übereinstimmung" im Fall mehrerer Beobachter definiert werden soll: Soll als Fall von "Übereinstimmung" gewertet werden, wenn zwei Beobachter dieselbe Kategorie wählen? Oder wollen wir diesen Begriff erst dann benutzen, wenn alle

Beobachter übereinstimmend urteilen? Im Fall der Mittelung aller möglichen Werte haben wir uns implizit für die erste Möglichkeit entschieden (zu den anderen Möglichkeiten vgl. Conger, 1980, und Hubert, 1977).

(4) Beobachterübereinstimmung bei ungleicher Gewichtung der Nicht-Übereinstimmungen

Nicht in allen Fällen wird man die Fälle von Nicht-Übereinstimmung gleich schwer gewichten. So läßt sich etwa der Fall denken, daß das Beobachtungssystem wenige Oberkategorien umfaßt, innerhalb derer feiner ausdifferenzierte Unterkategorien definiert werden. Ordnen die Beobachter eine Einheit verschiedenen Oberkategorien zu, würde man dies möglicherweise als „schwereren" Fall von Nicht-Übereinstimmung werten, als wenn lediglich unterschiedliche „Fein"-Kategorien innerhalb derselben Oberkategorie gewählt werden.

Um mit solchen Fällen adäquat umzugehen, gehen wir in zwei Schritten vor:

1. Wir definieren die bekannten Maße π und κ so um, daß in ihre Berechnung nicht die relativen Häufigkeiten der Übereinstimmungs-, sondern jene der Nicht-Übereinstimmungszellen eingehen. Das ist deshalb relativ einfach, weil sich die entsprechenden relativen Häufigkeiten zu eins addieren. Wenn wir mit P_{ne} und P_{no} die erwarteten bzw. tatsächlichen relativen Häufigkeiten der Nicht-Übereinstimmung bezeichnen, ergibt sich:

$$P_{no} = 1 - P_o = \frac{1}{n}\sum_{i \neq j} f_{ij}$$

$$P_{ne} = 1 - P_e$$

$$= \frac{1}{n^2} * \sum_{i \neq j} \left(\frac{f_{i.} + f_{.i}}{2}\right)\left(\frac{f_{j.} + f_{.j}}{2}\right) \qquad (\pi)$$

$$= \frac{1}{n^2} * \sum_{i \neq j} f_{i.} * f_{.j} \qquad (\kappa)$$

(Die Summierungen gehen sowohl für i als auch j von 1 bis N mit der Randbedingung i ungleich j; wir summieren also über alle Zellen außer den Übereinstimmungszellen.)

Damit ergibt sich

$$\pi / \kappa = \frac{P_o - P_e}{1 - P_e}$$

$$= \frac{(1 - P_{no}) - (1 - P_{ne})}{1 - (1 - P_{ne})} = \frac{1 - P_{no} - 1 + P_{ne}}{1 - 1 + P_{ne}} = \frac{P_{ne} - P_{no}}{P_{ne}}$$

$$= 1 - \frac{P_{no}}{P_{ne}}$$

2. Im zweiten Schritt führen wir für alle Zellen einen Gewichtsfaktor ein, der unsere Gewichtung der Nicht-Übereinstimmungen abbildet. Bleiben wir bei dem Beispiel der Ober- und Unterkategorien und stellen uns ein System aus vier Kategorien vor: zwei Oberkategorien, die wiederum jeweils ein Paar von feineren Kategorien umfassen. Wir vereinbaren, daß eine Nicht-Übereinstimmung, die die Grenzen der groben Kategorien „überschreitet", doppelt so stark gewichtet wird wie eine Nichtübereinstimmung innerhalb der Oberkategorien. Die nächste Kontingenztafel bildet die entsprechende Gewichtsmatrix ab (Abb. 36).

			B1		
			K		J
			K_1	K_2	J_1	J_2
B2	K	K_1	0	1	2	2
		K_2	1	0	2	2
	J	J_1	2	2	0	1
		J_2	2	2	1	0

Abbildung 36: Gewichtsmatrix

In die Berechnung der Koeffizienten gehen nun also nicht mehr die Häufigkeiten der Zellen, sondern die Produkte aus Gewicht (w) und Häufigkeit (f) ein.

$$P_{no} = \frac{1}{n} * \sum w_{ij} * f_{ij}$$

$$P_{ne} = \frac{1}{n^2} * \sum w_{ij} * \left(\frac{f_{i.} + f_{.i}}{2}\right)\left(\frac{f_{j.} + f_{.j}}{2}\right) \qquad (\pi_w)$$

$$P_{ne} = \frac{1}{n^2} * \sum w_{ij} * f_{i.} * f_{.j} \qquad (\kappa_w)$$

(Zur Erläuterung: Die Summierung läuft wieder über i und j; wir können die Randbedingung $i \neq j$ weglassen, da hier $w=0$ ist.)

Die Formel für die Standardabweichung soll hier nicht wiedergegeben werden; interessierte Leser seien auf Fleiss, Cohen und Everitt (1969) verwiesen.

(5) Beobachterübereinstimmung bei Ratingskalen

Wenn wir jede Beobachtungseinheit auf einer Ratingskala einstufen, um zu einem Meßergebnis zu gelangen, können wir prinzipiell ebenso eine Kontingenztafel erstellen wie in den anderen Fällen. Wir bekommen allerdings folgendes Problem: Bei einem Kategoriensystem wiegen die Fälle von Nicht-Übereinstimmung in der Regel gleich schwer. Wichtig ist in solchen Fällen lediglich, ob von den beiden Beobachtern dieselbe Kategorie gewählt wurde oder nicht; welcher Art die Nicht-Übereinstimmung war, ist nicht von Belang. Bei einer Ratingskala führt dieses Vorgehen offensichtlich zu unsinnigen Übereinstimmungsindizes. Nehmen wir etwa folgendes Beispiel: Wir beobachten Gesprächstherapiesequenzen, in denen wir - vielleicht neben anderen, stärker an Verhaltensindikatoren orientierten Zeichen - zu jeder Zeiteinheit ein Urteil über die „emotionale Beteiligung" des Klienten auf einer fünf-stufigen Ratingskala abgeben. Es versteht sich von selbst, daß ein so komplexes Urteil einer Reliabilitätsanalyse unterzogen werden muß. Betrachten wir als Beispiel wiederum ein fiktives Ergebnis zweier Beobachter (Abb. 37).

Zunächst fällt an der Tabelle auf, daß bei weitem nicht alle Einstufungen in den Übereinstimmungszellen wiederzufinden sind; allerdings gruppieren sich die Beobachtungen doch mehr oder weniger gut um die Haupt-Diagonale herum. Es gibt eine Häufung der Nicht-Übereinstimmungen links von der Diagonalen: Dahinter verbirgt sich lediglich die Tatsache, daß Beobachter 2 tendenziell höhere Werte gibt. Würden wir nun etwa den κ-Koeffizienten berechnen, würden alle Fälle, die nicht in der Diagonalen liegen, gleichermaßen als Nicht-Übereinstimmung gewertet. Vor allem an dem einen „Ausreißer" in der rechten oberen Ecke wird deutlich, daß dies kein sinnvolles Vorgehen sein kann: Dieser doch deutlich diskrepantes Rating signalisierende Fall wird ebenso behandelt wie die Einstufungen, in denen Beobachter 1 die '2', Beobachter 2 die '3' wählt. Entsprechend fällt der κ-Koeffizient mit 0.58 zwar nicht extrem niedrig, aber auch noch nicht gut aus.

		B1					
		1	2	3	4	5	
B2	1	3			1		4
	2	2	7				9
	3	1	2	12	1		16
	4		1	3	9	1	14
	5			2	2	3	7
		6	10	17	13	4	50

Abbildung 37: Fiktives Ergebnis einer Rating-Beobachtung

Aber was liegt näher, als die im letzten Abschnitt eingeführte Gewichtung zu nutzen? Sinnvoll ist es in diesem Zusammenhang, Gewichte zu wählen, die in einem geeigneten Verhältnis zur Differenz der beiden Beobachter-Skalenwerte stehen. Als zunächst einfachste Lösung bietet sich an, die absolute Differenz zu nehmen; d.h. wir würden folgende Gewichtstabelle erhalten (Abb. 38).

B2 \ B1	1	2	3	4	5
1	0	1	2	3	4
2	1	0	1	2	3
3	2	1	0	1	2
4	3	2	1	0	1
5	4	3	2	1	0

Abbildung 38: Gewichtsmatrix für das Rating-Beispiel (einfache Differenzgewichtung)

Tatsächlich ist es jedoch aus verschiedenen Gründen üblich, das *Quadrat der Differenz* zu wählen. Dieses Vorgehen wird plausibel, wenn man sich deutlich macht, daß bei quadratischer Gewichtung die „harmlosen" Fälle von Nicht-Übereinstimmung (eine Skaleneinheit Diskrepanz) relativ schwach eingehen, während die Differenzen, die entscheidend gegen die Güte einer Skala sprechen (der Fall in der oberen rechten Ecke!) überproportional eingehen. Die Gewichtstabelle, die wir dann erhalten, zeigt Abbildung 39.

B2 \ B1	1	2	3	4	5
1	0	1	4	9	16
2	1	0	1	4	9
3	4	1	0	1	4
4	9	4	1	0	1
5	16	9	4	1	0

Abbildung 39: Gewichtsmatrix für das Rating-Beispiel (quadratische Differenzgewichtung)

Die Berechnung dieses gewichteten κ (κ_w) ist nun ganz einfach:

$$\kappa_w = 1 - \frac{P_{no}}{P_{ne}}$$

$$= 1 - \frac{\frac{0*3+9*1+1*2+0*7+4*1+...+0*3}{50}}{\frac{0*6*4+1*10*4+4*17*4+...+0*4*7}{50*50}}$$

$$= 1 - \frac{.72}{2.61} = .72$$

(Die Zahlen im Doppelbruch ergeben sich, wenn man die Kreuztabelle in Abbildung 37 [für die Häufigkeiten] bzw. Abbildung 39 [für die Gewichte] von oben links nach unten rechts durchgeht.) Wir erhalten also einen Wert, der viel zufriedenstellender ist als das ungewichtete κ; er entspricht mehr dem Eindruck, daß die Nicht-Übereinstimmungen eher geordnet an der Diagonalen liegen, als daß sie zufällig über die Zellen verteilt wären. Wenn in den weiteren Ausführungen von κ_w die Rede ist, so ist immer das κ mit quadratischer Gewichtung gemeint.

Natürlich stellt sich auch bei dieser Anwendung die Frage, wie die Randverteilungen einbezogen werden sollen. Im besonderen geht es dabei um die Frage unterschiedlicher „Ankerpunkte" der Rater. Betrachten wir dazu ein anderes Ergebnisbeispiel einer Beobachtung (Abb. 40). Offensichtlich handelt es sich hierbei um einen besonders seltsamen Fall von Beobachtung. In gewisser Hinsicht urteilen die beiden Beobachter übereinstimmend: Könnten wir die Randverteilungen so verschieben, daß sie in beiden Fällen um den Wert 3 streuen, so hätten wir perfekte Übereinstimmung; andererseits liegt offenbar keine Beobachtungseinheit in den Diagonalzellen: Insofern haben wir einen Fall überzufälliger Nicht-Übereinstimmung.

		B1				
		1	2	3	4	5
B2	1			10		
	2				20	
	3					10
	4					
	5					

Abbildung 40: Fiktive Beobachtungsdaten

Da man derartig extreme „Anker"-Unterschiede wohl in der Regel als „Defekt" des Beobachtungssystems entlarven möchte, sollte ein geeignetes Maß sicherlich von 1 verschieden sein (vgl. dazu aber den Anhang). Tatsächlich ergibt sich ein κ_w von 0.2, obwohl - wie oben ausgeführt - bei κ die Unterschiedlichkeit der Randverteilung zugelassen ist. Einen noch deutlicheren Hinweis auf den Defekt dieses Systems erhalten wir dementsprechend, wenn wir analog zum κ_w das π_w berechnen: hierbei ergibt sich ein Wert von -0.333.

Mit κ_w und π_w haben wir damit einfache Erweiterungen der Nominalskala-Koeffizienten kennengelernt, die plausibel und rechnerisch sehr einfach sind. In der Literatur zur Beobachterübereinstimmung werden üblicherweise sogenannte „Intraklassen-Korrelationskoeffizienten" als Übereinstimmungsmaß bei Ratingskalen eingeführt. Auf sie werden wir in einem Anhang eingehen. Diese Einführung wird allerdings etwas voraussetzungsreicher sein als die bisherigen Erörterungen zur Beobachterübereinstimmung. Glücklicherweise gilt jedoch, daß κ_w und π_w bis auf einen vernachlässigbaren Term identisch mit zwei Varianten der Intraklassenkorrelation sind, so daß man genausogut diese recht einfach zu berechnenden Gütemaße benutzen kann.

(6) Beobachterübereinstimmung bei unklarer Segmentierung

Mehr oder weniger unausgesprochen sind wir bisher davon ausgegangen, daß die Einteilung des Geschehens *kein* kritischer Punkt der Beobachtung ist: Die Übereinstimmung in der Einheitenbildung wurde als unproblematisch vorausgesetzt, Grundlage der Reliabilitätsanalyse war immer nur die Zuordnung der jeweiligen Einheit zu einer Kategorie. Aus den vorherigen Kapiteln (s. vor allem Abschnitt 4.3) sollte jedoch deutlich geworden sein, daß auch die Segmentierung des Geschehens eine Aufgabe sein kann, die dem Beobachter überlassen wird und insofern ebenso mit in die Überprüfung eingehen muß. Stellen wir uns etwa ein Kategoriensystem vor, daß an „natürlichen" Handlungseinheiten orientiert ist. Wir erhalten etwa folgendes Ergebnis zweier Beobachter (Abb. 41).

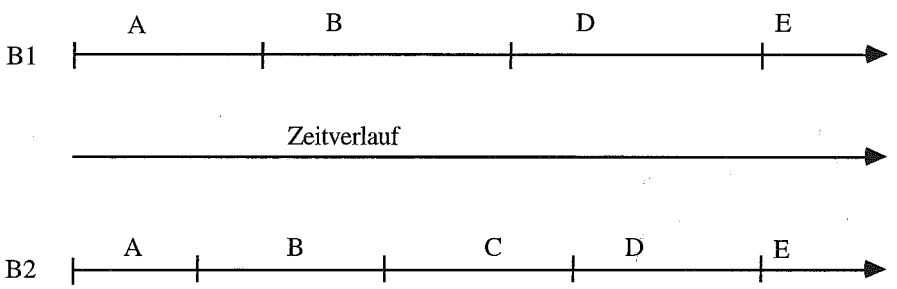


Abbildung 41: Segmentierung des Geschehens durch zwei Beobachter (mod. nach Asendorpf & Wallbott, 1979)

Beobachter 1 teilt das Geschehen in vier Handlungseinheiten im Sinne des Kategoriensystems ein; Beobachter 2 wählt dagegen eine Segmentierung in fünf Abschnitte. Wie können wir in ei-

nem solchen Fall ein Maß für die Beobachterübereinstimmung angeben? Asendorpf und Wallbott (1979; vgl. aber auch Bakeman & Gottman, 1986) machen den Vorschlag, nachträglich ein Zeitraster über den Geschehensablauf zu legen, um so die Erstellung von Kontingenztafeln zu ermöglichen; Abbildung 42 verdeutlicht diese Idee.

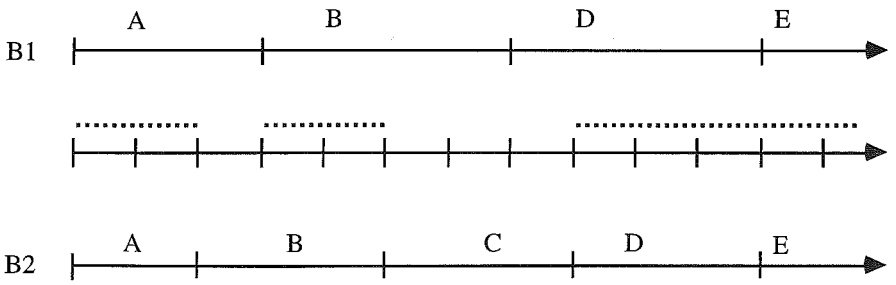


Abbildung 42: Nachträgliche Zeitrasterung; Übereinstimmungsintervalle sind durch Punktierung gekennzeichnet (mod. nach Asendorpf & Wallbott, 1979)

Wie Abbildung 41 zu entnehmen ist, können wir im Anschluß an die nachträgliche Rasterung eine Kontingenztafel erstellen, die zur Berechnung von π oder κ dient. Für das abgebildete Zeitintervall wäre etwa folgende Tafel zu erstellen (Abb. 43).

		B1					
		A	B	C	D	E	
B2	A	2					2
	B	1	2				3
	C		2		1		3
	D				3		3
	E					2	2
		3	4	0	4	2	13

Abbildung 43: Kontingenztafel zu Abb. 42

Es ergibt sich

$$\kappa = \frac{.69-.20}{1-.20} = .62$$

Die Wahl verschiedener Segmentierungspunkte sowie die partiell verschiedene Zuordnung zu den Kategorien führt zu einem nur mäßig hohen κ-Koeffizienten.

Wie eng sollte man nun das Zeitraster wählen? Asendorpf und Wallbott (1979) diskutieren auch diesen Punkt. Bevor wir jedoch darauf eingehen, sollten wir darauf hinweisen, daß wir hier nicht exakt dem Vorschlag Asendorpfs und Wallbotts gefolgt sind. Die Autoren sprechen tatsächlich von Punktmessungen; d.h. sie stellen nicht fest, ob ein Zeit*intervall*, sondern ob zu einem Zeit*punkt* übereinstimmend geratet wurde. Sie entgehen damit dem Problem, daß in Fällen einer Segmentierung *im* Intervall dieses nicht einer Kategorie zugeordnet werden kann. Ähnliche Probleme haben wir aber auch dann, wenn wir a priori eine Zeitrasterteilung wählen; es kommt eben in diesen Fällen darauf an, daß das gewählte Raster relativ klein zur Dauer der Verhaltensweisen ist. Insofern ist Asendorpf und Wallbott zuzustimmen, wenn sie die Wahl der Abtastrate (d.h. die Häufigkeit der Vergleichspunkte pro Zeiteinheit) als problematisch ansehen. Falsch scheint uns ihre Einschätzung, daß „bei zu großer Abtastrate ... und gleichzeitig nur langsam variierendem Verhalten im Zeitverlauf ... die Übereinstimmung überschätzt [wird]" (S. 252). Es ist offensichtlich, daß in unserem Beispiel eine Verdopplung der Abtastrate zu exakt demselben κ-Koeffizienten führen wird, da die relativen Häufigkeiten gleich bleiben (vgl. auch Bakeman & Gottman, 1986). Lediglich bei der inferenzstatistischen Beurteilung, ob κ signifikant von der Zufallserwartung abweicht, erhalten wir künstlich ein zu gutes Ergebnis, da aufgrund einer scheinbar höheren Gesamthäufigkeit geurteilt wird (vgl. die Formel für die Standardabweichung, Abschnitt 5.5.1: n steht im Nenner!). Allerdings gilt - worauf wir schon weiter oben hingewiesen haben -, daß ein signifikant von der Zufallserwartung abweichender Koeffizient noch kein Aufweis der Güte eines Meßinstrumentes ist. Vielmehr wird in der Regel die absolute Höhe des Index bewertet (vgl. Abschnitt 4.5, in dem „Faustregeln" angegeben werden).

Dieses Vorgehen ist natürlich nur möglich, wenn die Zeit während der Beobachtung mitkodiert wurde. Bakeman und Gottman (1986) machen einen Vorschlag zur Bestimmung der Beobachterübereinstimmung, wenn nur die Kodiersequenzen zweier Beobachter vorliegen.

Literaturempfehlungen

Zum Thema Messung in der Psychologie gibt es viele einschlägige Bücher, beispielsweise Gigerenzer (1981). In Kategoriensysteme führt z.B. Faßnacht (1995) ein, die IPA von Bales diskutiert Trolldenier (1985) ausführlich. Eine Einführung in Ratingsysteme bieten Langer und Schulz von Thun (1974), Ratingverfahren werden kritisch diskutiert etwa bei Beck (1987), praktische Hinweise zur Gestaltung von Ratingskalen finden sich auch bei Rohrmann (1978). Erweiterungen zur Beobachterübereinstimmung finden sich etwa bei Asendorpf und Wallbott (1979; dieser Aufsatz setzt allerdings etwas statistisches Wissen voraus).

Kapitel 6

Psychologische Beobachtung: Relikt der Vergangenheit oder Methode mit Zukunft?

Wie häufig wird Beobachtung in der Psychologie nun eigentlich tatsächlich eingesetzt? Wie angesehen ist sie als Methode der Datengewinnung? Wie kaum anders zu erwarten, gehen die Meinungen darüber ziemlich auseinander. Zitieren wir wenigstens ein paar positive Stimmen: „despite its problems, direct observation is perhaps the most valuable assessment tool of behavioral researchers" (Foster & Cone, 1980, S. 332). Beobachtung hat nach Einschätzung derselben Autoren in der klinischen Beurteilung sogar eine eher steigende Bedeutung (Cone & Foster, 1982). Auch nach der Ansicht von Manns et al. „... ist davon auszugehen, daß Beobachtungsverfahren in der Psychologie künftig an Bedeutung gewinnen werden. Das gilt ganz sicher für die Forschung, aber auch für die Praxis." (1987, S. 6). Kelly (1977) zählt in seinem Überblick über die in der Zeitschrift „Journal of Applied Behavioral Analysis" zwischen 1968 und 1975 publizierten Beobachtungsstudien immerhin 222 Arbeiten (von insgesamt 293, d.h. mehr als drei Viertel).

Die ursprüngliche Bedeutung von Beobachtung als Datenquelle hat im Vergleich zu den frühen Tagen des Behaviorismus dennoch nachgelassen. Dies hat sicher mehrere Gründe. Frenz und Frey (1981) weisen etwa darauf hin, daß zwar Beobachtung als Methode in den einschlägigen Artikeln und Handbüchern immer wieder empfohlen werde, aber die Summierung aller möglichen Beobachtungsfehler, die natürlich in diesen Quellen auch immer in aller Ausführlichkeit diskutiert werden (und so ja auch in unserem Buch; siehe Kapitel 3), klinge eher wie eine Sammlung gewichtiger Argumente, warum man Beobachtung als Methode *nicht* verwenden sollte (vgl. auch Foster & Cone, 1980). Eine Konsequenz davon war auch, daß die Literatur zur Berechnung von Beobachterübereinstimmung (als ein Maß zur Schätzung der Fehlerbelastung einzelner Beobachtungen einzelner Beobachter) stark anschwoll (wir haben in Abschnitt 4.5 und 5.5 einige wichtige Maße kennengelernt). Aber selbst die oft berichteten hohen Reliabilitätswerte, ein sonst hoch im Kurs stehendes Kriterium für die Güte von Studien, werden oft bezweifelt oder können häufig die Vorbehalte gegen derartige Untersuchungen nicht zerstreuen.

Zum anderen ist natürlich auch der Behaviorismus, als dessen paradigmatische Methode die objektive Verhaltensbeobachtung ihren Einzug in die Psychologie gehalten hatte, längst wieder

überholt – jedenfalls in Forschung und Wissenschaft, weniger in der Praxis. Neue „Wenden" haben auch ihn schon vor mehr als einem Jahrzehnt abgelöst, wie wir im ersten Kapitel (Abschnitt 1.3) kurz angedeutet haben. Über die handlungstheoretische Perspektive der Psychologie, die in jüngerer Zeit - auch im Kontext von Beobachtungsstudien (Mees, 1988; von Cranach et al., 1980) - an Bedeutung gewinnt, haben wir im Kapitel 2 gesprochen.

Es steht für uns jedoch außer Zweifel, daß auf systematische Beobachtung nicht verzichten kann, wer Wissenschaft betreiben will. Das gilt nicht nur, wenn ein sehr weites Verständnis des Begriffs zugrunde gelegt wird, sondern auch, wenn die engere Auffassung von deduktiver Beobachtung als Methode der Datenerhebung angesprochen ist (vgl. Abschnitt 1.4.2). Letztlich werden wir häufig nachsehen – und das heißt auch: beobachten – müssen, was tatsächlich der Fall ist, um entscheiden zu können, ob unsere Vermutungen darüber, wie die Welt (und also der Mensch) ist, auch Bestand haben. Wie wir im dritten Kapitel gesehen haben, sind auch scheinbar feinere oder differenziertere, anscheinend zuverlässigere und genauere Datenerhebungsmethoden oft nur dann plausible Verbesserungen, wenn sie der Überprüfung durch Beobachtungsdaten standhalten. Damit ist nicht ausgeschlossen, daß diese Verfahren dann, wenn wir bereit sind, ihnen zu vertrauen, Daten liefern können, die hinsichtlich Genauigkeit und Zuverlässigkeit über systematische Beobachtung hinausgehen.

Wir haben darüber hinaus eine Reihe von Beispielen kennengelernt, in denen in wissenschaftlicher Beobachtung der Königsweg zu den Daten liegt, die wir zur Beantwortung unserer Fragen brauchen (beispielsweise die Forschungsbemühungen in der Folge der Arbeit von Rosenthal und Jacobsen, 1968; siehe Abschnitt 1.3). Selbstverständlich aber – und das haben wir in Relation zu ihrer Bedeutung in diesem Buch vielleicht zu wenig betont – liegt ein riesiges Anwendungsfeld für systematische Beobachtung in der psychologischen Praxis, etwa in der Therapie, insbesondere wenn sie mit Personen arbeitet, die, aus welchen Gründen auch immer, über ihre psychische Situation nur in eingeschränktem Maße, nur in verzerrter Art und Weise oder sogar überhaupt nicht sprachlich Auskunft geben können. Systematische Beobachtung in diagnostischer Absicht ist für diese Arbeit unentbehrlich. Daß Beobachtung immer auch die Interpretation von Verhalten im Sinne der Erkundung seiner Bedeutung im konkreten Zusammenhang impliziert, wurde an verschiedenen Stellen dieses Bucher erörtert. Die Fähigkeit, dieses in der Interaktion mit dem Klienten zuverlässig leisten zu können, also in der Therapiesituation selbst als diagnostisches Instrument zu fungieren, ist ein wichtiges Ziel der Ausbildung zum Psychotherapeuten und setzt Kennerschaft nicht nur im Sinne der Erfahrung mit dem Gegenstand einer wissenschaftlichen Untersuchung, sondern nicht zuletzt auch der eigenen Person voraus. Insoweit bildet das Beobachten nicht nur eine Grundlage der wissenschaftlichen Forschung, sondern auch der praktischen psychologischen Arbeit, und zwar sowohl im Sinne der objektiven Verhaltensbeobachtung als auch im Sinne der – in diesem Buch nicht thematisierten – Funktion des „Gewahrseins" der eigenen psychischen Situation (vgl. dazu z.B. Perls, Hef-

ferline & Goodman, 1951). Allerdings ist mit der *Selbstbeobachtung* gleich ein ganzes Bündel von besonderen Problemen angesprochen, die seit der Kritik an der Introspektion als Datenerhebungsmethode ein permanentes Diskussionsthema in der Psychologie geblieben sind (vgl. hierzu z.B. Erdfelder, 1994, S. 55ff.; Bortz & Döring, 1995, S. 246f., 199f.; Greve, 1996).

Insbesondere, wenn sich Forderungen wie das zitierte Plädoyer von Norbert Bischof (1989; vgl. Abschnitt 1.4.2) durchsetzen, daß man Kennerschaft erwerben und sich auf seinen Untersuchungsgegenstand einlassen muß, bevor man mit der Wissenschaft beginnen kann, wird aber auch die Beobachtung als Haltung (heuristische Beobachtung) an Bedeutung gewinnen. Zwar ist es, wie wir angedeutet haben, aus logischer Sicht gleichgültig, ob uns die Vermutungen, Hypothesen und Theorien, die wir untersuchen, im Traum erschienen, unter der Dusche eingefallen sind oder ob wir sie uns in langer disziplinierter gedanklicher Arbeit oder Beobachtung erarbeitet haben. Es könnte aber andere Gründe geben (etwa ökonomische), bestimmte Wege zu präferieren, z.B. wenn sie erfolgreichere Ideen hervorbringen. In diesem Sinne ist Beobachtung als Haltung vermutlich tatsächlich empfehlenswert.

Die zahlreichen Fehlermöglichkeiten, die wir im dritten Kapitel diskutiert hatten, sollten uns dabei nicht entmutigen. Zum einen sind die Beobachtungssysteme, die wir in den Kapiteln 4 und 5 kennengelernt haben, ja auch als Antwort auf diese Schwierigkeiten entwickelt worden; mit ihnen oder durch sie sollen diese Fehler vermieden oder wenigstens verringert werden. Zum zweiten haben wir ja auch über verschiedene Wege der Kontrolle von Beobachtungsfehlern gesprochen, vor allem die wichtigsten Methoden zur Schätzung der Beobachterübereinstimmung. Und schließlich sind diese Fehlerquellen zum allergrößten Teil ja selbst potentielle oder tatsächliche Gegenstände psychologischer Forschung. Viele von ihnen sind ja auch gar nicht zuerst im Zusammenhang mit Beobachtungsstudien entdeckt worden, sondern beispielsweise im Zusammenhang sozialpsychologischer Fragestellungen (wie das Experiment von Asch, 1955). Je mehr wir über die ihnen zugrundeliegenden Phänomene wissen, desto eher werden wir in der Lage sein, sie zu vermeiden, sie auszugleichen oder zu verringern. Und in praktisch allen Fällen ist es nicht allein deswegen interessant, mehr über sie herauszufinden: Wir lernen auch dabei mehr über den Menschen.

Literatur

Asch, S.E. (1955). Opinions and social pressure. *Scientific American, 193*, 31-35.

Asendorpf, J. & Wallbott, H.G. (1979). Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie, 10*, 243-252.

Bain, R.K. (1968). Die Rolle des Forschers: Eine Einzelfallstudie. In R. König (Hrsg.), *Beobachtung und Experiment in der Sozialforschung* (6. Aufl.) (S. 115-128). Köln: Kiepenheuer & Witsch.

Bakeman, R. & Gottman, J.M. (1986). *Observing interaction.* Cambridge: Cambridge University Press.

Bales, R.F. (1950a). *Interaction process analysis. A method for the study of small groups* (zitiert nach dem Nachdruck 1976). Chicago: University of Chicago Press.

Bales, R.F. (1950b). A set of categories for the analysis of small group interaction. *American Sociological Review, 15*, 257-263.

Bales, R.F. (1968). Die Interaktionsanalyse: Ein Beobachtungsverfahren zur Untersuchung kleinerer Gruppen. In R. König (Hrsg.), *Beobachtung und Experiment in der Sozialforschung* (6. Aufl.) (S. 148-167). Köln: Kiepenheuer & Witsch.

Bales, R.F. (1982). Systematische mehrstufige Feldtheorie. In R.F. Bales & S.P. Cohen (Hrsg.), *SYMLOG: ein System für die mehrstufige Beobachtung von Gruppen* (S. 35-253). Stuttgart: Klett-Cotta.

Bales, R.F. & Cohen, S.P. (Hrsg.). (1982). *SYMLOG: ein System für die mehrstufige Beobachtung von Gruppen.* Stuttgart: Klett-Cotta.

Bales, R.F. & Gerbrands, H. (1948). The "interaction recorder". An apparatus and checklist for sequential content analysis of human interaction. *Human Relations, 1*, 456-463.

Bandura, A. (1966). Influence of models' reinforcement contingencies on the acquisition of imitative responses. *Journal of Personality and Social Psychology, 1*, 589-595.

Barash, D.P. (1972). Human ethology: The snack-bar security syndrome. *Psychological Reports, 31*, 577-578.

Barker, R.G. & Wright, H.F. (1961). *One boy's day* (Reprint der ersten Auflage von 1951). New York: Harper.

Barker, R.G. & Wright, H.F. (1971). *Midwest and its children* (Reprint der ersten Auflage von 1955). Hamden: Archon Books.

Bartko, J.J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83*, 762-765.

Bartlett, F.C. (1932). *Remembering.* Cambridge: Cambridge University Press (zit. n. Glass, Holyoak & Santa, 1979).

Baum, C.G., Forehand, R. & Zegiob, L.E. (1979). A review of reactivity in adult-child interactions. *Journal of Behavioral Assessment, 1*, 167-178.

Bayer, G. (1974). Verhaltensdiagnose und Verhaltensbeobachtung. In C. Kraiker (Hrsg.), *Handbuch der Verhaltenstherapie* (S. 255-275). München: Kindler.

Beck, K. (1987). *Die empirischen Grundlagen der Unterrichtsforschung. Eine kritische Analyse der deskriptiven Leistungsfähigkeit von Beobachtungsmethoden.* Göttingen: Hogrefe.

Berry, K.J. & Mielke, P.W.jr. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.

Bierhoff, H.W. (1984). *Sozialpsychologie. Ein Lehrbuch.* Stuttgart: Kohlhammer.

Bieri, P. (1981). Generelle Einführung. In P. Bieri (Hrsg.), *Analytische Philosophie des Geistes* (S. 1-28). Königstein/ Ts.: Hain.

Binet, A. (1897). Psychologie individuelle - la description d'un objet. *Année Psychologique, 3*, 296-332.
Bischof, N. (1989). Emotionale Verwirrungen oder: von den Schwierigkeiten im Umgang mit der Biologie. *Psychologische Rundschau, 40*, 188-205.
Block, I. (1961). *The Q-Sort method in personality assessment and psychiatric research.* Springfield: Thomas.
Blurton Jones, N. (Ed.). (1972). *Ethological studies of child behaviour.* Cambridge: Cambridge University Press.
Bohnen, A. (1972). Zur Kritik des modernen Empirismus. Beobachtungssprache, Beobachtungstatsachen und Theorien. In H. Albert (Hrsg.), *Theorie und Realität* (S. 171-190). Tübingen: Mohr.
Boice, R. (1983). Observational skills. *Psychological Bulletin, 93*, 3-29.
Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation.* Berlin Springer.
Bortz, J. (1993). *Lehrbuch der Statistik: Für Sozialwissenschaftler* (4. vollst. überarbeitete Aufl.). Berlin: Springer.
Brandtstädter, J. (1981). Entwicklung des moralischen Urteils. In H. Werbik & H.J. Kaiser (Hrsg.), *Kritische Stichwörter zur Sozialpsychologie* (S. 88-103). München: Fink.
Brandtstädter, J. (1982). Apriorische Elemente in psychologischen Forschungsprogrammen. *Zeitschrift für Sozialpsychologie, 13*, 267-277.
Brandtstädter, J. (1984). Apriorische Elemente in psychologischen Forschungsprogrammen: Weiterführende Argumente und Beispiele. *Zeitschrift für Sozialpsychologie, 15*, 151-158.
Brandtstädter, J. (1986). Normen und Ziele in der Entwicklungsintervention: Probleme der Konstruktion und Kritik. In K.H. Wiedl (Hrsg.), *Rehabilitationspsychologie. Grundlagen, Aufgabenfelder, Entwicklungsperspektiven* (S. 194-206). Stuttgart: Kohlhammer.
Brandtstädter, J. (Hrsg.) (1987). *Struktur und Erfahrung in der psychologischen Forschung.* Berlin: de Gruyter.
Brandtstädter, J. (1991). Psychologie zwischen Leib und Seele: Einige Aspekte des Bewußtseinsproblems. *Psychologische Rundschau, 42*, 66-75.
Bredenkamp, J. & Wippich, W. (1977). *Lern- und Gedächtnispsychologie* (Bd. 2). Stuttgart: Kohlhammer.
Broad, W. & Wade, N. (1984). *Betrug und Täuschung in der Wissenschaft.* Basel: Birkhäuser.
Brophy, J.E. & Good, T.L. (1976). *Die Lehrer-Schüler-Interaktion.* München: Urban & Schwarzenberg. (Original erschienen 1974: Teacher-student relationships.)
Bühler, K. (1927). *Die Krise der Psychologie.* Stuttgart: Fischer.
Bungard, W. & Lück, H.E. (1974). *Forschungsartefakte und nicht-reaktive Meßverfahren.* Stuttgart: Teubner.

Cairns, R.B. & Green, J.A. (1979). How to assess personality and social patterns: observations or ratings? In R.B. Cairns (Ed.), *The analysis of social interactions: methods, issues, and illustrations* (pp. 209-226). Hillsdale, NJ: Erlbaum.
Carmichael, L., Hogan, H.P. & Walter, A.A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology, 15*, 73-86.
Chalmers, A.F. (1989). *Wege der Wissenschaft.* Berlin: Springer.
Charlton, M., Liebelt, E., Sültz, J. & Tausch, A.-M. (1974). Auswirkung von Verhaltensmodellen aus einem Fernsehwestern auf Gruppenarbeitsverhalten und Aggressionsbereitschaft von Grundschülern. *Psychologie in Erziehung und Unterricht, 21*, 164-175.
Chomsky, N. (1959). [Review of] Verbal behavior. *Language, 35*, 26-58.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.
Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Cone, J.D. & Foster, S.L. (1982). Direct observation in clinical psychology. In P. Kendall & J. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 311-345). New York: Wiley.

Conger, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88*, 322-328.

Cranach, M. von & Frenz, H.-G. (1969). Systematische Beobachtung. In C.F. Graumann (Hrsg.), *Sozialpsychologie* (Handbuch der Psychologie, Bd. 7/1) (S. 269-331). Göttingen: Hogrefe.

Cranach, M. von, Kalbermatten, U., Indermühle, K. & Gugler, B. (1980). *Zielgerichtetes Handeln*. Bern: Huber.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Danziger, K. (1980). The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences, 16*, 241-262.

Dubey, D.R., Kent, R.N., O'Leary, S.G., Broderick, J.E. & O'Leary, K.D. (1977). Reactions of children and teachers to classroom observers. *Behavior Therapy, 8*, 887-897.

Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424.

Ehrhardt, K.J., Findeisen, P., Marinello, G. & Reinartz-Wenzel, H. (1981). Systematische Verhaltensbeobachtung von Aufmerksamkeit im Unterricht: Zur Prüfung von Objektivität und Zuverlässigkeit. *Diagnostica, 27*, 281-294.

Eibl-Eibesfeldt, I. (1984). *Die Biologie des menschlichen Verhaltens. Grundriß der Humanethologie*. München: Piper.

Elashoff, J.D. & Snow, R.E. (1972). *Pygmalion auf dem Prüfstand*. München: Kösel. (Original erschienen 1971: Pygmalion reconsidered.)

Erdfelder, E. (1994). Erzeugung und Verwendung empirischer Daten. In T. Herrmann & W.H. Tack (Hrsg.), *Methodologische Grundlagen der Psychologie* (=Enzyklpädie der Psychologie, Themenbereich B, Serie 1, Band 1; S. 47-97). Göttingen: Hogrefe.

Faßnacht, G. (1995[2]). *Systematische Verhaltensbeobachtung* (1. Aufl. 1979). München: Reinhardt.

Feger, H. (1972). *Skalierte Informationsmenge und Eindrucksurteil*. Bern: Huber.

Feger, H. (1983). Planung und Bewertung von wissenschaftlichen Beobachtungen. In H. Feger & J. Bredenkamp (Hrsg.), *Datenerhebung* (Enzyklopädie der Psychologie, Bd. 1, Forschungsmethoden) (S. 1-75). Göttingen: Hogrefe.

Feger, H. & Graumann, C.F. (1983). Beobachtung und Beschreibung von Erleben und Verhalten. In H. Feger & J. Bredenkamp (Hrsg.), *Datenerhebung* (Enzyklopädie der Psychologie, Bd. 1, Forschungsmethoden) (S. 76-134). Göttingen: Hogrefe.

Fieguth, G. (1977a). Die Entwicklung eines kategoriellen Beobachtungsschemas. In U. Mees & H. Selg (Hrsg.), *Verhaltensbeobachtung und Verhaltensmodifikation* (S. 33-42). Stuttgart: Klett.

Fieguth, G. (1977b). Beobachtertraining. In U. Mees & H. Selg (Hrsg.), *Verhaltensbeobachtung und Verhaltensmodifikation* (S. 78-87). Stuttgart: Klett.

Fisicaro, S.A. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology, 73*, 239-244.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J.L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

Fleiss, J.L. (1983). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley. (zit. nach Bakeman & Gottman, 1986).

Fleiss, J.L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Fleiss, J.L., Cohen, J. & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.
Foster, S.L. & Cone, J.D. (1980). Current issues in direct observation. *Behavioral Assessment, 2*, 313-338.
Frenz, H.-G. & Frey, S. (1981). Die Analyse menschlicher Tätigkeiten - Probleme der systematischen Verhaltensbeobachtung. In F. Frei & E. Ulich (Hrsg.), *Beiträge zur psychologischen Arbeitsanalyse* (S. 57-92). Bern: Huber.
Freud, S. (1938). Abriß der Psychoanalyse. In S. Freud, *Schriften aus dem Nachlaß* (Gesammelte Werke, Bd. 17) (S. 63-138). Frankfurt a.M.: Fischer.
Frick, T. & Semmel, M.I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research, 48*, 157-184.
Friedrichs, J. & Lüdtke, H. (1973). *Teilnehmende Beobachtung* (2. Aufl.). Weinheim: Beltz.

Gardner, H. (1989). *Dem Denken auf der Spur. Der Weg der Kognitionswissenschaft.* Stuttgart: Klett-Cotta. (Original erschienen 1985: The mind's new science. A history of the cognitive revolution.)
Gellert, E. (1955). Systematic observation. *Harvard Educational Review, 25*, 179-195.
Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie.* München: Reinhardt.
Glass, A.L., Holyoak, K.J. & Santa, J.L. (1979). *Cognition.* Reading, MA: Addison-Wesley.
Glennon, B. & Weisz, J.R. (1978). An observational approach to the assessment of anxiety in young children. *Journal of Consulting and Clinical Psychology, 46*, 1246-1257.
Glück, G. (1971). Methoden der Beobachtung. In G. Dohmen (Hrsg.), *Forschungstechnik für die Hochschuldidaktik* (S. 57-66). München: Beck.
Graumann, C.F. (1966). Grundzüge der Verhaltensbeobachtung. In E. Meyer (Hrsg.), *Fernsehen in der Lehrerbildung* (S. 86-107). München: Manz.
Greve, W. (1994). *Handlungsklärung. Die psychologische Erklärung menschlicher Handlungen.* Bern: Huber.
Greve, W. (1996). Erkenne Dich selbst? Argumente zur Bedeutung der „Perspektive der ersten Person". *Sprache & Kognition, 15.* (im Druck)
Greve, W. & Wippermann, D. (1990). *Nur Leib oder auch Seele? Mentale Phänomene als Problem der Psychologie* (Trierer Psychologische Berichte, 17, Heft 4). Trier: Universität, Fachbereich I - Psychologie.
Grümer, K.-W. (1974). *Beobachtung.* Stuttgart: Teubner.

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In G. Lüer (Hrsg.), *Allgemeine Experimentelle Psychologie* (S. 43-264). Stuttgart: Fischer.
Hanson, N.R. (1958/1972). *Patterns of discovery.* Cambridge: Cambridge University Press.
Hasemann, K. (1964). Verhaltensbeobachtung. In R. Heiss (Hrsg.), *Psychologische Diagnostik* (Handbuch der Psychologie, Bd. 6) (S. 807-836). Göttingen: Hogrefe.
Hasemann, K. (1983). Verhaltensbeobachtung und Ratingverfahren. In K.-J. Groffmann & L. Michel (Hrsg.), *Verhaltensdiagnostik* (Enzyklopädie der Psychologie, Bd. B/II/4, S. 434-488). Göttingen: Hogrefe.
Hay, L.R., Nelson, R.O. & Hay, W.M. (1980). Methodological problems in the use of participant observers. *Journal of Applied Behavior Analysis, 13*, 501-504.
Haynes, S.N. (Ed.). (1978). *Principles of behavioral assessment.* New York: Gardner Press.
Herrmann, T. (1973). *Persönlichkeitsmerkmale.* Stuttgart: Kohlhammer.
Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84*, 289-297.
Humpert, W. & Dann, H.-D. (1988). *Das Beobachtungssystem BAVIS.* Göttingen: Hogrefe.
Hussy, W. (1986). *Denkpsychologie. Ein Lehrbuch* (Bd. 2). Stuttgart: Kohlhammer.

Innerhofer, P. (1977). *Das Münchener Trainingsmodell. Beobachtung - Interaktionsanalyse - Verhaltensänderung.* Berlin: Springer.

Innerhofer, P. & Peterander, F. (1981). Quantitative Diagnostik von Mutter-Kind-Interaktionen. In H. Bommert & M. Hockel (Hrsg.), *Therapieorientierte Diagnostik* (S. 194-216). Stuttgart: Kohlhammer.

Isen, A.M. & Levin, P.F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology, 21*, 384-388.

Jacobson, N.S. & Moore, D. (1981). Spouses as observers of the events in their relationship. *Journal of Consulting and Clinical Psychology, 49*, 269-277.

Jäger, A.O. (1986). Validität von Intelligenztests. *Diagnostica, 32*, 272-289.

Jahoda, M., Deutsch, M. & Cook, S.W. (1968). Beobachtungsverfahren. In R. König (Hrsg.), *Beobachtung und Experiment in der Sozialforschung* (6. Aufl.) (S. 77-96). Köln: Kiepenheuer & Witsch.

Jarrett, R.B. & Nelson, R.O. (1984). Reactivity and unreliability of husbands as participant observers. *Journal of Behavioral Assessment, 6*, 131-145.

Johnson, S.M. & Bolstad, O.D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In L.A. Hamerlynck, K.C. Handy & E.J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice.* (pp. 7-67). Champaign, Il: Research Press.

Jorgensen, D.L. (1989). *Participant observation. A methodology for human studies*. London: Sage.

Kalbermatten, U. & Cranach, M. von (1981). Hierarchisch aufgebaute Beobachtungssysteme zur Handlungsanalyse. In P. Winkler (Hrsg.), *Methoden der Analyse von face-to-face-Situationen* (S. 83-127). Stuttgart: Metzlersche Verlagsbuchhandlung.

Kant, I. (1968). *Kritik der reinen Vernunft* (unveränderter photomechanischer Abdruck der von der preußischen Akademie der Wissenschaften 1902 begonnenen Ausgabe von Kants gesammelten Schriften, Bd. III. Original: 1781[A]/1787[B]). Berlin: de Gruyter.

Kazdin, A.E. (1977). Artifact, bias, and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

Kelly, M.B. (1977). A review of observational data-collection and reliability procedures reported in the JABA. *Journal of Applied Behavior Analysis, 10*, 97-101.

Kent, R.N. (1972). *Expectancy bias in behavioral observation.* Unpublished doctoral dissertation, State University of New York, Stony Brook, NY (zit. n. Kent & Foster, 1977).

Kent, R.N. & Foster, S.L. (1977). Direct observational procedures: Methodological issues in naturalistic settings. In A.R. Ciminero, K.S. Calhoun & H.E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 279-328). New York: Wiley.

Kent, R.N., O'Leary, K.D., Diamet, C. & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. *Journal of Consulting and Clinical Psychology, 42*, 774-780.

Klauer, K.C. (1991). *Einstellungen: Der Einfluß der affektiven Komponente auf das kognitive Urteilen*. Göttingen: Hogrefe.

Kluckhohn, F. (1968). Die Methode der teilnehmenden Beobachtung in kleinen Gemeinden. In R. König (Hrsg.), *Beobachtung und Experiment in der Sozialforschung* (6. Aufl.) (S. 97-114). Köln: Kiepenheuer & Witsch.

Koeck, R. & Strube, G. (1977). Beobachtung und Befragung. In G. Strube (Hrsg.), *Binet und die Folgen* (Psychologie des 20. Jahrhunderts, Bd. V) (S. 151-212). Zürich: Kindler.

Kohler, A. (1986). Seven years with SYMLOG: a review of research. *International Journal of Small Group Research, 2*, 83-89.

Krippendorf, K. (1970). Bivariate agreement coefficients for reliability of data. In E.F. Borgotta & G.W. Bohrnstedt (Eds.), *Sociological Methodology* (pp. 139-150). San Francisco: Jossey-Bass.

Kriz, J. (1979). Artefakte in der Sozialforschung. *Zeitschrift für Markt-, Meinungs- und Zukunftsforschung, 22*, 5039-5090.

Kriz, J., Lück, H.E. & Heidbrink, H. (1990). *Wissenschafts- und Erkenntnistheorie*. Opladen: Leske & Budrich.

Langer, I. & Schulz von Thun, F. (1974). *Messung komplexer Merkmale in Psychologie und Pädagogik. Ratingverfahren*. München: Reinhardt.

Lee, M.A.M. (1932). A study of emotional instability in nursery school children. *Child Development, 3*, 142-145. (zit. n. Faßnacht, 1979)

Lenk, H. (1987). Einige wissenschaftstheoretische Probleme der Sozialpsychologie. In H. Lenk, *Zwischen Sozialpsychologie und Sozialphilosophie* (S. 16-42). Frankfurt a.M.: Suhrkamp.

Light, R.J. (1971). Measures of response agreement for qualitative data. Some generalizations and alternatives. *Psychological Bulletin, 76*, 365-377.

Lindsay, P.H. & Norman, D.A. (1981). *Einführung in die Psychologie. Informationsaufnahme und -verarbeitung beim Menschen*. Berlin: Springer. (Original erschienen 1977: Human information processing, 2nd ed.)

Loftus, E.F. & Palmer, J.C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589.

Longabaugh, R. (1980). The systematic observation of behavior in naturalistic settings. In H. Triandis (Ed.), *The handbook of cross-cultural psychology: II. Methodology*. Boston, MA: Allyn & Bacon.

Lovaas, O.I., Koegel, R., Simmons, J.Q. & Lang, J.S. (1973). Some generalizations of follow-up measures on autistic children in behavior therapy. *Journal of Applied Behavior Analysis, 6*, 131-166.

Lüer, G. (Hrsg.). (1987). *Allgemeine experimentelle Psychologie*. Stuttgart: Fischer.

Manns, M., Schultze, J., Herrmann, C. & Westmeyer, H. (1987). *Beobachtungsverfahren in der Verhaltensdiagnostik. Eine systematische Darstellung ausgewählter Beobachtungsverfahren*. Salzburg: Müller.

Manz, W. (1974). Die Beobachtung verbaler Kommunikation im Laboratorium. In J. van Koolwijk & M. Wieken-Mayser (Hrsg.), *Erhebungsmethoden: Beobachtung und Analyse von Kommunikation* (Techniken der empirischen Sozialforschung. Bd. 3) (S. 27-65). München: Oldenbourg.

Margolin, G., Hattem, D., John, R.S. & Yost, K. (1985). Percentual agreement between spouses and outside observers when coding themselves and a stranger dyad. *Behavioral Assessment, 7*, 235-247.

Martin, E. & Wawrinowski, U. (1991). *Beobachtungslehre. Theorie und Praxis reflektierter Beobachtung und Beurteilung*. Weinheim: Juventa.

Masling, J. & Stern, G. (1969). Effect of observer in the classroom. *Journal of Educational Psychology, 60*, 531-540.

Mees, U. (1977a). Methodologische Probleme der Verhaltensbeobachtung in der natürlichen Umgebung: II. Beobachter und Beobachtete als mögliche Fehlerquellen von Beobachtungsdaten. In U. Mees & H. Selg (Hrsg.), *Verhaltensbeobachtung und Verhaltensmodifikation* (S. 66-77). Stuttgart: Klett.

Mees, U. (1977b). Zur Validität von Verhaltensbeobachtungen. In U. Mees & H. Selg (Hrsg.), *Verhaltensbeobachtung und Verhaltensmodifikation* (S. 88-95). Stuttgart: Klett.

Mees, U. (Hrsg.). (1988). *Beobachtung, Interaktionsanalyse und Modifikation aggressiven Kindverhaltens*. Oldenburg: Universitäts-Verlag.

Mees, U. & Selg, H. (Hrsg.). (1977). *Verhaltensbeobachtung und Verhaltensmodifikation*. Stuttgart: Klett.

Mercatoris, M. & Craighead, W.E. (1974). Effects of nonparticipant observation on teacher and pupil classroom behavior. *Journal of Educational Psychology, 66*, 512-519.

Merkens, H. & Seiler, H. (1978). *Interaktionsanalyse*. Stuttgart: Kohlhammer.

Mittenecker, E. (1987). *Video in der Psychologie* (Methoden der Psychologie, Bd. 9). Bern: Huber.

Moya, C.J. (1990). *The philosophy of action. An introduction.* Cambridge: Polity Press.

Nafe, J.P. (1924). An experimental study of the affective qualities. *American Journal of Psychology, 35*, 507-544.
Nisbett, R.E. & Wilson, T.D. (1977). The Halo-effect: Evidence for unconscious alteration of judgements. *Journal of Personality and Social Psychology, 35*, 250-256.
Nissen, G. (1989). Autistische Syndrome. In C. Eggers, R. Lempp, G. Nissen & P. Strunk (Hrsg.), *Kinder- und Jugendpsychiatrie* (5. Aufl.) (S. 518-534). Berlin: Springer.

O'Leary, K.D.& Kent, R. (1973). Behavior modification for social action: research tactics and problems. In L.A. Hamerlynck, K.C. Handy & E.J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice* (pp. 69-96). Champaign, IL: Research Press.
O'Leary, K.D., Kent, R.N. & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis, 8*, 43-51. (zit. n. Kent & Foster, 1977).

Patterson, G.R. & Cobb, J.A. (1973). Stimulus control for classes of noxious behavior. In J.F. Knutson (Ed.), *Control of aggression* (pp. 145-201). Chicago. (zit. n. Fieguth, 1977b).
Pawlik, K. & Buse, L. (1982). Rechnergestützte Verhaltensregistrierung im Feld: Beschreibung und erste psychometrische Überprüfung einer neuen Erhebungsmethode. *Zeitschrift für Differentielle und Diagnostische Psychologie, 3*, 101-119.
Perls, F.S., Hefferline, R. & Goodman, P. (1951). Gestalt therapy. New York: Julian Press.
Perrez, M. & Reicherts, M. (1989). Belastungsverarbeitung: Computerunterstützte Selbstbeobachtung im Feld. *Zeitschrift für Differentielle und Diagnostische Psychologie, 10*, 129-139.
Pfungst, O. (1977). *Der kluge Hans, ein Beitrag zur nicht-verbalen Kommunikation* (2. Aufl.). Frankfurt a.M.: Fachbuchhandlung für Psychologie.
Pinther, A. (1972). Grundprobleme der Beobachtungsmethode. In W. Friedrichs (Hrsg.), *Methoden der marxistisch-leninistischen Sozialforschung* (S. 118-138). Berlin: VEB Deutscher Verlag der Wissenschaften.
Polley, R.B., Hare, A.P. & Stone, P.J. (Eds.). (1988). *The SYMLOG practitioner: applications of small group research.* New York: Praeger.

Rheinberg, F. & Minsel, B. (1986). Psychologie des Erziehers. In B. Weidenmann & A. Krapp (Hrsg.), *Pädagogische Psychologie* (Kap. 9). München: Psychologie Verlags Union.
Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie, 9*, 222-245.
Rommetveit, R. (1980). On "meanings" of acts and what is meant and made known by what is said in a pluralistic social world. In M. Brenner (Ed.), *The structure of action* (pp. 108-149). Oxford: Blackwell.
Rosenshine, B. & Furst, N. (1973). The use of direct observation to study teaching. In R.M.V. Travers (Ed.), *Second handbook of research on teaching* (pp. 122-184). Chicago: Rand McNally.
Rosenthal, R. (1976). *Experimenter effects in behavioral research* (2nd ed.). New York: Appleton.
Rosenthal, R. (1977). Einleitung. Der kluge Hans: Eine Fallstudie für Forschungsfragen. In O. Pfungst, *Der kluge Hans, ein Beitrag zur nicht-verbalen Kommunikation* (2. Aufl.) (S. 7-34). Frankfurt a.M.: Fachbuchhandlung für Psychologie. (Original erschienen 1965: Clever Hans).
Rosenthal, R. & Jacobson, L. (1971). *Pygmalion im Unterricht.* Weinheim: Beltz. (Original erschienen 1968: Pygmalion in the classroom).
Rosenthal, R. & Rosnow, R.L. (Eds.). (1969). *Artifact in behavioral research.* New York: Academic Press.

Ryle, G. (1969). *Der Begriff des Geistes*. Stuttgart: Reclam. (Original erschienen 1949: The concept of mind.)

Schaller, S. (1980). Beobachtungsmethoden in der Verhaltensdiagnostik. In W. Wittling (Hrsg.), *Handbuch der Klinischen Psychologie. Bd. 1: Methoden der klinisch-psychologischen Diagnostik* (S. 130-157). Hamburg: Hoffmann & Campe.

Scharpf, U. (1988). *Entscheidungsfindung im Gruppenprozeß*. Konstanz: Hartung-Gorre.

Scherer, K.R. (1974). Beobachtungsverfahren zur Mikroanalyse non-verbaler Verhaltensweisen. In J. van Koolwijk & M. Wieken-Mayser (Hrsg.), *Techniken der empirischen Sozialforschung* (S. 66-109). München: Oldenbourg.

Schmitt, M. (1990). *Konsistenz als Persönlichkeitseigenschaft?* Berlin: Springer.

Schwartz, S. (1988). *Wie Pawlow auf den Hund kam ...: Die 15 klassischen Experimente der Psychologie*. Weinheim: Beltz. (Original erschienen 1987: Pavlov's heirs. Classic psychology experiments that changed the way we view ourselves)

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Seiffert, H. (1973). *Einführung in die Logik*. München: Beck.

Shuller, D.Y. & McNamara, J.R. (1976). Expectancy factors in behavioral observation. *Behavior Therapy, 7*, 519-527.

Simon, A. & Boyer, E.G. (1974). *Mirrors for behavior. An anthology of observation instruments*. Communication Materials Center, Wyncote, Pennsylvania.

Skinner, B.F. (1957). *Verbal behavior*. New York: Appleton.

Smedslund, J. (1978). Bandura's theory of self-efficacy: A set of common-sense theorems. *Scandinavian Journal of Psychology, 19*, 1-14.

Smedslund, J. (1984). What is necessarily true in psychology? *Annals of Theoretical Psychology, 2*, 241-272.

Stegmüller, W. (1974). *Theorie und Erfahrung. Erster Halbband: Begriffsformen, Wissenschaftssprache, empirische Signifikanz und theoretische Begriffe* (Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Bd. II) (verbesserter Neudruck). Berlin: Springer.

Stegmüller, W. (1979). *Rationale Rekonstruktion von Wissenschaft und ihrem Wandel*. Stuttgart: Reclam.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

Strunk, P. (1989). Emotionale Störungen mit vorwiegend somatischer Symptomatik. In C. Eggers, R. Lempp, G. Nissen & P. Strunk (Hrsg.), *Kinder- und Jugendpsychiatrie* (5. Aufl.) (S. 189-262). Berlin: Springer.

Suppes, P. & Zinnes, J.L. (1963). Basic measurement theory. In R.D. Luce, R.B. Bush & E. Galanter (Eds.), *Mathematical Psychology,* (Vol. I, pp. 1-77). New York: Wiley.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin, 52*, 1-23.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29.

Trolldenier, H.-P. (1985). *Verhaltensbeobachtung in Erziehung und Unterricht mit der Interaktionsprozeßanalyse*. Frankfurt a.M.: Fachbuchhandlung für Psychologie.

Uebersax, J.S. (1982). A generalized kappa coefficient. *Educational and Psychological Measurement, 42*, 181-183.

Velden, M. (1982). *Die Signalentdeckungstheorie in der Psychologie*. Stuttgart: Kohlhammer.

Watson, J.B. (1913). Psychology as a behaviorist views it. *Psychological Review, 20*, 158-177.

Weinrott, M.R., Garrett, B. & Todd, N. (1978). The influence of observer presence in classroom behavior. *Behavior Therapy, 9*, 900-911.

Werner, J. (1976). Varianzanalytische Maße zur Reliabilitätsbestimmung von ratings. *Zeitschrift für Experimentelle und Angewandte Psychologie, 23*, 489-500.
Whipple, G.M. (1909). The observer as a reporter: A survey of the 'psychology of testimony'. *Psychological Bulletin, 6*, 153-170.
Wildman, B.G. & Erickson, M.T. (1977). Methodological problems in behavioral observation. In J.D. Cone & R.P. Hawkins (Eds.), *Behavioral assessment. New directions in clinical psychology* (pp. 241-282). New York: Brunner & Mazel.
Winer, B.J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill (zit. nach Asendorpf & Wallbott, 1979).
Winer, B.J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.
Wittgenstein, L. (1984). *Philosophische Untersuchungen.* Frankfurt a.M.: Suhrkamp.

Young, P.T. (1927). Studies in affective psychology. *American Journal of Psychology, 38*, 157-193.

Zapf, D. (1989). *Selbst- und Fremdbeobachtung in der psychologischen Arbeitsanalyse: methodische Probleme bei der Erfassung von Streß am Arbeitsplatz.* Göttingen: Hogrefe.
Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 374-378.

Anhang

Die Intraklassen-Korrelation als Übereinstimmungsmaß

Üblicherweise wird in der Literatur zur Beobachterübereinstimmung als Maß der Übereinstimmung für Ratingskalen die sogenannte Intraklassenkorrelation eingeführt (vgl. z.B. Asendorpf & Wallbott, 1979; Feger, 1983). Genauer gesagt werden verschiedene Formen dieser Korrelation vorgestellt, die - genau wie wir es bei den einfachen Kreuztabellenmaßen π und κ kennengelernt haben - die einzelnen Fehlermöglichkeiten unterschiedlich stark berücksichtigen.

In diesem Abschnitt soll nun kurz die Intraklassenkorrelation eingeführt werden. Wir halten dies für notwendig, um einen Anschluß an die weiterführende Literatur zu schaffen; dabei wird es sich allerdings nicht ganz vermeiden lassen, einige Begriffe etwas knapp unter Verweis auf entsprechende Vertiefungslektüre einzuführen; daher diese Form des Anhangs. Darüber hinaus ist es sinnvoll, den Zusammenhang zwischen der Intraklassenkorrelation und der in der Psychologie häufig genutzten „Produkt-Moment-Korrelation" zu klären. Einerseits wird so dem Eindruck entgegengewirkt, die Statistik der Beobachterübereinstimmung stehe unverbunden neben der Statistik, wie sie in den Einführungskursen des Psychologiestudiums gelehrt wird; andererseits läßt sich die Problematik der Beobachterübereinstimmung noch einmal sehr schön aufzeigen, wie wir gleich sehen werden.

(1) Die Intraklassen-Korrelation ohne Beobachteradjustierung

Zunächst ist es wichtig, einige wenige Begriffe einzuführen. Als erstes müssen wir den Übergang von der Kontingenztafel als Rohdatendarstellung zur Variablenbetrachtung leisten. Nehmen wir etwa - analog zu unserem Beispiel im Abschnitt 5.5 - eine Ratingskala mit den Stufen von 1 bis 5. Statt nun jede Beobachtungseinheit in das entsprechende Feld der Kontingenztafel einzutragen, schreiben wir einfach die Ratings der beiden Beobachter nebeneinander, wie es in der Tabelle 4 zu sehen ist.

Tab. 4: Variablendarstellung der Ratings zweier Beobachter

	B1	B2
Einheit 1	3	5
Einheit 2	4	2
Einheit 3	3	3
Einheit 4	2	1
Einheit 5	3	4
Einheit 6	1	1

Es liegen also als Ausgangsbasis für die Korrelationsberechnung zwei Variablen B1 und B2 vor.

Der zentrale Begriff in den Überlegungen zur Intraklassen-Korrelation ist der der Varianz. Die Varianz ist ein Maß für die Schwankung einer Reihe von Werten um ihren Mittelwert. Genauer gesagt ist sie die mittlere quadrierte Abweichung der Werte vom Mittelwert:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Diese Formel gilt allerdings nur, wenn wir die Varianz als deskriptives Maß zur Beschreibung der vorliegenden Stichprobe von Meßwerten heranziehen. In der Regel sind wir aber daran interessiert, eine Schätzung der Verhältnisse in der Population, der Gesamtheit aller Werte zu erfahren. In unserem Beispiel bedeutet das, daß nicht die Schwankung der 6 Meßwerte von B1 interessiert, sondern welche Varianz bei allen (also im Prinzip unendlich vielen) Ratings von B1 in der definierten Beobachtungssituation zu erwarten ist. Um diese Schätzung zu erhalten, teilt man den Zähler (die sogenannte Quadratsumme) nicht durch die Anzahl der Werte, sondern durch die sogenannten Freiheitsgrade. Was heißt das? Wir müssen uns bei der Berechnung einer statistischen Größe immer vergegenwärtigen, wieviele der eingehenden Werte unabhängig voneinander variieren können. Im Fall der Varianzformel gehen die n Meßwerte x_i und der Mittelwert $\bar{x}$ in die Formel ein. Wenn wir $\bar{x}$ als gegebene Größe betrachten, kennen wir aufgrund von n-1 Meßwerten den letzten, den n-ten Meßwert. Es können also nur n-1 Werte unabhängig voneinander variieren. Wir sagen, die Varianz habe n-1 Freiheitsgrade, so daß sich die Formel modifiziert zu:

$$\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

(Zur exakten Herleitung der Freiheitsgrade der Varianz vgl. Bortz, 1993, Anhang B, S. 630f.) So beträgt etwa die Varianz von B1 im Beispiel 1.07, die von B2 2.67.

Für die Herleitung der Intraklassenkorrelation stellen wir jetzt folgende Überlegung an: Die Gesamtvarianz aller Werte (also der Werte von B1 *und* B2) läßt sich nach dem gleichen Muster berechnen. Wir berechnen zunächst den Gesamtmittelwert über die beiden Beobachtermeßreihen, bilden die Quadratsumme und teilen durch die Freiheitsgrade. Um dies ebenfalls in einer Formel auszudrücken, müssen wir zunächst wieder eine Notation einführen.

Tab. 5: Notation für die Varianzberechnung

	B_1	B_2	...	B_j	...	B_p	
E_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}	e_1
E_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}	e_2
...	...	...	...	...	...	...	...
E_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}	e_i
...	...	...	...	...	...	...	...
E_n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}	e_n
	b_1	b_2	...	b_j	...	b_p	g

In Tabelle 5 stehen E_1 bis E_n für die Beobachtungseinheiten, die durch die Beobachter B_1 bis B_p durch die Werte x_{11} bis x_{np} eingeschätzt werden. Im einfachsten Fall handelt es sich natürlich um nur zwei Beobachter. Die nachfolgenden Formeln machen hierzu aber keinerlei Voraussetzung, so daß wir schon hier diesen Vorzug einer solchen Betrachtung der Intraklassen-Korrelation gegenüber π_w und κ_w zwanglos einführen können. Die Platzhalter b_1 bis b_p sowie e_1 bis e_n stehen für die Mittelwerte der Beobachter (über alle Einheiten hinweg) bzw. die Mittelwerte der Einheiten (über alle Beobachter hinweg). Schließlich ist g der Gesamtmittelwert aller Meßwerte.

Die Varianz aller Meßwerte ergibt sich in dieser Notation nach folgender Formel:

$$\hat{\sigma}^2_{gesamt} = \frac{\sum_{i,j}(x_{ij} - g)^2}{n * p - 1}$$

(Zur Erläuterung: Im Zähler steht eine Doppelsumme; i „läuft" von 1 bis n und j von 1 bis p. Der Nenner ergibt sich dadurch, daß wir insgesamt n*p Werte haben; die Subtraktion von 1 ergibt sich aus dem gleichen Grund wie oben.)

Die nächste Überlegung betrifft die „Quellen" dieser gesamten Varianz. Wie kommen alle diese Abweichungen vom globalen Mittelwert zustande? Wir können folgendes Modell aufstellen: Jeder einzelne Meßwert x_{ij} ergibt sich als Summe aus globalem Mittelwert (μ), einem Wert, der die Einwirkung der Beobachtungseinheit darstellt (τ), und einem Meßfehler (ε):

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Zur Erläuterung wollen wir ein einfaches Beispiel nehmen: Denken wir etwa an eine einfache Ratingskala zur Einschätzung von Angst (vgl. Abschnitt 5.4) und an eine Untersuchung, in der die einzelne Versuchsperson Einheit ist. Wir lassen etwa die „Ängstlichkeit" verschiedener

Kinder von mehreren Beobachtern einschätzen: Welches Kind zeigt mehr, welches weniger Angst? Der Parameter τ_i steht dann für die Ängstlichkeit des Kindes i, ausgedrückt als Abweichung vom Gesamtmittelwert (es gilt dabei, daß die Summe der τ_i Null ergibt). Der Parameter ε_{ij} ist ein Wert einer Zufallsvariablen, die als um Null verteilt angenommen wird. Sie steht für alle unsystematischen Fehlereinflüsse, mit denen der Meßwert belastet ist. In dieses Modell geht nicht ein, daß dieselben Beobachter alle Einheiten einschätzen; d.h. es wird lediglich berücksichtigt, daß pro Einheit p beliebige, ungeordnete Ratings vorliegen.

Die Gesamtvarianz der Meßwerte ergibt sich aus der Varianz der τ (σ_τ^2) und der Varianz der ε (σ_ε^2). Um ein Gütemaß für die Beobachtung zu bekommen, müssen diese beiden Varianzanteile aufeinander bezogen werden. Tatsächlich ist die Intraklassen-Korrelation so definiert (vgl. z.B. Feger, 1983):

$$IC_u = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}$$

Wir beziehen also die Varianz der tatsächlichen Einheitenunterschiede auf die Summe aus dieser und der Fehlervarianz. Wird die Fehlervarianz Null, nimmt IC_u den Wert 1 an, wird die Varianz der tatsächlichen Einheitenunterschiede Null, wird IC_u auch Null. (Da wir - wie oben schon angekündigt - mehrere Arten der Intraklassen-Korrelation einführen wollen, haben wir schon hier einen Index vergeben: IC_u = „Intraclass, unjustiert"; diese Terminologie übernehmen wir von Asendorpf und Wallbott, 1979.) Die nächste Aufgabe ist nun, die beiden verschiedenen Varianzanteile aufgrund der Stichprobendaten zu schätzen. Tatsächlich ist es möglich, die „Quellen" der Varianz zu trennen.

Wie erhalten wir zunächst eine Schätzung der Fehlervarianz? Hierzu betrachten wir die Meßreihen für die einzelnen Einheiten: Offenbar sind die Unterschiede der einzelnen Beobachter bei der Beurteilung genau das, was wir als Nicht-Übereinstimmung oder Fehler bezeichnen wollen. Wir berechnen also die Varianzen für die einzelnen Meßreihen (d.h. die Varianzen, die auf den Abweichungen vom Einheitenmittelwert beruhen) und bilden ihren Durchschnitt. Um die „berechenbaren" von den theoretischen Varianzen (σ_τ^2, σ_ε^2) abzuheben, wollen wir in Übereinstimmung mit vielen Veröffentlichungen von „mittleren Quadratsummen", MS, sprechen (vgl. Asendorpf & Wallbott, 1979; Werner, 1976; beachte dagegen Bortz, 1993). Die Varianz für eine einzelne Einheit E_i ist:

$$MS_{Einh.i} = \frac{\sum_j (x_{ij} - e_i)^2}{p - 1}$$

Die Freiheitsgrade ergeben sich wieder aus der gleichen Überlegung wie oben.

Man mittelt mehrere Varianzen, indem die Summe der Quadratsummen durch die Summe der Freiheitsgrade geteilt wird (vgl. Bortz, 1993, S. 231); also ist die Formel für die Fehlervarianz:

$$MS_{Fehler} = \frac{\sum_{i,j}(x_{ij} - e_i)^2}{n*(p-1)}$$

Ist die MS_{Fehler} nun eine gute Schätzung des von uns gesuchten Populationsparameters σ_ε^2? Man kann zeigen, daß unter der Annahme gleicher Varianzen $MS_{Einh.i}$ gilt (vgl. Bortz, 1995, Kap. 12):

$$E(MS_{Fehler}) = \sigma_\varepsilon^2$$

Diese Formel liest sich so: Der Erwartungswert der MS_{Fehler} ist gleich der σ_ε^2; d.h. ziehen wir unendlich häufig Stichproben und berechnen für jede Stichprobe MS_{Fehler}, dann ist der Mittelwert dieser Werte die Populationsvarianz σ_ε^2. Wir dürfen also die aus den Stichprobendaten berechnete MS_{Fehler} als Schätzung für σ_ε^2 nutzen.

Wenden wir uns nun σ_τ^2 zu. Um sie zu berechnen, gehen wir von folgender Überlegung aus: Wenn es ausschließlich Unterschiede zwischen den Beobachtungseinheiten, aber keine Fehlervarianz gäbe, hätte jeder Beobachter die gleiche Einschätzung bei je einer Einheit vorgenommen. Die Werte einer Beobachtungseinheit x_{i1} bis x_{ip} wären also jeweils alle gleich. Um diese Situation zu „simulieren", nehmen wir als beste Schätzung jeweils den Stichprobenmittelwert e_i; Tabelle 6 verdeutlicht dies.

Tab. 6: „Simulation" der Situation ohne Beobachterunterschiede

	B_1	B_2	...	B_j	...	B_p	
E_1	e_1	e_1	...	e_1	...	e_1	e_1
E_2	e_2	e_2	...	e_2	...	e_2	e_2
...	...	...	...	...	...	...	...
E_i	e_i	e_i	...	e_j	...	e_i	e_i
...	...	...	...	...	...	...	...
E_n	e_n	e_n	...	e_n	...	e_n	e_n
	g	g	...	g	...	g	g

Welche Gesamtvarianz würden wir erhalten, wenn sie aufgrund dieser fiktiven Meßwerte berechnet würde? Der Zähler in der Varianzformel (die Quadratsumme $QS_{Beob.einheit}$) ergibt sich ganz einfach: für jede Beobachtungseinheit geht p-mal die quadrierte Abweichung des Einheitenmittelwertes vom Gesamtmittelwert in die Summe ein. Als Formel sieht das so aus:

$$QS_{Beob.einheit} = \sum_i p*(e_i - g)^2$$

Die Freiheitsgrade (also der Nenner der Varianzformel) ergeben sich nach folgender Überlegung: nach unserer Annahme, daß keine Fehlervarianz existiert, können die einzelnen Werte der Beobachter für eine Einheit nicht mehr variieren; dagegen sind die Mittelwerte der Einheiten unabhängig voneinander. Wir müssen allerdings daran denken, daß bei gegebenem Gesamtmittelwert dies wiederum nur für n-1 Einheiten gilt. Die Varianz der Beobachtungseinheiten ergibt sich somit zu:

$$MS_{Beob.einheit} = \frac{\sum_i p*(e_i - g)^2}{n-1}$$

Ist $MS_{Beo.einheit}$ eine Schätzung der von uns gesuchten σ_τ^2? Nein, da wir in unserer „Simulation" die meßfehlerbelasteten Einheitenmittelwerte e_i benutzt haben. Es läßt sich zeigen, daß folgende Beziehung gilt (vgl. Bortz, 1993, Kap. 12):

$$E(MS_{Beob.einheit}) = p*\sigma_\tau^2 + \sigma_\varepsilon^2$$

Die $MS_{Beob.einheit}$ setzt sich also aus σ_τ^2 und σ_ε^2 zusammen. Wir können den letzten Ausdruck umformen und erhalten:

$$\sigma_\tau^2 = \frac{E(MS_{Beob.einheit}) - \sigma_\varepsilon^2}{p}$$

Da wir die σ_ε^2 schon kennen, können wir also auch unsere σ_τ^2 berechnen:

$$\sigma_\tau^2 = \frac{E(MS_{Beob.einheit}) - E(MS_{Fehler})}{p}$$

Damit ist es auch möglich, die Intraklassenkorrelation IC_u aufgrund der Stichprobenwerte zu schätzen:

$$IC_u = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}$$

$$= \frac{\frac{E(MS_{Beob.einheit}) - E(MS_{Fehler})}{p}}{\frac{E(MS_{Beob.einheit}) - E(MS_{Fehler})}{p} + E(MS_{Fehler})}$$

$$= \frac{E(MS_{Beob.einheit}) - E(MS_{Fehler})}{E(MS_{Beob.einheit}) + (p-1) * E(MS_{Fehler})}$$

Wir dürfen in diese Formel für $E(MS_{Beob.einheit})$ die Berechnungsformel für $MS_{Beob.einheit}$ einsetzen; für $E(MS_{Fehler})$ gilt das Entsprechende.

Man nennt dieses Aufsplitten nach Varianzquellen „Varianzanalyse". Dieses Verfahren spielt in der psychologischen Forschung eine große Rolle, da dort häufig die Frage zu entscheiden ist, ob Unterschiede zwischen den Stichproben eines Experiments (z.B. die Kinder in Banduras Untersuchung, die das bestrafte Modell gesehen haben, versus die Kinder, denen die belohnte Person gezeigt wurde; vgl. Abschnitt 1.1) größer sind als die Unterschiede innerhalb der Gruppen. Das wesentliche Prinzip dabei ist die „Quadratsummenzerlegung", denn es gilt:

$$QS_{gesamt} = QS_{Beob.einheit} + QS_{Fehler}$$

Ebenso setzen sich die Freiheitsgrade (df, „degrees of freedom") additiv zusammen:

$$df_{gesamt} = df_{Beob.einheit} + df_{Fehler}$$

Bortz (1993) erläutert die Varianzanalyse in seinem Statistik-Lehrbuch sowohl recht pragmatisch mit Berechnungsregeln (Kap. 7f.) als auch in ihren theoretischen Grundlagen (Kap. 12).

Welche Eigenschaft hat die Intraklassen-Korrelation IC_u? Da jede Art der Nicht-Übereinstimmung als Beitrag zur Fehlervarianz gilt, ist IC_u ein sehr strenges Gütemaß. So werden alle systematischen Beobachterunterschiede, etwa unterschiedliche „Anker", wie sie sich in den unterschiedlichen Beobachtermittelwerten b_j ausdrücken, als Fehler gewertet. Es deutet sich eine Parallele zum π-Koeffizienten an: Auch dort hatten wir den „Fehler der unterschiedlichen Bereitschaft" als Fehler angesehen (vgl. Abschnitt 4.5). Tatsächlich läßt sich nachweisen, daß IC_u bis auf einen Term, der mit wachsendem n gegen Null geht (also in der Regel vernachlässigbar ist), äquivalent zu π_w ist.[34]

Gerade bei Ratingskalen ist es häufig nicht so entscheidend, ob Beobachter verschiedene Ankerpunkte haben: Wenn stets dieselben Personen das Geschehen beurteilen, mitteln wir zur weiteren Datenverarbeitung die Ratings dieser Beobachter, so daß globale Mittelwertsunterschiede nicht so problematisch sind und nicht als Fall von mangelnder Reliabilität angesehen zu werden brauchen. Deshalb wollen wir einen zweiten Koeffizienten herleiten, bei dem die Mittelwertsunterschiede berücksichtigt werden.

---

[34] Hier ist nicht der Raum, um den Beweis der Äquivalenz auszuführen. Er ist in einem kleinen Arbeitspapier dargelegt, das bei D. Wentura angefordert werden kann. Darüber hinaus weist Krippendorf (1970) die Äquivalenz von π_w mit einer formal etwas anders hergeleiteten IC_u nach.

(2) Die Intraklassen-Korrelation mit Beobachteradjustierung

Wir können die Unterschiedlichkeit der Mittelwerte der Beobachter ebenfalls als Quelle der Varianz ansehen. Wir gehen dazu einfach von einem erweiterten Modell aus:

$$x_{ij} = \mu + \tau_i + \beta_j + \rho_{ij}$$

Wir nehmen in die Bestimmungsformel für den einzelnen Meßwert noch einen Term β_j auf, der für die tatsächlichen Unterschiede zwischen den Beobachtern steht, wie sie sich in den Mittelwertsunterschieden ausdrücken. Wir übernehmen damit das Modell der „Varianzanalyse mit Meßwiederholung", da explizit der Gedanke mit aufgenommen wird, daß jede Einheit von jedem Beobachter geratet wird. Aus diesem Grund wird der letzte Term der Gleichung „ρ_{ij}" genannt, da der Fehleranteil in zwei Komponenten zerlegt werden kann (vgl. Bortz, 1989, Kap. 12). Wir wollen darauf nicht weiter eingehen, da dies für die folgenden Überlegungen irrelevant ist; die σ^2_ε des oberen Modells entspricht der σ^2_ρ des neuen Modells.

Zusätzlich zu den Varianzen σ^2_τ und σ^2_ρ müssen wir aber nun auch noch die σ^2_β mit einbeziehen. Diese weitere Varianz berechnet sich ganz analog zu σ^2_τ: Wir können eine Varianz der Beobachter ähnlich wie die Varianz der Beobachtungseinheiten bestimmen. Hierzu gehen wir von folgender Überlegung aus: Welche Gesamtvarianz würden wir erhalten, wenn alle Beobachtungseinheiten von jedem Beobachter mit seinem persönlichen Mittelwert b_j eingeschätzt würden?

$$MS_{Beobachter} = \frac{\sum_j n * (b_j - g)^2}{p - 1}$$

Unter einer für die Beobachtung zutreffenden Annahme[35] steht die $MS_{Beobachter}$ in folgendem Verhältnis zur σ^2_β (vgl. Bortz, 1993, S. 390ff.):

$$E(MS_{Beobachter}) = n * \sigma^2_\beta + \sigma^2_\rho$$

Wir können den Ausdruck umformen und erhalten:

$$\sigma^2_\beta = \frac{E(MS_{Beobachter}) - \sigma^2_\rho}{n}$$

Es verändert sich jetzt auch die σ^2_ρ bzw. die $MS_{Residual}$. Die Herleitung der entsprechenden Quadratsumme ist jetzt nicht mehr ganz so einfach vorführbar. Wir nutzen deshalb das Prinzip

[35] Wir meinen die Annahme, daß die Beobachtungseinheiten eine zufällige Auswahl aus den gesamten Einheiten darstellen. In der Varianzanalyse spricht man von einem „random factor".

der Quadratsummenzerlegung; wie oben schon beschrieben, kann man die gesamte Quadratsumme in ihre Teilkomponenten zerlegen:

$$QS_{gesamt} = QS_{Beob.einheit} + QS_{Beobachter} + QS_{Residual}$$

Durch Umformung erhalten wir:

$$QS_{Residual} = QS_{gesamt} - QS_{Beob.einheit} - QS_{Beobachter}$$

Ebenso setzen sich die Freiheitsgrade additiv zusammen, so daß gilt:

$$df_{Residual} = df_{gesamt} - df_{Beob.einheit} - df_{Beobachter}$$

$$= (n*p-1) - (n-1) - (p-1)$$

$$= (n-1)\,(p-1)$$

So ergibt sich:

$$MS_{Residual} = \frac{QS_{gesamt} - QS_{Beob.einheit} - QS_{Beobachter}}{(n-1)*(p-1)}$$

Es gilt unter den Annahmen der Varianzanalyse mit Meßwiederholung (s. Bortz, 1993, Kap. 12.3):

$$E(MS_{Residual}) = \sigma_\rho^2$$

Wie definieren wir nun die Intraklassenkorrelation nach dem neuen Modell? Es gibt vor allem zwei Möglichkeiten, die sich danach unterscheiden, wie die σ_β^2 einbezogen wird. Die erste Möglichkeit geht davon aus, daß σ_β^2 als systematische Ratervarianz nicht die Gütebeurteilung belasten sollte. Dementsprechend lautet die Intraklassen-Formel in diesem Fall (IC_a für „Intraclass, adjustiert"):

$$IC_{a1} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\rho^2}$$

Die σ_τ^2 wird also nur zur Residualvarianz in Beziehung gesetzt. Die Berechnungsformel für IC_{a1} ist dann:

$$IC_{a1} = \frac{E(MS_{Beob.einheit}) - E(MS_{Residual})}{E(MS_{Beob.einheit}) + (p-1) * E(MS_{Residual})}$$

Dieser Koeffizient ist in der Reihe der Intraklassen-Koeffizienten der am wenigsten strenge: Unterschiede der Ratermittelwerte verändern diesen Koeffizienten nicht. Ebel (1951), Winer (1971) und Werner (1976) geben eine solchermaßen beobachteradjustierte Formel an. Vor allem bei der Verwendung einer festen Ratergruppe für die gesamte Untersuchung scheint uns dieser Koeffizient pragmatisch begründbar.

Die zweite Variante geht von folgender Logik aus: Wir behalten die abstrakte Form der IC_u bei; d.h. die systematische Beobachtungseinheitenvarianz wird relativiert auf die gesamte Varianz, die sich aus systematischer und Fehlervarianz zusammensetzt. Bei unserem neuen Modell, welches die Beobachter als weitere Quelle von Varianz betrachtet, müßte dementsprechend der Nenner der Intraklassenkorrelation drei Komponenten enthalten; die nächste Formel drückt diesen Gedanken aus.

$$IC_{a2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\beta^2 + \sigma_\rho^2}$$

Als erstes ist festzustellen, daß IC_{a2} stets kleiner ist als IC_{a1}, da ja im Nenner ein zusätzlicher Ausdruck erscheint. IC_{a2} ist also der strengere Test der Reliabilität, da auch Mittelwertsunterschiede den Koeffizienten nicht unbelastet lassen. Er ist dann angemessen, wenn verschiedene Ratergruppierungen benutzt werden. Wir können dann aufgrund der einen Ratergruppe, die die Daten für die IC-Berechnung geliefert hat, besser die Reliabilität beliebiger Gruppen einschätzen. Diese Form der Adjustierung bevorzugen Bartko (1976) und Fleiss und Cohen (1973).

Es stellt sich aber auch die Frage, in welchem Verhältnis IC_{a2} zu IC_u steht. Man könnte meinen, daß die beiden Koeffizienten gleich sein müßten, da in beiden Fällen die gesamte Varianz im Nenner steht. Tatsächlich ist aber IC_u in der Regel kleiner als IC_{a2} , da σ_ρ^2 (ebenfalls in der Regel) kleiner ist als σ_ε^2 und somit in der Formel zur Berechnung von σ_τ^2 ein kleinerer Wert subtrahiert wird. Der Zähler im Ausdruck von IC_{a2} wird also größer. Wir wollen auch für IC_{a2} eine Berechnungsformel angeben (die Herleitung ist analog zu der entsprechenden Herleitung bei IC_u):

$$IC_{a2} = \frac{E(MS_{Beob.einheit}) - E(MS_{Residual})}{E(MS_{Beob.einheit}) + (p-1) * E(MS_{Residual}) + \frac{p}{n}(E(MS_{Beobachter}) - E(MS_{Residual}))}$$

Fleiss und Cohen (1973) konnten zeigen, daß κ_W bis auf einen Term, der bei wachsendem n gegen Null geht, identisch mit IC_{a2} ist.[36]

Wir verlassen jetzt die Intraklassen-Koeffizienten und führen die Produkt-Moment-Korrelation ein. Es sei schon hier gesagt, daß die Produkt-Moment-Korrelation in der Regel nicht als Gütemaß empfohlen wird; wir werden gleich sehen, aus welchem Grund. Die Produkt-Moment-Korrelation ist allerdings der Korrelationstyp, der in der psychologischen Forschung die größte Rolle spielt und deswegen in den Statistik-Anfängerkursen gelehrt wird. Wir wollen deshalb nicht versäumen, Bezüge zwischen ihr und den Intraklassen-Korrelationen herzustellen.

(3) Die Produkt-Moment-Korrelation

Die Produkt-Moment-Korrelation ist ein Maß für den (linearen) Zusammenhang zweier Variablen. Was heißt das? Angenommen, wir interessieren uns für den Zusammenhang von Intelligenz mit der Schulleistung. Wir nehmen eine Stichprobe von Schülern, lassen sie einen Intelligenztest bearbeiten und erfragen etwa ihre Zeugnisdurchschnittsnote. Von jedem Schüler liegt dann ein Wert der Variable „Testintelligenz" und ebenso ein Wert der Variable „Durchschnittsnote" vor. Wir wollen nun wissen, ob ein hoher (niedriger) Testintelligenzwert im Durchschnitt mit einer niedrigen (hohen) Durchschnittsnote einhergeht. Das kann man noch präzisieren: Korrespondiert einer Abweichung vom Mittelwert auf der Intelligenzvariable eine entsprechende Abweichung vom Mittelwert auf der Notenvariable? Das ist offenbar immer noch nicht präzise genug: Was heißt „entsprechende Abweichung"? Wenn wir bei dem Intelligenztest von den üblichen Werten, die deutlich um den Wert 100 schwanken, ausgehen und uns verdeutlichen, daß die Noten prinzipiell nur von 1.0 bis 6.0 gehen können, so wird klar, daß wir ein Maß für die Schwankung der Werte berücksichtigen müssen. Üblicherweise wird die Varianz (s.o.) oder die sogenannte Standardabweichung (SD), die Wurzel aus der Varianz, benutzt. Die Produkt-Moment-Korrelation wird so definiert (x und y seien unsere beiden Variablen):

$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n}}{SD_x * SD_y}$$

Der Zähler dieser Formel - die sogenannte „Kovarianz" - ist so konstruiert, daß einander im Vorzeichen entsprechende Abweichungen die Summe vergrößern, entgegengesetzte Abweichungen die Summe verkleinern. Die Summe wird also z.B. vom Betrag sehr groß, wenn die

---

[36] Um Verwirrungen vorzubeugen: Asendorpf und Wallbott (1979) schreiben, daß Fleiss und Cohen (1973) die Äquivalenz von κ_W und IC_u nachgewiesen hätten (S. 249f.). Es handelt sich offenbar um einen Druckfehler. Es müßte heißen: κ_W äquivalent zu IC_a' (ihre Notation für IC_{a2}).

Feger (1983, S. 34) zieht unter Berufung auf Krippendorf (1970) eine Parallele zwischen κ_W und „Spearmans Rangkorrelationskoeffizienten". Das ist nicht ganz korrekt: Krippendorf (1970, S. 144) weist nach, daß unter der Annahme, daß jede Rating-Kategorie nur einmal benutzt wird, sowohl π_W als auch κ_W identisch mit diesem Rangkorrelationskoeffizienten sind.

Abweichungen (vor allem die großen) sich immer im Vorzeichen entsprechen oder stets entgegengesetzt sind; sie wird sehr klein, wenn die Abweichungen in ihrer Vorzeichenentsprechung zufällig gemischt sind. Diese Summe wird dann auf n - die Anzahl der Versuchspersonen - relativiert. Der Wert der Kovarianz ist jedoch noch unnormiert; d.h. Kovarianzen verschiedener Variablenpaare sind nicht ohne weiteres vergleichbar. Deshalb wird die Kovarianz auf das Produkt der Standardabweichungen bezogen. Tatsächlich ist der so entstandene Ausdruck ein normierter Index, der zwischen -1 und 1 schwanken kann. Die Korrelation würde in unserem Beispiel dann -1, wenn eine perfekte Vorhersage der Noten aufgrund der Intelligenz möglich wäre: Ein Schüler mit einem Intelligenztestwert, der genau eine Standardabweichung über dem Mittelwert liegt, hätte dann exakt eine Durchschnittsnote, die eine Standardabweichung besser (also niedriger) als die mittlere Note wäre.

Wenn wir die Übereinstimmung von nur zwei Ratern bestimmen wollen, haben wir ebenfalls zwei Variablen vorliegen, deren Zusammenhang wir berechnen können. Die Produkt-Moment-Korrelation ist allerdings das am wenigsten strenge Gütemaß: Da sie für die Zusammenhangsberechnung beliebig skalierter Variablen konstruiert wurde, dürfen sowohl die Mittelwerte der Beobachter als auch ihre Varianz unterschiedlich sein, ohne daß dies Einfluß auf die Korrelation hätte.

In welchem Verhältnis steht die Produkt-Moment-Korrelation zur Intraklassen-Korrelation? Wir wollen dies an einer anderen Problemstellung deutlich machen: Nehmen wir an, wir erforschen die Erblichkeit bestimmter Merkmale anhand von Zwillingsuntersuchungen. Wir wollen wissen, wie gut wir die Testintelligenz eines Zwillings aufgrund der Kenntnis der Ausprägung des Geschwisters vorhersagen können. Auf den ersten Blick eine ähnliche Fragestellung wie im vorherigen Beispiel. Doch was sind hier die Variablen? Nun, offenbar „Intelligenz des ersten Zwillings" und „Intelligenz des zweiten Zwillings". Doch wer ist der erste Zwilling, wer der zweite? Sicherlich wäre dies in jedem Fall eine beliebige Entscheidung (vgl. zu diesem Beispiel Feger, 1983). Bei der Berechnung der Produkt-Moment-Korrelation gehen wir jedoch immer davon aus, daß eindeutig festliegt, welcher Wert zu welcher Variablen gehört, da in die Formel Mittelwerte und Standardabweichungen dieser Variablen eingehen. Bei unserem Zwillingsproblem können wir also die Korrelation in ihrer Höhe dadurch schwanken lassen, daß wir bei einigen Zwillingspaaren die Zuordnung „erster Zwilling" und „zweiter Zwilling" austauschen. Das ist sicherlich eine unerwünschte Eigenschaft.

Mit einem einfachen Trick können wir jedoch zu einem Index gelangen, der nicht mehr von der willkürlichen Zuordnung der Paarlinge zu den Variablen abhängt: Wir verdoppeln unsere Meßwertreihen, indem wir jedes Paar einmal in der Reihenfolge a,b und ein zweites Mal in der Folge b,a eingehen lassen. Wir erhalten auf diese Art zwei Variablen mit identischem Mittelwert und identischer Standardabweichung, so daß das Zuordnungsproblem verschwindet. Wenn wir nun diese neuen Variablen in die Produkt-Moment-Formel einsetzen, erhalten wir (wieder bis auf einen vernachlässigbaren Term) die Intraklassen-Korrelation IC_u (vgl. Krippendorf, 1970).

Asendorpf und Wallbott (1979, S. 245[37]) zeigen überdies an einer alternativen Berechnungsformel für IC_{a1} sehr schön auf, daß nur im Fall gleicher Varianzen der beiden Beobachter $IC_{a1} = r$ ist; in allen anderen Fällen gilt $IC_{a1} < r$.

(4) Ein Anwendungsbeispiel

Wir wollen die Eigenschaften der verschiedenen Korrelationskoeffizienten noch einmal an einem Beispiel deutlich machen, welches weitgehend Bartko (1976, S. 762f.) entnommen ist.

Tab. 7: Beobachtungsergebnisse von vier Beobachtern (mod. nach Bartko, 1976, S. 762)

B1	B2	B3	B4
1	1	5	2
2	2	6	4
3	3	7	6
4	4	8	8
5	5	9	10

Den vier Beobachtern stand eine Ratingskala mit den Stufen 1 bis 10 zur Verfügung. Die Daten zweier Beobachter (B1, B2) weisen perfekte Übereinstimmung auf. Dagegen haben die Beobachter B3 und B4 die Einheiten recht verschieden eingeschätzt: B3 weist einen viel höheren Mittelwert auf als B1, Beobachter 4 hat im Gegensatz zu den anderen fast die volle Breite der Skala ausgenutzt; seine Werte haben also eine viel größere Varianz.

Es sollen im folgenden alle vier Korrelationstypen (IC_u , IC_{a1} , IC_{a2} und r) für alle Korrelationen von B1 mit den übrigen Beobachtern berechnet werden. Tabelle 8 gibt die Ergebnisse wieder.

Tab. 8: Die verschiedenen Korrelationen für das Beispiel (mod. und erweitert nach Bartko, 1976, S. 763)

	B1-B2	B1-B3	B1-B4
IC_u	1.00	-.23	.34
IC_{a2}	1.00	.24	.48
IC_{a1}	1.00	1.00	.80
r	1.00	1.00	1.00

[37] Leider hat sich hier ein kleiner Fehler bei Asendorpf und Wallbott (1979) eingeschlichen: Sie zitieren Winer (1962) für eine alternative Berechnung von IC_u ; zumindest in der (uns vorliegenden) nächsten Auflage (1971) gibt Winer dieselbe Berechnungsformel für IC_{a1} an. Es entsteht so bei Asendorpf und Wallbott fälschlicherweise der Eindruck, die erwähnten Beziehungen zu r würden für IC_u (und nicht IC_{a1}) gelten.

Wir sehen also, daß im Fall perfekter Übereinstimmung (B1-B2) alle Korrelationen den Wert 1 annehmen; dieses Ergebnis wurde natürlich in die Konstruktion der Formeln „eingebaut" und überrascht nicht. Die größte Schwankung weist die Spalte B1-B3 auf. Die Produkt-Moment-Korrelation r und IC_{a1} bleiben von dem starken Mittelwertsunterschied unberührt. Wenn die beiden Beobachter wirklich konsistent darin bleiben, daß B1 nur den unteren Teil, B3 aber nur den oberen Teil der Skala nutzt *und* beide Beobachter das gesamte Geschehen einer Untersuchung einordnen *und* dieser enorme „Anker"-Unterschied berichtet wird, dann ist IC_{a1} durchaus akzeptabel als Reliabilitätswert: Abgesehen von den Mittelwertsunterschieden ist die Übereinstimmung perfekt. IC_{a2} reflektiert stärker diese Unterschiede, hebt sich aber immer noch deutlich von der unjustierten IC_u ab. Je nachdem, wie die Mittelwertsunterschiede zu bewerten sind, würde man die Übereinstimmungsbeurteilung durch IC_u in vielen Fällen als zu streng zurückweisen.

Die Spalte B1-B4 weist Mittelwerts- und Varianzunterschiede auf. Dementsprechend sinken alle Intraklassen-Koeffizienten unter den Maximalwert. Die Unterschiede zwischen den drei Gütemaßen resultieren hier aus der unterschiedlichen Behandlung der Mittelwertsdifferenzen, nicht der Varianzen.

Sachregister

Personenregister