5. Kapitel

Fragebogenkonstruktion

Ulrich Tränkle

1. Einführung

Die Konstruktion eines Fragebogens wird entscheidend davon bestimmt, welche Arten von Informationen (Inhalten) erfaßt, welche Arten von Aussagen gemacht, wofür und auf welcher Grundlage Validität für die Antworten in Anspruch genommen, in welcher Kommunikations-(Befragungs-)form der Fragebogen verwendet werden soll und welche Determinanten des Antwortverhaltens in der Befragungssituation mutmaßlich wirksam sind. Es ist deshalb unerläßlich, in einem einleitenden Kapitel relativ ausführlich auf diese grundlegenden Sachverhalte einzugehen, bevor die mehr technischen Aspekte der Fragenkonstruktion und des Fragebogenaufbaus in der Differenziertheit behandelt werden können, die einem häufig verwendeten Instrument wissenschaftlicher Datenerhebung angemessen ist.

1.1 Versuch einer Systematik von Fragebogen

1.1.1 Einteilungsgesichtspunkte für Fragebogen

Fragebogen lassen sich nach einer Vielzahl von Gesichtspunkten einteilen bzw. charakterisieren. Die wichtigsten sind im folgenden zusammengestellt.

a. Nach dem Grad der Standardisierung lassen sich unterscheiden

- nicht oder schwach standardisierte Fragebogen
  Sie enthalten die Befragungsthemen(-inhalte), aber weder eine genaue Festlegung der Fragen, der Fragenreihenfolge, noch die Antwortmöglichkeiten. Im allgemeinen spricht man hier weniger von Fragebogen als von Interviewerleitfaden, wie sie z.B. in freien Explorationen Verwendung finden.
- Teilstandardisierte Fragebogen
  Sie enthalten in der Regel eine Festlegung von Fragenformulierungen

und Fragenreihenfolge, nicht aber eine Formulierung von Antwortmöglichkeiten.

- Vollstandardisierte Fragebogen
  Bei diesen Fragebogen sind die Fragenformulierungen, die Fragenreihenfolge und die Antwortformulierungen festgelegt.

Mischformen zwischen diesen Typen sind möglich und gebräuchlich.

b. Nach der Kommunikationsform werden meist Fragebogen für schriftliche und solche für mündliche Befragung (Atteslander 1971) oder solche für persönliches Interview und schriftliche Befragung (Scheuch 1973) unterschieden. Eine erschöpfende Klassifizierung nach Kommunikationsformen müßte aber mindestens unterscheiden (Tränkle 1974)

- Fragebogen zur Bearbeitung in Anwesenheit eines Interviewers,
  - im Einzelversuch (ein Befragter),
    - mit mündlicher Vorgabe der Fragen,
      - mit mündlicher Beantwortung (persönliches mündliches Interview),
      - mit schriftlicher Beantwortung,
    - mit schriftlicher Vorgabe der Fragen und schriftlicher Beantwortung (persönliches schriftliches Interview),
  - im Gruppenversuch (mehrere Probanden gleichzeitig),
    - mit mündlicher Vorgabe der Fragen und schriftlicher Beantwortung,
    - mit schriftlicher Vorgabe der Fragen und schriftlicher Beantwortung,
- Fragebogen zur Bearbeitung in Abwesenheit eines Interviewers,
  - im Einzelversuch,
    mit mündlicher Vorgabe der Fragen und mündlicher Beantwortung (z.B. Telefoninterview),
  - im Einzel- oder Gruppenversuch (undefiniert),
    mit schriftlicher Vorgabe der Fragen und schriftlicher Beantwortung (‚postalische' Befragung).

Wiederum sind Mischformen gebräuchlich, außerdem enthält das Klassifikationsschema nicht alle möglichen Kombinationen der beteiligten Gesichtspunkte, sondern nur diejenigen, für die dem Autor Anwendungen bekannt sind.

c. Nach dem angestrebten Gültigkeitsbereich der Aussagen lassen sich unterscheiden

- individual-diagnostische Fragebogen, die Aussagen über Individuen zum Ziel haben, und
- Fragebogen, die Aussagen über Gruppen (Populationen) anstreben und bei denen das Antwortverhalten des einzelnen Individuums als solches nicht interessiert. Solche Fragebogen werden im folgenden der Kürze

halber als ‚demoskopische' oder sozialwissenschaftliche Fragebogen angesprochen.

Hier sind insofern Übergänge möglich, als Aussagen über Populationen auch ausgehend von solchen über Individuen gemacht werden können, in manchen Fällen sogar müssen (vgl. Feger 1974).

d. Nach dem Inhalt der angestrebten Aussagen kann man unterscheiden
- fakten-, wissens- oder kenntnisorientierte Fragebogen mit individualdiagnostischer (wie bestimmte Intelligenztests) oder demoskopischer Intention,
- meinungs- bzw. einstellungsorientierte Fragebogen, ebenfalls entweder mit individual-diagnostischer oder demoskopischer Zielsetzung, und
- persönlichkeitsorientierte diagnostische Fragebogen; hierunter fallen Z.B.
  - Problemfragebogen (adjustment inventories), bei denen es darum geht, das ‚Problemniveau' (die Auffälligkeit) einer Person festzustellen,
  - eigenschafts-(trait-)orientierte Fragebogen, die die Messung des Ausprägungsgrades bestimmter Persönlichkeitsmerkmale zum Ziel haben, und
  - Interessenfragebogen, die einen eng umschriebenen inhaltlichen Aspekt der Persönlichkeit, nämlich Vorlieben bzw. Bevorzugungen von Tätigkeiten, Situationen, Berufen zu erfassen trachten (Mittenecker 1971).

Auch im Hinblick auf den Inhalt des Fragebogens sind natürlich Mischformen möglich.

e. Nach dem Grundkonzept der Fragebogenkonstruktion (dem der Konstruktion zugrundeliegenden Validitätskonzept) kann man unterscheiden
- rationale (Anastasi 1968, Edwards 1970), inhaltsorientierte, ‚sample-approach'-Fragebogen (Cronbach 1970),
- empirische (Anastasi 1968), statistische, ‚sign-approach'-Fragebogen (Cronbach 1970),
- konstrukt-valide, theoriegeleitete (Cronbach 1970) Fragebogen.

Auf diese Grundkonzeptionen von Fragebogen wird im folgenden etwas ausführlicher eingegangen.

1.1.2 Grundkonzeptionen von Fragebogen

Rationale Fragebogenkonstruktionen bestehen in einer Zusammenstellung von Items nach inhaltlichen Gesichtspunkten. Sie wird als repräsentative Stichprobe (Sample) aus einem Universum interessierender Inhalte angesehen. Der Befragte soll die Items verstehen und sie wahrheitsgemäß beantworten. Dem-

entsprechend werden die Antworten ihrer inhaltlichen Bedeutung nach interpretiert und gegebenenfalls zu einem Gesamtscore zusammengefaßt. Dabei werden unter Umständen auch die Interkorrelationen der Items in Betracht gezogen.

Dieser Konstruktionsansatz liegt üblicherweise (vgl. 1.1.3) demoskopischen Fragebogen zugrunde, war aber auch Ausgangspunkt der ersten diagnostischen Fragebogen von Woodworth und Mitarbeitern (Cronbach 1970, Mittenecker 1971).

Innerhalb der rationalen Fragebogenkonstruktion unterscheiden manche Autoren (vgl. Burisch 1976) ein intuitiv-rationales von einem deduktiven Vorgehen bei der Formulierung der Items (Ableitung der Items aus einem - möglicherweise spekulativen - Persönlichkeitsmodell). Sie heben davon einen sogenannten internalen (z.B. Hornick et al. 1977) bzw. induktiven (z.B. Burisch 1976) Ansatz der Konstruktion von Fragebogen auf der Basis der Ergebnisse von Faktorenanalysen der Iteminterkorrelationen ab. Dabei handelt es sich u.E. jedoch nicht um eine eigenständige Fragebogenkonzeption, sondern um eine Technik in der Regel inhaltlicher Validierung. Bekanntlich hängt das Ergebnis einer Faktorenanalyse entscheidend von den einbezogenen Variablen (Items) ab (für den Fall von Persönlichkeitsfragebogen und von zugehörigen Persönlichkeitstheorien hat Coan 1964 dies auch empirisch demonstriert, vgl. auch Scheier & Cattell 1965). über die Einbeziehung eines Items in die Faktorenanalyse wird aber nach inhaltlichen Gesichtspunkten oder nach seiner kriteriumsbezogenen Validität entschieden. Im letztgenannten Fall ist Ziel der Faktorenanalyse ebenfalls die nähere Untersuchung des Fragebogeninhalts.

Prinzipielles Problem des rationalen Konstruktionsansatzes ist, daß er mit durchschaubaren Items arbeitet und arbeiten muß und daß dadurch die Antworten leicht verfalschbar sind (vgl. 1.2.1).

Dies war historisch gesehen auch der Anlaß für die Entwicklung des *empirischen* (bzw. externalen, vgl. Hornick et al. 1977) Konstruktionsansatzes. Wird er in reiner Form verwirklicht, so orientiert sich die Zusammenstellung der Items ausschließlich an ihren Korrelationen zu externen Kriterien. Die Antworten werden als verbales Verhalten betrachtet, das als Zeichen (sign) bzw. Indikator für einen Sachverhalt anzusehen ist, d.h. die Bedeutung einer Antwort ergibt sich allein aus der Korrelation zu Außenkriterien. Auf dieser Grundlage wurden z.B. der MMPI und der Interessenfragebogen von Strong entwickelt. Da die Iteminhalte prinzipiell unerheblich sind, ist es hier möglich, nicht durchschaubare Items zu verwenden und dadurch die Möglichkeit von Verfälschungen erheblich zu reduzieren. Inhaltlich valide Items sind aus diesem Grund für einen streng empirischen Fragebogen geradezu unerwünscht. Die Zusammenfassung von Antworten zu Gesamtscores erfolgt gegebenenfalls nach Maßgabe gemeinsamer Korrelationen mit Außenkriterien. Hauptpro-

bleme dieses Ansatzes sind einerseits die fehlende oder geringe face-validity der Items, wodurch die Motivation der Probanden beeinträchtigt werden kann, andererseits die Risiken, die in einer inhaltsblinden Suche nach empirisch validen Items stecken: Recht häufig wird man dadurch Items mit überhöhten Validitäten in den Fragebogenentwurf aufnehmen und bei Kreuzvalidierungen erhebliche Validitätsrückgänge feststellen. Gelegentlich wird sogar die Auffassung vertreten, „. . . daß substantielle und stabile Zusammenhänge immer nur für Merkmale mit plausibler inhaltlicher Beziehung zu finden . . .“ seien (Burisch 1976, 28). Jedenfalls sind von empirischen Konstruktionen auch nur mittlere Validitäten erreicht worden (Cronbach 1970). Versuche, zur Verbesserung der Akzeptabilität den Items eine von der wirklichen Validität verschiedene face-validity zu verleihen, sind ihrerseits problematisch, da sie evtl. Verfälschungen nach Maßgabe der face-validity begünstigen.

Für praktische diagnostische Anwendungen ist es in jedem Falle unerläßlich, daß ein Fragebogen auch empirisch validiert und nicht, wie bei ausschließlich rationalen Konstruktionen, nur eine „. . . spekulativ halbwegs sinnvoll erscheinende Zusammenstellung . . .“ von Items (Wottawa 1980, 211) ist. Umgekehrt kann kaum ein Testanwender der Versuchung widerstehen, entgegen den Intentionen des Konstrukteurs einen rein empirischen Fragebogen auch inhaltlich zu interpretieren (Cronbach 1970), so daß Überlegungen zur inhaltlichen Validität erforderlich werden. In der Praxis reduziert sich der Unterschied der Grundkonzeptionen häufig auf einen solchen beim ersten Schritt der Item-Selektion (Trennschärfe vs. Kriteriumskorrelation). Darüber hinaus ließ sich auch bei strenger Verwirklichung eine generelle Überlegenheit des einen oder anderen Ansatzes nicht nachweisen (Hase & Goldberg 1967, Burisch 1976, Hornick et al. 1977).

Für *theoriegeleitete* bzw. *konstrukt-valide* Fragebogen (Cronbach & Meehl 1955) ist zum Zwecke der Validierung der Nachweis zu führen, daß es sich bei dem, was sie messen, um ein im Rahmen einer Theorie definiertes Konstrukt handelt. Dieser Nachweis erfolgt vor allem durch Ableitung von Beziehungen zu weiteren Konstrukten aus der Theorie und Überprüfung der Verträglichkeit dessen, was der Fragebogen erfaßt, mit diesen Vorhersagen. Ansätze zu derartigen Fragebogen sind zunächst vor allem von Eysenck (vgl. Eysenck 1953) vorgelegt worden, dessen Intention stets die Messung von Konstrukten war, die in eine umfassende Theorie kortikaler Prozesse eingebaut sind. Demgegenüber kann man allein aufgrund von Versuchen einer Abstraktion von traits aus Fragebogendaten z.B. mittels Faktorenanalyse (vgl. vor allem Guilford 1965) wohl noch nicht von theoriegeleiteter Fragebogenkonstruktion oder von einer Konstruktvalidierung von Fragebogen sprechen (Cronbach 1970). Die faktorielle Reinheit von Fragebogenitems hat - entgegen Holm (1974 a, b) - nichts mit ihrer Theorieorientiertheit zu tun. Außerdem gibt es kaum sachliche (allenfalls technische) Gründe, sie zu fordern (Cattell 1974),

zumal sie ja nicht an sich, sondern immer nur innerhalb einer gegebenen Variablen-(Item-)stichprobe existiert. Von erheblichem Nutzen kann die Faktorenanalyse sein, wenn ein inhaltlich (rational) konstruierter Fragebogen z.B. durch Einbeziehen von Markierungsvariablen daraufhin überprüft werden soll, ob die angestrebten Inhalte tatsächlich enthalten sind (so schon Eysenck 1953). Außerdem läßt sich mit ihrer Hilfe bei einem empirischen Fragebogen, besonders wenn die Kriterien mit in die Analyse einbezogen werden, die zunächst nur statistische Beziehung zwischen Antwort und Kriterien auch in ihrer inhaltlichen Bedeutung erhellen, d.h. ein zunächst nur empirisches in ein auch inhaltliches Validitätskonzept überführen.

1.1.3 Hauptanwendungsgebiete für Fragebogen

Diagnostische Fragebogen sind Tests, insofern orientiert sich ihre psychometrische Konstruktion an einem Test- bzw. Meßmodell. Bezüglich der damit zusammenhängenden Fragen muß auf die einschlägige Literatur, z.B. Gulliksen (1950), Magnussen (1966), Lienert (1969), Fischer (1974), verwiesen werden. Faßt man einen Fragebogen ganz allgemein als eine Zusammenstellung von Fragen auf, so enthalten die meisten Tests auch Fragebogen. Für Fragen in Leistungstests ist charakteristisch, daß es für sie eine objektiv richtige Antwort gibt, so daß besondere Überlegungen zum Problem des Ratens erforderlich werden (vgl. Wottawa 1980). Diagnostische Fragebogen im engeren Sinne sind z.B. Persönlichkeits- und Interessenfragebogen, für ihre Items gibt es höchstens subjektiv richtige Antworten. Die folgenden Ausführungen beschränken sich in der Regel auf Fragebogen dieses Typs. Fast immer sind Persönlichkeits- und Interessenfragebogen vollstandardisierte Verfahren, mit schriftlicher Vorgabe von Fragen und schriftlicher Beantwortung in Anwesenheit eines Untersuchers, die Durchführung erfolgt teils im Einzelversuch, teils in Gruppen.

Demoskopische Fragebogen können alle möglichen Standardisierungsgrade aufweisen. Die Befragungen werden meist als persönliche mündliche Interviews im Einzelversuch oder ‚postalische' Befragungen, seltener als persönliche schriftliche Interviews durchgeführt. Die Fragebogen haben Fakten, Wissen, Kenntnisse, Meinungen oder Einstellungen zum Inhalt und sind vom Validitätskonzept her fast ausschließlich rationale bzw. inhaltsorientierte Konstruktionen. Überlegungen zu empirischen (kriteriumsorientierten) Validierungen werden vor allem im Zusammenhang mit Meinungs- und Einstellungsfragen und ihrer Indikatorfunktion angestellt (Friedrich 1971), soweit Fakten- und Wissensbereiche thematisch sind, interessieren eher Verfälschungen bzw. Fehler (Lansing et al. 1961) und ihre Hintergründe, bzw. Fragen der Reproduktion oder des Wiedererkennens von Gedächtnisinhalten (Cannell et al. 1977).

In gewisser Hinsicht eine Zwischenstellung zwischen diagnostischen (Persönlichkeits- und Interessen-) und demoskopischen Fragebogen besitzen die sogenannten Einstellungs-(Attitüden-)skalen, die eingehend etwa bei Edwards (1957), Scheuch (1962), Süllwold (1969), Scheuch & Zehnpfennig (1974) behandelt werden. Wie demoskopische Fragebogen werden sie fast ausnahmslos mit dem Endziel von Aussagen über Gruppen und nicht zu individual-diagnostischen Zwecken eingesetzt, doch sind sie - von Eigentümlichkeiten der Itemselektion abgesehen - formal mit diagnostischen Fragebogen identisch. Die Konstruktion von Einstellungsskalen geht stets von Iteminhalten aus, erfordert aber mindestens bei Skalen des Thurstone-Typs auch eine empirische Validierung. Während für die Lickert-Skalen die Trennschärfe unter Zugrundelegung von Gesamtscore-Extremgruppen (also ein inhaltliches Konzept) das Selektionskriterium für Items darstellt, werden bei den Thurstone-Skalen die Items unter Verwendung einer ‚Eichstichprobe' hinsichtlich der Extremität der durch sie zum Ausdruck gebrachten Einstellung skaliert. Skalenwert des Individuums ist im erstgenannten Falle die Summe der graduell abgestuften Zustimmungen zu den Items, im letztgenannten Falle (Modifikationen dieser Vorgehensweise einmal unberücksichtigt gelassen) der Skalenwert des Items, das der Proband am ehesten für zutreffend hält. Abgesehen von Einwänden, die sich auf die ungeprüft unterstellte Eindimensionalität der gemessenen Sachverhalte beziehen und denen Guttman mit der Skalogrammanalyse zu begegnen versuchte (vgl. Edwards 1957), wäre bei der Verwendung von Zustimmungsgraden in Lickert-Skalen die Dimensionalität der Antworten, z.B. durch Zugrundelegung eines mehrkategoriellen probabilistischen Meßmodells zu berücksichtigen (Wottawa 1980); jedenfalls läßt sich die mehr oder weniger willkürliche Verwendung von (gleichabständigen) Gewichten für die Zustimmungsgrade bei der Bildung eines Gesamtscores kaum rechtfertigen. Für Thurstone-Skalen ist ungeklärt, wie sich Personen und an einer Eichstichprobe skalierte Items in einem gemeinsamen psychologischen Raum darstellen lassen könnten. Eine kritische Auseinandersetzung mit den Ansätzen von Thurstone, Lickert und Guttman sowie alternative Vorgehensweisen finden sich z.B. bei Feger (1974) und Lantermann & Gehlen (1977).

Trotz der vorstehend beschriebenen Akzentuierungen gibt es hinsichtlich zahlreicher Probleme keine prinzipiellen Unterschiede zwischen diagnostischen Fragebogen, Einstellungsskalen und demoskopischen Fragebogen, so daß die nachfolgenden Ausführungen sich nur ausnahmsweise explizit auf bestimmte Anwendungssituationen beziehen.

1.2 Ansätze zu einer Theorie des Beantwortungsprozesses

1.2.1 Determinanten des Antwortverhaltens

Als erster Schritt auf dem Wege zu einer Theorie des Beantwortungsprozesses bietet sich die Analyse dessen an, was bei der Entstehung einer Antwort in der Vp vor sich geht. Entsprechende Untersuchungen sind für den Fall von Persönlichkeitsfragebogenitems mehrfach durchgeführt worden (vgl. etwa die Nachweise bei Cronbach 1970, Schneider-Düker & Schneider 1977, Kalinowsky-Czech 1979), ihre Ergebnisse dürften sich prinzipiell aber auch auf demoskopische Fragen übertragen lassen. Turner & Fiske (1968) und in Fortführung dieses Ansatzes Kuncel (1973, vgl. auch Fiske 1978) und ebenso Nowakowska (1971) verwendeten einen ‚Mets-Fragebogen' zur nachträglichen Erfassung der Beantwortungsprozesse, Schneider-Düker & Schneider (1977) und Kalinowsky-Czech (1979) bedienten sich der Methode des ‚Lauten Denkens', die letztgenannte Autorin ließ ihre Vpn außerdem frei zu den Items assoziieren. Rogers (1974 a, b) variierte bestehende Charakteristika der Items experimentell und zog ausgehend von den Veränderungen der Beantwortungszeiten (Reaktionszeiten) Schlüsse auf den Beantwortungsprozeß. Cliff et al. (1973) und Cliff (1977) versuchten, Beziehungen zwischen der Beantwortung von Items und ihrer Bedeutung (ausgehend von der MDS ihrer Ähnlichkeitsstruktur) herzustellen. übereinstimmend zeigten diese Untersuchungen, daß einerseits das Verständnis ein- und derselben Frage von Vp zu Vp beträchtlich variiert und andererseits die Beantwortungsprozesse ein- und derselben Vp itemspezifisch recht unterschiedlich ablaufen (vgl. auch Crutchfield & Gordon 1947). Turner & Fiske (1968) etwa klassifizierten ausgehend von MMPI-Items nur etwa 50% der von Vpn beschriebenen Beantwortungsprozesse als adäquat im Sinne der Intention des Fragebogens bzw. seiner Autoren. Dazu dürfte u.a. beitragen, daß die Vorstellungen, die Vpn mit den in Fragebogenitems häufig anzutreffenden unbestimmten Zahlen- und Häufigkeitsangaben (‚einige', ‚manchmal') verbinden, sehr unterschiedlich sind (Simpson 1944, Strahan & Gerbasi 1973, Schriesheim & Schriesheim 1974, Rohrmann 1978, Bradburn & Sudman 1979).

Weitere Erkenntnisse betreffend die Determination des Antwortverhaltens kommen sodann von experimentellen Untersuchungen zur Verfälschbarkeit (faking) von Antworten auf Fragebogenitems, z.B. in vorgestellten Situationen, und von Untersuchungen zur Wirksamkeit von Antworttendenzen bei der Bearbeitung von Fragebogen in Ernstsituationen. Auf die kaum noch überschaubare Fülle der dazu vorliegenden empirischen Befunde kann an dieser Stelle nicht näher eingegangen werden, zusammenfassend orientieren darüber z.B. Block (1965), Berg (1967), Anastasi (1968), Cronbach (1970), Edwards (1970). Im deutschen Sprachraum sind insbesondere Untersuchungen von Cohen & Carl (1964), Carl (1968), Fürntratt (1969), Tholey (1976), Ham-

pel & Klinkhammer (1978), Jannssen (1978), Häcker et al. (1979) und mehrere Arbeiten von Hoeth und Mitarbeitern (zusammenfassend Hoeth 1980) zu nennen.

Mit Möglichkeiten der Vermeidung bzw. Erfassung und Elimination der Einflüsse von Antworttendenzen (Reaktionseinstellungen, response sets) und den Schwierigkeiten ihrer Realisierung bei der Fragebogenkonstruktion setzen sich u.a. Ehlers (1973), Janke (1973) und Keil (1973) auseinander.

Als *Antworttendenzen* könnte man allgemein diejenigen systematischen Anteile im Antwortverhalten der Vpn bezeichnen, die nicht auf den jeweiligen (subjektiv) wahren Sachverhalt, sondern auf die Form der Frage bzw. der Befragung insgesamt zurückzuführen sind. Untersucht wurden derartige Antworttendenzen vor allem im Zusammenhang mit diagnostischen Fragebogen, doch lassen sie sich nach Hoeth (1980) auch für demoskopische (sozialwissenschaftliche) Fragebogen leicht aufzeigen.

Bei den *frageninhaltsorientierten* Antworttendenzen wie Simulation, Dissimulation und der besonders intensiv untersuchten Tendenz zu sozial erwünschten Antworten (SD-Tendenz) werden die Antworten auf Fragen im Hinblick auf ganz bestimmte Zwecke (z.B. einen ‚guten Eindruck' zu machen) verfälscht.

Demgegenüber erfolgt bei den *antwortinhaltsorientierten* Antworttendenzen (den response sets im engeren Sinne) eine Bevorzugung von Antworten ganz bestimmten Inhalts ohne Rücksicht auf die Inhalte der Fragen. Am meisten Aufmerksamkeit auch im Hinblick auf Beziehungen zu bestimmten Persönlichkeitsmerkmalen hat dabei die Bejahungs- oder Zustimmungstendenz (acquiescence) gefunden, außerdem wurden Verneinungstendenzen, Mittententendenzen, Extremtendenzen und Variationstendenzen aufgezeigt.

Schließlich gibt es *nicht-inhaltsorientierte* Antworttendenzen, dazu gehören Positionseffekt und formale Antwortstereotypien (Antwortmuster).

Antworttendenzen und Unterschiede im Verständnis von Fragebogenitems stellen im Rahmen einer rationalen (inhaltsorientierten) Fragebogenkonstruktion (sample-approach) sicherlich ein schwerwiegendes Problem dar (Eysenck 1953), das durch geeignete Formulierung und Zusammenstellung von Items bzw. durch Verwendung spezieller Fragen oder Fragentypen (forced-choice-items) bestenfalls gemildert, nicht jedoch überwunden werden kann (Ehlers 1973, Keil 1973).

Ihrer Natur nach sind Verständnisunterschiede und Antworttendenzen nicht als intraindividuell unkorrelierte Zufallsfehler mit einem Erwartungswert von Null anzusehen, so daß entsprechend dem Reliabilitätskonzept der klassischen Testtheorie erwartet werden dürfte, daß sie sich mit steigender Zahl homoge-

ner Items zur Erfassung des jeweiligen Sachverhaltes tendenziell aufheben. Ein probabilistisches Meßmodell bietet zwar prinzipiell den Vorteil, daß sich hier im individual-diagnostischen Anwendungsfall von den Eigentümlichkeiten der verwendeten Items (Item-Parametern) befreite spezifisch objektive Personenkennwerte (Personen-Parameter) bestimmen lassen, im Falle von gruppendeskriptiven Zielsetzungen Kennwerte @em-Parameter), die von den verwendeten Vpn (Personen-Parametern) bereinigt sind (Sixtl 1972, Andersen 1973, Fischer 1974), doch würden Antworttendenzen und Verständnisunterschiede der Items bei verschiedenen Vpn gravierend gegen die Modellannahme der Unabhängigkeit der Item- von den Personen-Parametern verstoßen. Innerhalb eines inhaltsorientierten (rationalen) Ansatzes zeichnen sich also keine Möglichkeiten ab, das Problem der Antworttendenzen und Verständnisunterschiede für Items zu lösen (im Zusammenhang mit Überlegungen zur Fragenformulierung wird auf diesen Punkt noch einmal zurückzukommen sein, vgl. 3.).

Im Rahmen des rein empirischen Validitätskonzeptes für Fragebogen (sign-approach) stellen die Unterschiede im Verständnis der Fragen kein prinzipielles Problem dar. Vielmehr ist durchaus denkbar, daß sie eine wesentliche Grundlage der empirischen Validität sind und daß diese sinken würde, wollte man die Verständnisunterschiede zwischen Vpn reduzieren (Cronbach 1970, Mittenecker 1971).

So berichten Strahan & Gerbasi (1973) tatsächlich deutliche Korrelationen zwischen Interpretationen von Items und Persönlichkeitsdimensionen (die sie freilich anders interpretieren). Die Minimalbedingung für die Brauchbarkeit von Antworten in Fragebogen ist nur, daß sie „. . . irgendwie psychologisch bedeutsame Reaktionen . . .“ sein müssen (Mittenecker 1971, 480).

Solche Reaktionen liegen nur dann nicht vor, wenn die Vp auf Items eines Fragebogens ohne Bezug zum Inhalt von Frage- oder Antwortmöglichkeiten reagiert, d.h. z.B. zufällig oder willkürlich antwortet. Verfälschungstendenzen (wie soziale Erwünschtheit) bzw. responsesets (wie acquiescence) können dagegen als solche psychologisch bedeutungsvoll sein und schließen die Verwendbarkeit von Reaktionen nicht a priori aus (Cattell 1974). Unter dem Begriff ‚response style‘ hat man insbesondere die Zustimmungstendenz (acquiescence) selbst zum Persönlichkeitsmerkmal erhoben, also das, was ursprünglich eine Verfälschung der Antworten zu sein schien, zu einer inhaltlich interessierenden Fragebogenvariablen gemacht (vgl. als kritische Übersicht Rorer 1965).

1.2.2 Antwortgenese

Die prinzipiellen Möglichkeiten für das Zustandekommen der Antwort auf ein Fragebogenitem sind in Abb. 1 in Form eines Flußdiagramms dargestellt, in

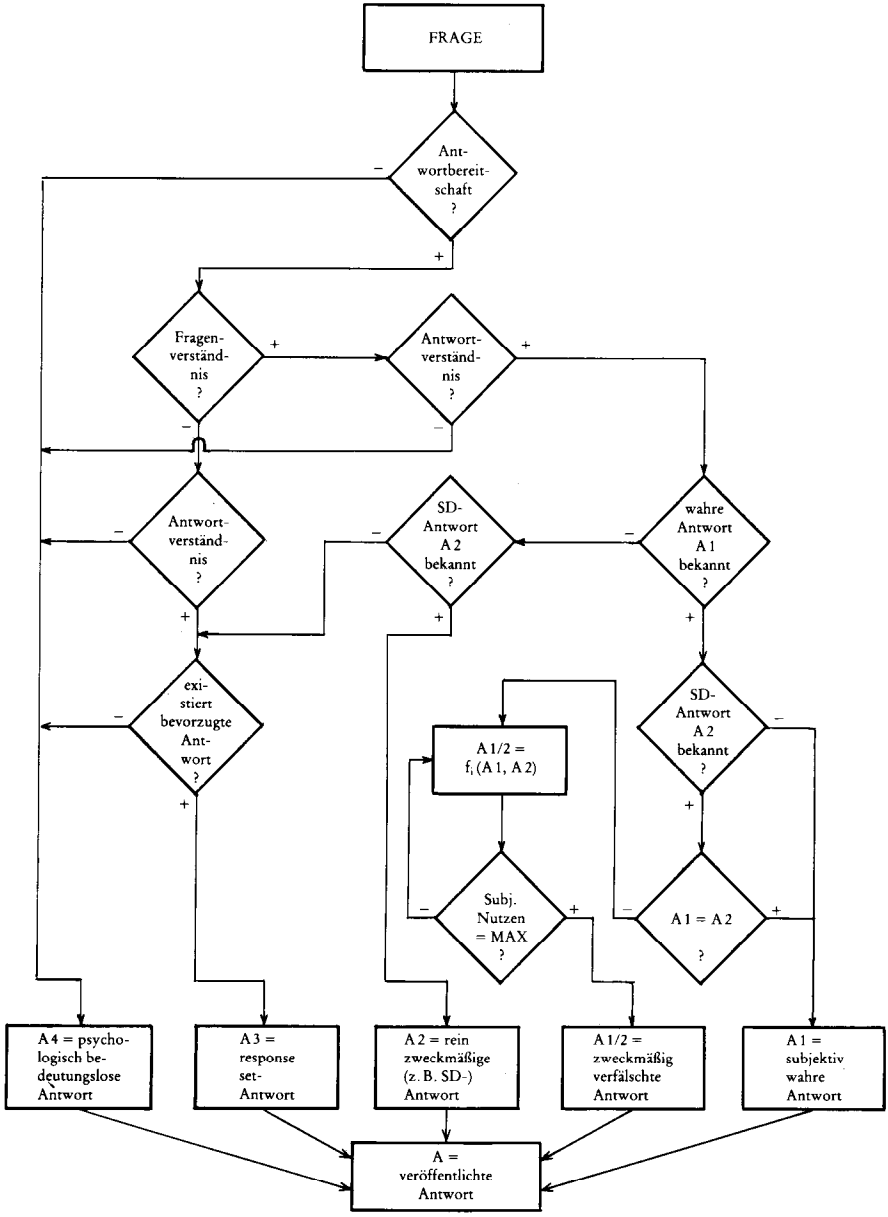

Abb. **1:** Modell der Antwortgenese

das Vorstellungen von Getzels (1954), Damarin (1970), Nowakowska (1971), Schneider (1972) eingegangen sind. Dabei handelt es sich um ein Modell der Antwortgenese, nicht etwa um eine phänomenologische Beschreibung, d.h. es wird nicht angenommen, daß die Antwortgenese sich im Bewußtsein der Vp so darstellt, wie das Modell sie abbildet.

Eine Antwort in dem hier zugrundegelegten Sinne ist auch die Nichtbeantwortung (das Ausbleiben einer inhaltlichen Antwort) bzw. die Enthaltung (Wahl einer Neutralkategorie).

Bei der Entstehung einer Antwort gibt es zunächst die Möglichkeit, daß die Vp, z.B. weil sie nicht zu angemessener Mitarbeit bereit ist, vom Inhalt des Items bzw. von den Inhalten der Antwortmöglichkeiten gänzlich unbeeinflußt reagiert. Dies führt zu einer mindestens im Zusammenhang mit dem Item psychologisch nicht bedeutungsvollen, ‚zufälligen' oder willkürlich systematischen Antwort (A4). Derartige Antworten können die Grundlage sogenannter Positionseffekte sein. Natürlich besteht die Möglichkeit, aus dem Beantwortungsprozeß ‚auszusteigen' und nach A4 zu verzweigen, auf allen nachgeordneten Stufen. Im Interesse der Übersichtlichkeit haben wir dies im Flußdiagramm nicht eigens vorgesehen (wollte man es tun, könnte man nach jedem Schritt eine weitere Abfrage ‚Motivation noch ausreichend' o.ä. einbauen).

Sodann kann eine Antwort zustande kommen durch die Verarbeitung von Frage und explizit oder implizit vorgegebenen Antwortmöglichkeiten unter Heranziehung von Gedächtnisinhalten (Wissen, Normen). Dies setzt voraus, daß ein subjektives (d.h. von der Vp als solches erlebtes) Verständnis von Frage und/oder Antwortmöglichkeiten erzielt worden ist. Sieht man von A4, der psychologisch nicht bedeutungsvollen Antwort, ab, so kann die Vp ihre Antwort nach dem Kriterium der subjektiven Richtigkeit (Wahrheit) als richtige Antwort (A1), in Verfolgung eines bestimmten Zweckes z.B. entsprechend der (subjektiven) sozialen Erwünschtheit (A2) oder nach Maßgabe einer bei ihr vorherrschenden Reaktionstendenz (response set), also ohne Rücksicht auf den Fragen-, aber mit Bezug auf den Antwortinhalt auswählen (A3). Alle diese Möglichkeiten entsprechen der Minimalbedingung der psychologischen Bedeutsamkeit (Mittenecker 1971) für Antworten.

Ist eine Frage für eine Vp bedeutsam, d.h. trifft sie auf ihre eigene Situation zu und verbindet sie mit der Frage kognitive Inhalte oder positive bzw. negative Assoziationen, so ist sie auch in der Lage, zu bestimmen, welche Antwort die subjektiv richtige (A1) ist (Nowakowska 1971). Diese Antwort wird sie unmittelbar als endgültige Antwort (A) verlautbaren, wenn sie keine Kriterien (spezifische Erfahrung mit Antwortkonsequenzen, verinnerlichte gesellschaftliche Normen) für die Zweckmäßigkeit (z.B. für die soziale Erwünschtheit) einer Antwort besitzt. Besitzt sie solche Kriterien, so wird sie zusätzlich eine ‚zweckmäßige' (d.h. meist sozial erwünschte) Antwort (A2) entwerfen, die

mit der subjektiv richtigen Antwort übereinstimmen (A1 = A2) oder von dieser abweichen kann. Daneben gibt es auch die Möglichkeit, daß die Vp eine subjektiv richtige Antwort (z.B. mangels Betroffensein) nicht kennt, jedoch weiß, welche Antwort zweckmäßig, z.B. sozial erwünscht ist. In diesem Fall wird sie eine sozial erwünschte Antwort (A2) wählen (die natürlich auch in einer Nichtbeantwortung oder Enthaltung bestehen kann). Einerseits unterscheiden sich Probanden in der Neigung, relativ unabhängig von der konkreten Frage sozial erwünscht zu antworten, andererseits unterscheiden sich aber auch Fragen in der Neigung, relativ unabhängig von den Probanden sozial erwünscht beantwortet zu werden. Holm (1974 b) differenziert dementsprechend im Rahmen seiner faktorenanalytischen Theorie der Frage bzw. Fragenbatterie zwischen einer ‚allgemeinen sozialen Erwünschtheit' und einer (fragen-) ‚spezifischen sozialen Erwünschtheit'. Letztere wird häufig auch als Suggestivwirkung einer Frage bezeichnet.

Stehen sich zwei verschiedene Antwortmöglichkeiten, eine subjektiv richtige (A1) und eine zweckmäßige, z.B. sozial erwünschte (A2) gegenüber, so trifft die Vp nach den Ergebnissen von Nowakowska (1971), die sie mittels Faktorenanalyse von Zusammenhängen in den Beschreibungen der Beantwortungsprozesse bei 28 Items aus dem 16 PF von Cattell gewann, die Entscheidung in Abhängigkeit von der subjektiven Nützlichkeit (N) der Antwort. Die subjektive Nutzenfunktion wird dabei durch die erwarteten materiellen und gesellschaftlichen Konsequenzen der Antwort einerseits und die Konsequenzen für das Selbstwertgefühl (z.B. bei Abweichung von der Wahrheit) andererseits bestimmt. Die Optimierung unter dem Kriterium des subjektiven Nutzens kann zur Wahl von A1 oder A2 führen oder eine kombinierte Antwort (A1/2) = f (A1, A2) erzeugen, also eine durch ‘Zweckmäßigkeitsüberlegungen verfälschte richtige bzw. eine in dem Bestreben nach Wahrheit veränderte Zweckantwort. Diese Veränderung der Antwortentwürfe kommt für die Vp natürlich nur in Betracht, wenn A1 und A2, d.h. die wahre und die zweckmäßige Antwort, divergieren.

Fragen, bei denen die Vp erhebliche gesellschaftliche Konsequenzen im Falle einer sozial nicht erwünschten Beantwortung erwartet, dürften den größten Teil dessen abdecken, was in der Literatur unter den Bezeichnungen ‚schwierige', ‚heikle', ‚unangenehme' Fragen abgehandelt wird. Für solche Fragen gilt als typisch, daß sie relativ hohe Nichtbeantwortungsquoten aufweisen. Vermutlich handelt es sich hier um Fälle, in denen der Konflikt zwischen wahrer und zweckmäßiger (sozial erwünschter) Antwort durch Nichtbeantwortung gelöst wird, d.h. in denen die Nichtbeantwortung den höchsten subjektiven Nutzen verspricht. In diese Richtung deuten die Ergebnisse von Koolwijk (1968, 1969), denenzufolge Fragen nicht an sich unangenehm sind, sondern in Abhängigkeit davon, wie die wahre Antwort der Vp ausfallen müßte, in sehr verschiedenem Ausmaß als unangenehm empfunden werden.

Gelingt es der Vp, z.B. aus Mangel an Aktualität (subjektiver Bedeutsamkeit) der Frage und an Vorstellungen über gesellschaftliche Konsequenzen der Antworten nicht, eine am Frageninhalt orientierte Antwort (A1, A2, A1/2) zu entwickeln, so tritt eine vom Frageninhalt unabhängige und - soweit solche vorhanden sind - durch Antwortbevorzugungstendenzen bestimmte Antwort auf. Derartige response sets kommen demnach also nur ins Spiel, wenn eine Vp weder Kriterien für eine richtige, noch solche für eine zweckmäßige Antwort hat. Holm (1974a) spricht in diesem Falle von Fragen ohne klare Zieldimension. Die Tendenz zu sozial erwünschten Antworten hat demgegenüber andere Qualität. Sie tritt nicht nur auf, wenn die Vp eine wahrheitsgemäße Antwort nicht zur Verfügung hat, sondern konkurriert mit dieser.

Sollte bei Unmöglichkeit frageninhaltsorientierten Antwortens (A1, A2, A1/2) die Vp auch nicht über Antwortbevorzugungstendenzen verfügen, die A3 determinieren könnten, tritt eine willkürliche Antwort (A4) auf, bei der es sich natürlich auch um eine Auslassung handeln kann.

Dem Untersucher steht im Normalfall nur die endgültige Antwort (A) der Vp zur Verfügung. Durch Konstruktion des Fragebogens sollte er soweit als möglich sicherstellen, daß bei zugrundegelegtem inhaltlichen Validitätskonzept (sample approach) A = A1, bei empirischem Validitätskonzept (sign approach) im Sinne obiger Minimalbedingungen $A \neq A4$ ist.

Für einen Teil der Fragen in demoskopischen Interviews existieren objektiv richtige Antworten. In diesen Fällen ist es gerechtfertigt, von ‚Beantwortungsfehlern' (response-errors) zu sprechen. Unter (im Vergleich zu obigem Modell der Antwortgenese) stärkerer Betonung der Ursachen für die Fehler unterscheiden Lansing et al. (1961)

- motivationsbedingte Beantwortungsfehler, die vorliegen, wenn der Proband nicht motiviert ist, die richtige Antwort zu geben, selbst wenn er das könnte (vgl. Cattell 1974),
- kommunikationsbedingte Beantwortungsfehler, die vorliegen, wenn
  - der Proband nicht versteht, welche Information von ihm erwartet wird, d.h. der Fragesteller sich nicht verständlich gemacht hat,
  - der Untersucher die vom Probanden übermittelte Information nicht versteht, d.h. der Proband sich nicht verständlich gemacht hat,
- Unwissenheitsfehler, die vorliegen, wenn dem Probanden die erbetene Information nicht zur Verfügung steht.

Diesen *Beantwortungsfehlern* (response-errors) sind noch die *Antwortverweigerungsfehler* (errors-of-non-response) an die Seite zu stellen, also Fehler, die die Verallgemeinerungsfähigkeit von Befragungsergebnissen betreffen und die ganz besonders im Zusammenhang mit unpersönlichen (z.B. postalischen) Befragungen ein gravierendes Problem sind. Als dritter Fehlertypus neben

Beantwortungs- und Antwortverweigerungsfehlern sind *Stichprobenfehler* in Rechnung zu stellen (Lansing et al. 1961), die - da es sich nicht um spezifisch mit Befragungen bzw. Fragebogen verbundene Fehler handelt - im Rahmen dieser Abhandlung unberücksichtigt bleiben sollen.

1.2.3 Die Frage als Suchbegriff

Aus der eher makroskopischen Betrachtungsweise des Modells für die Antwortgenese soll ein Aspekt wegen seiner Wichtigkeit herausgegriffen und etwas näher beleuchtet werden: die Auffindung der für die subjektiv richtige Antwort erforderlichen Gedächtnisinhalte. Insbesondere bei einer Ausrichtung von Fragen auf inhaltliche Validität wird man sich nicht mit der Abfrage ‚wahre Antwort bekannt' (siehe Abb. 1) zufrieden geben, man wird sich vielmehr überlegen müssen, wie die Vp versucht, die wahre Antwort aufzufinden, und wie der Fragebogenkonstrukteur ihr dabei helfen kann. Cannell et al. (1977) haben im Zusammenhang mit der Diskussion von Faktenfragen, also von Fragen, für die inhaltliche Validität angestrebt wird, deutlich gemacht, daß für solche Fragen einerseits die Speicherung der Information im Gedächtnis der Vp und die zu ihrer Auffindung erforderlichen Suchprozesse, andererseits die Vorstellungen des Fragebogenkonstrukteurs von den zu erfragenden Fakten näher untersucht und in einer formulierten Frage zur Deckung gebracht werden müssen. Abb. 2 (in Anlehnung an Cannell et al. 1977, 53) gibt die dabei mindestens zu beachtenden Schritte wieder.

Eine Gedächtnisspur nimmt nicht vom wahren Sachverhalt, sondern vom Phänomen, d.h. von dem im Erleben des Probanden realisierten Sachverhalt ihren Ausgang, wohingegen der Untersucher seine Vorstellung des Sachverhaltes zugrundelegt. Da der Untersucher zunächst nicht weiß, wie ein Proband einen bestimmten Sachverhalt erlebt und in welcher Kodierung er ihn abgespeichert hat, ist es - um die Wahrscheinlichkeit der Auffindung zu erhöhen - erforderlich, möglichst viele alternative Vorstellungen von dem Sachverhalt zu entwickeln. Um ein Beispiel von Cannell et al. (1977) zu verwenden: Was sich für den Untersucher als ‚zahnärztliche Behandlung' darstellt, kann im Erleben des Befragten in erster Linie eine ‚besonders schmerzhafte Erfahrung' oder eine ‚hohe finanzielle Belastung' gewesen sein. Fragt man nicht nur nach ‚zahnärztlicher Behandlung', sondern verwendet auch die beiden alternativen Vorstellungen vom Sachverhalt, so erhöht man die Wahrscheinlichkeit, daß der Proband die gesuchte Information in seinem Speicher auffindet. Aber auch *ein* bestimmter erlebter Sachverhalt kann in sehr verschiedene *Bezugssysteme* eingebettet sein. Bei der Operationalisierung des Sachverhaltes als Untersuchungsvariable kommt es darauf an, möglichst viele der in Frage kommenden Bezugssysteme zu berücksichtigen. Fragt man nach Krankheiten des Probanden nicht nur im Kontext eines Klassifikationsschemas für Krankheiten, son-

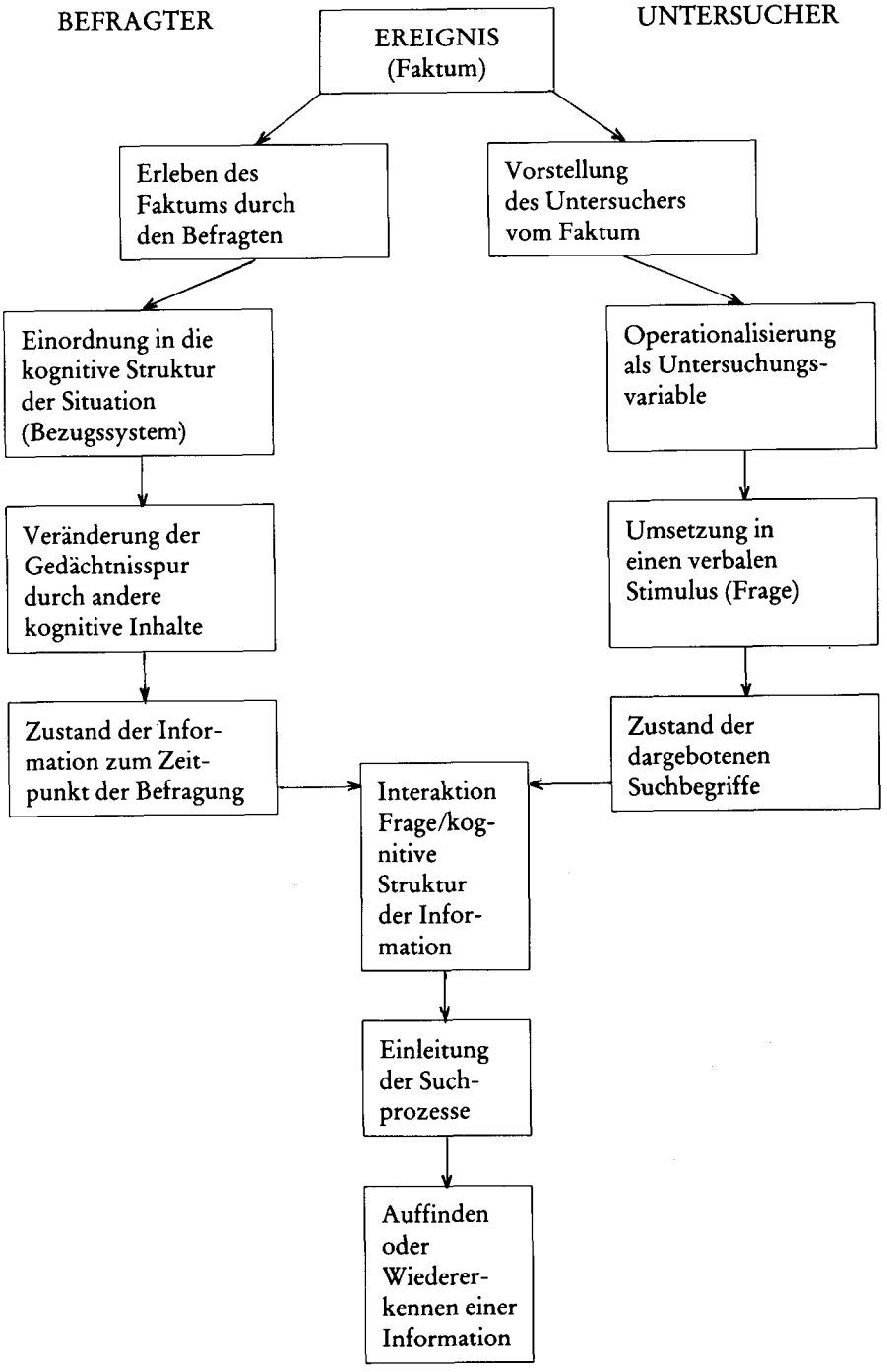

Abb. 2: Interaktion von Frage und kognitiver Struktur der Information

dern z.B. auch auf dem Hintergrund der Bezugssysteme ‚Symptome' (Schmerzen), ‚Ereignisse' (Krankheitsfalle) und ‚Lebensgewohnheiten' (Diät, Medikamenteneinnahme etc.), so lassen sich die Nennungen tatsächlich vorliegender Krankheiten ganz beträchtlich erhöhen (Cannell et al. 1977). Veränderungen der Gedächtnisspur durch andere kognitive Inhalte müssen dabei in Betracht gezogen werden. So kann z.B. ein Krankenhausaufenthalt in der Kindheit durch spätere Krankenhausaufenthalte an subjektiver Bedeutsamkeit verlieren. Bei der Umsetzung der Untersuchungsvariable ‚Krankenhausaufenthalte' wäre dies z.B. dadurch zu berücksichtigen, daß man den Probanden fragt, ob ein berichteter Krankenhausaufenthalt tatsächlich der erste Aufenthalt in einem Krankenhaus war.

Die in Abb. 2 aufgelisteten Verarbeitungsprozesse auf seiten des Befragten sollten vom Untersucher stets im Auge behalten werden, wenn er versucht, einen Sachverhalt mit Hilfe von Fragen zu erfassen. Soweit es sich um Wissenstatbestände handelt, muß dabei eine Optimierung der Befragungssituation im Hinblick auf die Wiedergabe (Reproduktion bzw. Wiedererkennen) angestrebt werden, ein Gebiet, das im Vergleich zum optimalen Lernen in der empirischen Forschung bisher recht stiefmütterlich behandelt worden ist (Cannell et al. 1977). Nur dann wird man erreichen können, daß die Beantwortungsfehler, von denen der Untersucher gerne spricht und die er nur in Ausnahmefällen (methodologischen Studien) als solche erkennen kann, nicht in Wahrheit Befragungsfehler sind.

1.3 Einordnung der Fragebogenkonstruktion in die Stadien einer Befragung

Die Konstruktion eines Fragebogens ist eingebettet in den Prozeß der Planung, Vorbereitung, Durchführung und Auswertung einer Befragung oder Testung. Um diese Einbettung deutlich zu machen, führen wir nachstehend die wichtigsten Stadien dieses Prozesses für demoskopische Befragungen auf. Sinngemäß sind sie auch auf diagnostische Untersuchungen übertragbar. Die Fragebogenkonstruktion im engeren Sinne umfaßt dabei Entwurf, Erprobung und Revision eines Fragebogens, nötigenfalls mehrfach wiederholt.

a. Sichtung und Aufarbeitung der zum Themenbereich vorliegenden theoretischen Ansätze und empirischen Befunde, besonders soweit es sich um Ergebnisse früherer Befragungen handelt.
b. Formulierung der genauen Fragestellung, soweit es sich um eine Befragung mit hypothesenprüfendem Anspruch handelt auch der Hypothesen, dabei auch explizite Festlegung der Grundgesamtheit, auf die sich die Hypothesen beziehen.
c. Differenzierung der Fragestellung in einzelne, als Untersuchungsvariablen

geeignete Aspekte (Definition der abhängigen Variablen, der ‚Programmfragen' i. S.v. Noelle 1963).

d. Festlegung der interessierenden und zu erfassenden Kovarianten.

e. Erweiterung der Informationsbasis z.B. durch Expertenbefragungen, freie Explorationen mit Betroffenen, Gruppendiskussionen, aber auch durch Analyse von Medien u.ä. Dieser Schritt kann evtl. auch schon an früherer Stelle erfolgen.

f. Soweit erforderlich Revision der in b.-d. getroffenen Festlegungen aufgrund der Erkenntnis in e.

g. Festlegung der genauen Untersuchungsmethode, vor allem Entscheidung zwischen mündlicher, persönlich-schriftlicher oder unpersönlich-schriftlicher Befragung auf dem Hintergrund der inhaltlichen Festlegungen in den Schritten a.-f.

h. Erstellung eines Fragebogenentwurfs durch
   - Operationalisierung der Untersuchungsvariablen, d.h. ihre Umsetzung in ‚Ermittlungsfragen' (Noelle 1963) unter Berücksichtigung möglicher Formulierungseffekte,
   - Festlegung des Befragungsverlaufes durch Definition der Reihenfolge der Fragen unter Berücksichtigung möglicher Reihenfolgeeffekte,
   - Festlegung der Fragebogengestaltung (Layout) unter Berücksichtigung möglicher Einflüsse auf Antwortbereitschaft und Art der Beantwortung.

   Bei der Erstellung des Fragebogenentwurfs sind außer den inhaltlichen Gesichtspunkten und den genannten möglichen Einflüssen auf das Antwortverhalten besonders auch die jeweilige Zielpopulation (Grundgesamtheit) und die genaue Befragungsmethode zu bedenken. Außerdem muß schon in diesem Stadium im Hinblick auf Kodierungen die spätere Auswertung genau geplant werden.

i. Erprobung des Fragebogens (Pretest) an einer im Vergleich zur Hauptuntersuchung kleineren, aber für die Grundgesamtheit ebenfalls repräsentativen Stichprobe. Im Rahmen dieser Erprobung sollten einerseits Interviewer die Vpn bei der Beantwortung der Fragen möglichst systematisch beobachten, um Verständnisschwierigkeiten besonders durch Fragenformulierungen und Fragenreihenfolge (Verzweigungen) aufzudecken. Zum anderen sollten die Antworten der Vpn verwendet werden
   - zur Entdeckung häufig ausgelassener Fragen,
   - zur Ermittlung und Analyse von Antwortverteilungen (mehrgipflige Verteilungen weisen häufig auf Mehrdeutigkeit der Fragen hin),
   - zu Itemanalysen bzw. -Validierungen, je nach verfolgtem Validitätskonzept (dabei können z.B. Konsistenzanalysen, Stabilitätsbestimmungen, Kriteriumskorrelationen der Items und Faktorenanalysen der Iteminterkorrelationen angezeigt sein),
   - zur Entdeckung von Verfälschungsmöglichkeiten (soziale Erwünschtheit, response sets) und Tendenzen zu unsorgfältiger (zufälliger) Be-

antwortung (dabei ist auch eine Bestimmung des akzeptablen Fragebogenumfangs vorzunehmen und zu ermitteln, ob zusätzliche Instruktionen erforderlich sind).

Auf dieser Stufe, evtl. auch schon bei der Erstellung des ersten Fragebogenentwurfes, können evtl. Sprachanalysen der Fragenformulierungen (vgl. z.B. Ash & Edgell 1975) oder Untersuchungen zur Unangenehmheit bestimmter Fragen in Subpopulationen (vgl. Koolwijk 1968) nützlich sein.

j. Revision des Fragebogenentwurfs (h), evtl. auch der Entscheidung über die Befragungstechnik (g), die Fragestellung (b) bzw. die Untersuchungsvariablen (c) und Kovarianten (d) und erneute Erprobung des veränderten Fragebogenentwurfs. Veränderungen und erneute Erprobungen sind solange zu wiederholen, bis eine befriedigende ‚Endfassung' erstellt ist.

k Wahl eines angemessenen Verfahrens der Stichprobenziehung und Durchführung der Stichprobenziehung.

l. Soweit persönliche Befragungen durchgeführt werden sollen, Auswahl und Schulung der Interviewer.

m. Durchführung der Befragung, im Falle unpersönlicher (z.B. postalischer) Befragung mit mehrstufigem ‚Nachfassen'.

o. Formale Auswertung einschließlich statistischer Hypothesenprüfung.
Bei der Auswertung eines Fragebogens empfiehlt es sich, Plausibilitätskontrollen der Antworten durchzuführen, um Fälle zu entdecken, in denen z.B. die Eintragung der Antwort durch Interviewer oder Befragten an der falschen Stelle oder ohne Berücksichtigung des Inhalts nach einem bestimmten System erfolgt ist. Das gilt besonders beim Einsatz von Fragebogen zu diagnostischen Zwecken, da andernfalls gravierende Folgen für den Probanden eintreten können. Eintragungen an falscher Stelle lassen sich in der Regel natürlich nur entdecken, wenn richtige Antworten existieren.

p. Interpretation der Ergebnisse unter Berücksichtigung der allgemeinen methodenspezifischen Beschränkungen bzw. Validitätsvorbehalte und gegebenenfalls auch der tatsächlich aufgetretenen methodischen Unzulänglichkeiten (z.B. Rücklauf bei postalischer Befragung).

q. Einordnung der Befunde in den Wissensbestand, vor allem Darlegung von Abweichungen und Übereinstimmungen mit
- fremden Befunden mit vergleichbarer Methode,
- Befunden aufgrund andersartiger Methoden.

r. Gegebenenfalls Erarbeitung von Hinweisen auf Probleme und Sachverhalte, die bei weiteren Untersuchungen innerhalb des Themenbereichs beachtet werden sollten.

Viele der o.a. Gesichtspunkte sind nicht für die Befragung als Untersuchungsmethode spezifisch oder berühren nicht die Konstruktion des Fragebogens im engeren Sinne. Ihre Auflistung macht deutlich, daß die Fragebogenkonstruktion nur ein (wenn auch wichtiger) Schritt im Zusammenhang mit einer empirischen Untersuchung oder diagnostischen Urteilsbildung ist. In den folgen-

den Abschnitten werden vor allem die bei der Erstellung des Fragebogenentwurfs erforderlichen Überlegungen zur Fragenformulierung, zur Festlegung der Fragenreihenfolge und zur äußeren Gestaltung des Fragebogens (Layout) eingehender behandelt. Ein Handbuchartikel muß sich dabei wegen des begrenzten verfügbaren Raumes mehr oder weniger auf eine Aufzählung von Problemen und Lösungsansätzen beschränken. Mindestens in einzelnen Aspekten weitergehende, teilweise auch stärker praxisbezogene Darstellungen sind unter vielen anderen Jonsson (1957), Noelle (1963), Richardson et al. (1965), Stroschein (1965), Oppenheim (1966), Phillips (1966, 1970), Richter (1969), Atteslander (1971), Friedrich (1971), Mayntz et al. (1971), Münch (1971), Friedrich (1973), Muccielli (1973), Scheuch (1973), Koolwijk & Wieken-Mayser (1974), Kreutz & Titscher (1974), Friedrich & Hennig (1975), Holm (1975b), Kirschhofer-Bozenhardt & Kaplitza (1975), Burisch (1976), Karmasin & Karmasin (1977). Über Befragungsmethoden allgemein informieren außerdem Jetzschmann & Kallabis (1966), Cannell & Kahn (1968), Anger (1969), König (1972), Maccoby & Maccoby (1972), Sheatsley (1972), Behrens (1974), Schreiber (1974), Steward & Cash (1974), Wilk (1974), Holm (1975a).

2. Fragentypen

2.1 Zielsetzungen von Fragen

Nicht jede in einem Fragebogen verwendete Frage hat die Aufgabe, inhaltlich oder im Sinne von ‚Zeichen' (signs) interessierende Informationen zu erheben. Bestimmte Fragen haben Merkmale im Auge, die nicht als solche interessieren und nur im Hinblick auf Interpretationen oder Erklärungen der eigentlich thematischen Sachverhalte von Bedeutung sind (dazu gehören in der Regel die Angaben zur Person des Interviewten). Außerdem gibt es Fragen, die nur innerhalb des Fragebogens bzw. der Befragung bestimmte Aufgaben zu erfüllen haben. In leichter Abwandlung der Terminologie vor Stroschein (1965) sei die erste Gruppe als die der ‚*Ergebnisfragen*', die zweite als die der ‚*Korrelationsfragen*' und die dritte als die der ‚*instrumentellen Fragen*' angesprochen.

Diese letztgenannte Gruppe, für die auch die Bezeichnung ‚Funktionsfragen' gebräuchlich ist (z.B. Anger 1969), läßt sich weiter unterteilen in:

a. *Kontrollfragen,* und zwar einmal *Erhebungskontrollfragen* (z.B. Fragen nach Ort und Zeitpunkt des Interviews) zur Gewährleistung der Nachprüfbarkeit und *Auskunftskontrollfragen* (z.B. Wiederholungsfragen), mit dem Ziel der Ermittlung der Konsistenz des Antwortverhaltens,
b. *Ablauf-Ordnungsfragen,* insbesondere *Filter-Fragen* mit der Aufgabe, Befragte mit bestimmten Merkmalen von bestimmten Fragen auszuschließen,

und *Gablungs-* oder *Verzweigungs-Fragen* mit der Aufgabe, antwortabhängig zu verschiedenen Folgefragen zu verzweigen,

c. *befragungstaktische Fragen* wie *Einleitungsfragen* mit der Aufgabe, den Kontakt zur Befragungsperson herzustellen, *Unterweisungsfragen* zur Information der Versuchsperson über evtl. nicht hinreichend bekannte Sachverhalte, *Ablenkungs- und Pufferfragen* zur Verdeckung der Zusammenhänge zwischen Fragen bzw. zur Vermeidung von Einflüssen vorangegangener auf nachfolgende Fragen (Halo-Effekte) und *Füllfragen* mit dem Ziel, dem Fragebogen in der Wahrnehmung des Befragten eine von der tatsächlichen abweichende inhaltliche Ausrichtung zu verleihen.

Erdos (1970) hält in postalischen Befragungen sogenannte ‚*return getters*', also Fragen mit der alleinigen Aufgabe der Rücklaufsteigerung für angebracht. Für schwach- oder teilstandardisierte mündliche Befragungen werden z.B. von Stollberger (1966) und Atteslander (1971) außerdem ‚*Sondierungsfragen*' (das sind Nachfragen bei unzureichenden ersten Antworten) und ‚*Rangierfragen*' (Fragen, die im Falle von Abschweifungen den Befragten wieder auf das eigentliche Thema lenken sollen) als Typen taktischer Fragen angeführt. Weitere, zum Teil kurios anmutende Arten von Funktionsfragen beschreibt Noelle (1963, 74, sowie 1974, z.B. sogenannte Spielfragen nach der Beurteilung von Frisuren, Kleidern etc. mit dem Ziel, das Interesse der Vp am Interview zu erhalten).

Auch in diagnostischen Fragebogen wird von instrumentellen Fragen mehr oder weniger Gebrauch gemacht (vgl. Mittenecker 1971). So enthält etwa die englische Variante des MPI von Eysenck 12 Pufferfragen, die nicht in die Auswertung eingehen. Von den 566 Items des MMPI sind immerhin 166 instrumentelle, genauer ‚Auskunfts-Kontroll-Items', die unterschiedliche Kontrollstrategien verfolgen: Itemwiederholungen zur Bestimmung der Konsistenz, Lügenitems (die wahrheitsgemäß nur in ganz bestimmter Richtung beantwortbar sind) und Sorgfaltsitems (die von fast allen Probanden in einer Richtung beantwortet werden, vgl. Hathaway & Mc Kinley 1963).

Kreutz & Titscher (1974) fordern, daß Itemwiederholungen nur in so großen Abständen erfolgen, daß die Antworten stochastisch voneinander unabhängig sind. Sie warnen - allerdings ohne dies empirisch zu begründen - vor den Folgen für die Motivation und das Antwortverhalten der Vpn, wenn diese die Wiederholung (und damit die Kontrollabsicht) bemerken.

Dem wäre entgegenzuhalten, daß die Entdeckung eingebauter Kontrollmechanismen durch die Vpn auch positive Wirkungen (größere Sorgfalt, größere Ehrlichkeit) haben kann. Nach den Ergebnissen von Hoeth & Köbler (1967) scheint es u.U. sogar vorteilhaft zu sein, wenn Vpn auf solche Mechanismen eigens hingewiesen werden (vgl. auch Ehlers 1973).

Ablauf-Ordnungsfragen (Filter- und Verzweigungsfragen) werden dagegen in diagnostischen Fragebogen nicht eingesetzt, da diese Fragebogen in der Regel schriftlich beantwortet werden und dadurch die Verwendung dieser Instrumente mit Schwierigkeiten verbunden wäre (Richter 1969). Überlegungen zu individualisiertem (antwortabhängigem) Testen im Bereich der Persönlichkeitsdiagnostik - allerdings unter Aufgabe des traditionellen Fragebogenkonzeptes - sind damit natürlich nicht ausgeschlossen.

2.2 Frageninhalte

Als mögliche Inhalte von Ergebnisfragen (vgl. 2.1) kommen vor allem in Betracht (Stollberger 1966, Holm 1975b):

- Fakten (z.B. Lebensalter, Besitz eines Farbfernsehgerätes).
- Wissen (‚Wie heißt der Bundesfinanzminister?'). Während bei Faktenfragen das Interesse des Untersuchers sich auf das Faktum richtet, d.h. er etwas über das Faktum (Verbreitung von Farbfernsehgeräten) erfahren will, interessiert bei Wissensfragen letztlich nicht das Faktum (der Name des Politikers), sondern die Informiertheit des Befragten.
- Beurteilungen, Bewertungen, Meinungen bzw. Einstellungen (‚Was halten Sie von Kernkraftwerken?'). Soweit der Untersucher an Informationen über Sachverhalte (Kernkraftwerk) interessiert ist, spricht man von Beurteilungs- oder Bewertungsfragen, steht dagegen der Befragte im Mittelpunkt des Interesses, von Meinungs- bzw. Einstellungsfragen (Holm 1975 b).
- Verhalten bzw. Handlungen (‚Treiben Sie regelmäßig Sport?'). Dabei handelt es sich nur dann um Faktenfragen, wenn nach gegenwärtigem oder früherem Verhalten, nicht aber wenn nach zukünftigem oder hypothetischem Verhalten gefragt wird.
- Motive (‚Warum sind Sie dieser Meinung?').

Nach Cannell et al. (1977) ist davon auszugehen, daß unabhängig von der Fragenform Einstellungs- bzw. Motiv- im Vergleich zu Fakten- oder Wissensfragen ‚schwieriger' sind: Sie führen' häufiger zu ausweichenden Antworten (‚weiß nicht' o.ä.), häufiger zu Rückfragen das Fragenverständnis betreffend und häufiger zu qualifizierten (eingeschränkten) Antworten.

Das bedeutet allerdings nicht, daß Fakten- oder Wissensfragen auch zu ‚richtigeren' Antworten führen müßten bzw. daß es einfacher sei, Wissen und Fakten durch geeignete Fragen zu erfassen (Mauldin & Marks 1950).

Eine Grundforderung an Fragen ist die nach Eindeutigkeit in einem gegebenen Zusammenhang. Daraus ergibt sich, daß eine Frage sich stets nur auf einen bestimmten Inhalt beziehen darf. In diesem Sinne wäre z.B. das Item (MMPI Nr. 307) ‚Bei einigen Spielen lehne ich es ab, mich zu beteiligen, weil ich sie

nicht gut kann. richtig/falsch?‘ nicht eindeutig, da hier gleichzeitig nach einem Faktum und nach einem Motiv gefragt wird. (Trotz dieses inhaltlichen Mangels ist nicht ausgeschlossen, daß im Rahmen eines empirischen Validitätskonzeptes ein solches Item sich als brauchbar erweist.)

2.3 Direktheit einer Frage

Neben der direkten Frage (‚Wie alt sind Sie?‘) zu einem bestimmten Sachverhalt gibt es meist mehrere Möglichkeiten zur Formulierung indirekter Fragen, die nach verschiedenen Gesichtspunkten eingeteilt werden können. Holm (1974a) bezeichnet als indirekte Fragen solche, die sich direkt auf einen verwandten Sachverhalt beziehen, besonders

- ‚Fragen durch die Hintertür‘ (‚Welcher Jahrgang sind Sie?‘),
- ‚Fragen über Ersatzdimensionen‘ (‚Wieviele Zähne fehlen Ihnen?‘).

Stroschein (1965), der von ‚unmittelbaren‘ und ‚mittelbaren‘ Fragen spricht, führt als Beispiele an:

- Assoziationsfragen, d.h. Fragen, die darauf abzielen, die mit einem bestimmten Gegenstand verbundenen Vorstellungen zu ermitteln,
- Projektionsfragen, d.h. Fragen, die den Probanden veranlassen sollen, in eine Situation oder Person eigene Gefühle oder Stimmungen hineinzuverlagern.

Weiterhin werden ‚Dialogfragen‘ (Stroschein 1965, ‚A sagt . . ., B sagt. . . .‘) häufig so gestellt, daß der Befragte seine Meinung indirekt in Form eines Schiedsspruches (‚Wer hat recht?‘) zwischen A und B zum Ausdruck bringen soll. Ähnliches gilt für sogenannte ‚hypothetische Situationen‘ (‚Stellen Sie sich bitte vor, Herr X . . .‘), die prinzipiell in eine direkte (‚Wie würden Sie sich verhalten?‘) oder eine indirekte Frage (‚Wie glauben Sie, daß Herr X sich verhält?‘) münden können (vgl. Friedrichs 1973). In gewisser Hinsicht als indirekt muß man wohl auch ‚Fragen mit Zitaten‘ ansehen, bei denen eine bestimmte Meinung vom Probanden nicht unmittelbar, sondern über den Umweg der Stellungnahme zu einem Zitat erfragt wird, das der Untersucher einer in der Regel bekannten Persönlichkeit in den Mund legt. Wie auch empirisch vielfach demonstriert wurde, hängt die Antwort außer vom Inhalt des Zitates natürlich stark von der Einstellung des Probanden gegenüber der zitierten Persönlichkeit ab (vgl. z.B. Roslow et al. 1940, Stroschein 1965). Weitere Beispiele indirekter Fragen finden sich u.a. bei Karmasin & Karmasin (1977) und bei Maccoby & Maccoby (1972).

Indirekte Fragenformulierungen werden meist für sogenannte schwierige, heikle, unangenehme, peinliche Sachverhalte verwendet. Dazu gehören für viele Befragte Fragen nach dem Einkommen, der Kindererziehung, der Allge-

meinbildung, der Sexualität, den Familienverhältnissen (Friedrichs 1973) und vor allem nach der körperlichen Sauberkeit (Scheuch 1973). Aber auch für die Leserschafts-Forschung werden indirekte Techniken vorgeschlagen (vgl. z.B. Schyberger 1968), um Verfälschungen durch die Auflagenstärke zu vermeiden. Barton (1958) hat ironisierend am Beispiel der (direkten) Frage ‚Haben Sie Ihre Frau umgebracht?' die verschiedenen mehr oder weniger indirekten Ansätze zusammengestellt:

- die Möglichkeitsfrage (‚Könnte es sein, daß . . .'),
- die Kartenfrage (Identifizierung der Kennung einer Karte mit der zutreffenden Antwort durch die VP),
- der Jedermann-Ansatz (‚Viele haben in letzter Zeit . . . und Sie?'),
- der Andere-Ansatz (‚Kennen Sie Leute, die . . . und Sie?'),
- die Urnentechnik (Antwort auf direkte Frage kommt in verschlossenem Umschlag in eine Urne),
- die projektive Technik (‚Welche Gedanken kommen Ihnen bei diesen Bildern . . .?'),
- die ‚Kinsey-Technik' (dem evtl. peinlichen Verhalten wird durch die Formulierung die Eigenschaft des Selbstverständlichen verliehen).

Inwieweit indirekte Fragen, vor allem Projektions- und Assoziationsfragen, die in sie gesetzten Erwartungen erfüllen, scheint weitgehend ungeklärt zu sein (Stroschein 1965, Friedrichs 1973).

Karmasin & Karmasin (1977) bemerken kritisch, daß solche Fragen für die Vpn grundsätzlich mehrdeutig seien. Die Vp kann bei einer indirekten Frage (vgl. o. ‚Wer hat recht?')

- ihre eigene Meinung äußern,
- sich überlegen, wie das ‚mehrheitlich' wohl gesehen wird, oder
- unter Zugrundelegung irgendwelcher Normen (wie es sein sollte) entscheiden.

Die Interpretation der Antwort auf eine indirekte Frage ist demnach nur auf dem Hintergrund einer Theorie möglich, die den Zusammenhang zwischen Sachverhalt und Frage herstellt (z.B. eine Theorie der Projektion, vgl. Anger 1969, aber auch ein empirisches Validitätskonzept). Dabei ist die Indirektheit einer Frage als Kontinuum anzusehen. Der Grad der Indirektheit bestimmt sich nach der Komplexität der Mechanismen, die die Theorie zur Vermittlung zwischen Sachverhalt und Antwort annimmt (Cannell & Kahn 1968). Indirekte Fragen sind demnach höchstens so brauchbar, wie es die ihnen zugrundeliegende Theorie ist.

Für Items in diagnostischen Fragebogen hat Ellis (1947) in einer umfangreichen Untersuchung die direkte (hier personalisierte) Formulierung (‚Ich . . .') mit der indirekten (unpersönlichen, ‚Leute, die . . . sind . . .') verglichen. Dabei

zeigte sich für die indirekte (unpersönliche) Form eine stärkere Abhängigkeit von anderen Fragebogenmerkmalen (positive oder negative Formulierung) als für direkte (personalisierte) Items. Vor allem aber war auch bei heiklen (sozial mutmaßlich unerwünschten) Sachverhalten keine Überlegenheit der indirekten Formulierung nachweisbar.

2.4 Formale Fragenkonstruktion

Prinzipiell kann ein Fragebogenitem als Frage (‚Besitzen Sie ein Farbfernsehgerät? ja/nein') oder als Statement mit Aufforderung zur Stellungnahme (‚Ich besitze ein Farbfernsehgerät. stimmt/stimmt nicht') formuliert sein. In der Literatur über Fragebogenkonstruktion scheint „. . . der formale Unterschied zwischen Behauptungen (Statements) und Fragen noch gar nicht recht bewußt geworden . . ." zu sein (Kreutz & Titscher 1974, 52). Über Auswirkungen dieses Unterschiedes läßt sich deshalb derzeit nur spekulieren. Kreutz & Titscher (1974) vermuten z.B., daß die in vielen diagnostischen Fragebogen beobachtete Zustimmungstendenz (aquiescence) mit der üblichen Formulierung der Items als Statements zusammenhängt und sich weniger ausgeprägt zeigen würde, hätten diese als Fragen in stärkerem Maße den Charakter des ‚Unentschiedenen'.

2.4.1 Offene und geschlossene Fragen

Ausgehend von den Antwortmöglichkeiten auf eine Frage unterscheidet man *offene* von geschlossenen Fragen (z.B. Stollberger 1966, Friedrichs 1973, Cannell et al. 1977). Eysenck (1953) spricht vom ‚kreativen' und vom ‚selektiven' Antworttyp, und Stroschein (1965) hebt die ‚inkategorialen Fragen', d.h. die Fragen, bei denen die Auswertungsgesichtspunkte für die Vpn nicht erkennbar sind, von den ‚kategorialen Fragen' ab, die in irgendeiner Form (meist durch Antwortvorgaben) Informationen über die Auswertungsgesichtspunkte enthalten.

Geschlossene Fragen lassen sich ‚kategorie-neutral' (Stroschein 1965), d.h. ohne durch die Vorgabe der Kategorien das Antwortverhalten gravierend zu beeinflussen, zu einem bestimmten Themenbereich nur formulieren, wenn alle möglichen Antworten bereits bekannt sind. Aus diesem Grund wird man bei geringem Vorwissen zwangsläufig eher zu offenen Fragen greifen (Friedrichs 1973). Daneben ist aber auch zu bedenken, daß offene Fragen ‚freie Reproduktion', geschlossene Fragen nur ‚Wiedererkennen' fordern. Schon dadurch sind offene Fragen schwieriger. Cannell et al. (1977) berichten von ca. 30% unbrauchbaren Antworten (z.B. Auslassungen und nicht fragenbezogene Ausführungen) bei offenen im Vergleich zu nur 6% bei geschlossenen Fragen

zum gleichen Themenkomplex, merken allerdings selbstkritisch an, daß man der geschlossenen Antwort ihre Unbrauchbarkeit oft auch nur nicht ansieht. In die Richtung größerer Schwierigkeit der offenen Frage deuten aber (abgesehen von Plausibilitätsüberlegungen hinsichtlich erforderlicher Ausdrucksfähigkeit und evtl. Schreibgewandtheit) auch Ergebnisse über Rückläufe bei postalischen Befragungen: Falthzik & Carroll (1971) erzielten bei einem aus nur einer Frage bestehenden Fragebogen einen Rücklauf von 78%, wenn die Frage geschlossen formuliert war, und von nur 27%, wenn es sich um eine offene Frage handelte. Einen Unterschied von immerhin noch 60% zu 50% berichtet Erdos (1970). Andererseits fand Richter (1969) für umfangreichere Fragebogen zwar ebenfalls eine Senkung des Rücklaufs durch eine große Zahl offener Fragen, stellt aber günstige Auswirkungen auf den Rücklauf fest, wenn zu jedem Themenbereich neben geschlossenen Fragen auch eine offene Frage vorgesehen ist, was er auf eine ‚Ventilfunktion' offener Fragen und auf verminderte ‚Ermüdung' durch vereinzelte Zwischenschaltung solcher Fragen zurückführt.

Zu bedenken ist auch, daß die ‚Abdeckung' eines Themenbereiches in der Regel erheblich mehr geschlossene als offene Fragen erfordert (Cannell et al. 1977).

Letztlich wird aber entscheidend für die Wahl offener oder geschlossener Fragen sein, ob die Reproduktions- oder die Wiedererkennungsleistung dem untersuchten Inhalt angemessener ist. So demonstrierten z.B. Roslow et al. schon 1940, daß bei der Ermittlung von Kaufgewohnheiten und Verbreitungsgraden in offenen Fragen (freie Reproduktion) häufiger als in geschlossenen Fragen (Wiedererkennen) und auch häufiger als objektiv zutreffend die Produkte mit hohen Marktanteilen genannt wurden (ähnliche Ergebnisse berichtet Stroschein 1965). Wie genau hier die geschlossene Frage den wahren Sachverhalt trifft, hängt entscheidend von der Vollständigkeit der Vorgaben ab (Roslow et al. 1940). Ungeeignet sind offene Fragen auch zur Erfassung ‚alltäglicher' Sachverhalte (Payne 1951), die in der Regel unscheinbar, d.h. nicht als Figur abgehoben sind (auf eine offene Frage nach dem Tagesablauf hin werden viele Vpn z.B. das ‚Zähneputzen' nicht erwähnen).

Dagegen ist die offene Frage wegen geringen Vorwissens, hoher Differenziertheit bzw. Komplexität des Sachverhaltes (Friedrichs 1973), hoher problemspezifischer Validität der Reproduktionsleistung oder Unangenehmheit bzw. SD-Empfindlichkeit des Inhalts (Sudman & Bradburn 1974, Bradburn & Sudman 1979) in anderen Fällen durchaus angezeigt, allerdings bereitet ihre Auswertung erhebliche Schwierigkeiten. Die freien Antworten werden üblicherweise nach Art einer systematischen Inhaltsanalyse (vgl. z.B. Friedrichs 1973) mit Hilfe eines eigens erstellten Kategoriensystems klassifiziert (grundsätzliche Überlegungen speziell für den Fall offener Fragen finden sich z.B. bei Lazarsfeld & Barton 1955, Vorschläge zur ‚Automatisierung' unter Einsatz von EDV

berichten Friesbie & Sudman 1968). Dabei zeigt sich allerdings häufig, daß relativ viele Antworten (mindestens 10%) bei Zugrundelegung eines noch überschaubaren Systems nicht klassifizierbar sind und daß intraindividuelle Stabilität und interindividuelle Objektivität der Klassifizierung zu wünschen übrig lassen (Stroschein 1965). Außerdem führen vorgegebene Antwortkategorien und Klassifikationen freier Antworten unter Zugrundelegung dieser ‚Antwortvorgaben' häufig zu recht unterschiedlichen Ergebnissen (Stroschein 1965), was sich allerdings ebensogut als Argument gegen offene wie gegen geschlossene Fragen verwenden läßt.

Die Unterscheidung zwischen offenen und geschlossenen bzw. inkategorialen und kategorialen Fragen ist entgegen dem ersten Anschein nicht streng durchzuführen (Maccoby & Maccoby 1972). Einmal ist es möglich, in der Fragenformulierung nur bestimmte Antwortkategorien aufzuführen, aber klarzumachen, daß es weitere gibt (‚Haben Sie gestern abend ferngesehen oder . . .?'). Noelle (1963) spricht dann von ‚halboffenen' Fragen. Zum anderen kann eine Frage auch nur ‚scheinbar inkategorial' (Stroschein 1965) sein, weil die möglichen Kategorien jedem Befragten evident sind (das gilt z.B. für die ‚offene' Frage: ‚Welche Haarfarbe haben Sie?').

Im Falle mündlicher Befragung muß schließlich berücksichtigt werden, daß eine Frage aus der Sicht des Probanden offen sein, die Antwort vom Interviewer aber direkt klassifiziert werden kann. Gegen diese sogenannte ‚Feldverschlüsselung' (vgl. auch Noelle 1963) werden allerdings gravierende Einwände erhoben (Anger 1969), da sie den Interviewer häufig überfordere und vor allem nicht nachprüfbar sei. Andererseits werden auch bei der schriftlichen Erfassung freier Antworten durch Interviewer erhebliche ‚Verluste' beklagt, so daß Quantifizierungen mit Vorsicht behandelt und Hypothesenprüfungen auf der Grundlage offener Fragen nicht vorgenommen werden sollten (Anger 1969).

Da offene Fragen im Rahmen eines mündlichen Interviews eher der alltäglichen Konversation entsprechen und daher natürlicher wirken (Maccoby & Maccoby 1972, Karmasin & Karmasin 1977), werden sie auch aus vorwiegend befragungstaktischen Gründen eingesetzt.

2.4.2 Arten geschlossener Fragen

Streng kategorial sind Fragen, wenn sie explizit alle Antwortmöglichkeiten enthalten. Bei der Formulierung ist im allgemeinen sicherzustellen, daß die Fragen kategorie-neutral sind, d.h. daß keine der Antwortkategorien durch die Formulierung begünstigt wird.

Neutralität im Hinblick auf die *Reihenfolge* der Vorgaben ist allerdings nur erreichbar, wenn mit verschiedenen Fassungen des Fragebogens gearbeitet und

dabei die Reihenfolge der Vorgaben variiert wird (Split-ballot-verfahren, gegabelte Befragung).

Zur Vermeidung von Einflüssen z.B. der sozialen Erwünschtheit kann es in Ausnahmefällen angezeigt sein, bewußt und geplant von der Ausgewogenheit der Kategorien abzuweichen. So berichtet z.B. Stroschein (1965), daß die Frage ‚Werden Sie bestimmt mitwählen oder werden Sie vielleicht nicht zur Wahl gehen?' zu einer sehr guten Prognose der Wahlbeteiligung führte, vermutlich gerade weil sie nicht kategorie-neutral ist (vgl. auch 3.1.4).

Nach der Zahl der Antwortvorgaben lassen sich kategoriale Fragen weiter unterteilen in *Alternativfragen* und in *Selektivfragen* (Listenfragen, Auswahlfragen, auch Katalogfragen, vgl. Anger 1969, oder Multiple-Choice-Fragen, vgl. Payne 1951, obschon der letztgenannte Begriff vorwiegend bei Verwendung entsprechender Fragen im Rahmen diagnostischer Fragebogen gebraucht zu werden scheint). Ein Spezialfall sind sogenannte Eigenschaftswörterlisten (adjective check lists). Dabei ist zu berücksichtigen, ob

- die Zahl zulässiger Nennungen unbestimmt bleiben,
- die Zahl zulässiger Nennungen nach unten und/oder nach oben begrenzt werden oder ob
- zu jeder der aufgeführten Kategorien eine Einzelantwort gefordert werden soll.

Problem der Alternativ-, vor allem der Listenfragen ist die Bevorzugung von Vorgaben in Abhängigkeit von ihren Positionen (darauf wird in 3.1.3 näher eingegangen).

Selektivfragen mit vielen Vorgaben sind schwierige Fragen (nach Richter 1969 senken sie den Rücklauf in postalischen Befragungen) und machen im mündlichen Interview Hilfsmittel (Vorlagen, Kartensätze) erforderlich. Werden solche nicht verwendet, erweisen sich umfangreiche Selektivfragen als besonders anfällig gegenüber Interviewereinflüssen (Cahalan et al. 1947).

Sonderfälle selektiver Fragen sind sogenannte *Skalafragen* (ordinale oder im engeren Sinne quantitative Fragen, vgl. Holm 1975 b), d.h. Fragen, deren Antwortkategorien geordnet bzw. graduell abgestuft sind. Verwendet man solche Fragen im Zusammenhang mit Beurteilungen, Bewertungen oder Einschätzungen, so spricht man auch von *Ratingskalen*. über diesen Ansatz orientieren zusammenfassend Guilford (1954) und Clauss (1968), neuere Ergebnisse zur Frage der optimalen Zahl von Skalenstufen referiert Mc Kelvie (1978), Varianten des innerhalb demoskopischer Befragungen besonders beliebten graphischen Rating diskutiert Narayana (1977). ‚Geeichte' numerisch-verbale Skalen für Häufigkeiten, Intensitäten, Wahrscheinlichkeiten und Beurteilungen des Zutreffens finden sich bei Rohrmann (1978).

Ratingskalen werden z.B. auch zur Erfassung des Zustimmungsgrades in diagnostischen Fragebogen verwendet, wobei die zweistufige Variante (ja/nein; richtig/falsch) wiederum als Spezialfall angesehen werden kann. Die ebenfalls verbreitete dreistufige Form (ja/?/nein) bietet zwar die Möglichkeit, über die Häufigkeit von ‚?' - Antworten die Aktualität der Fragen für den Probanden zu ermitteln (Heller & Krüger 1976), wirft andererseits aber wie alle mehrstufigen Kategorienskalen Probleme im Hinblick auf die Dimensionalität der Antworten auf (vgl. 1.1.3).

2.4.3 Sonderformen

Soweit nicht ohnehin schriftliche Befragung erfolgt, empfiehlt sich bei umfangreichen Selektivfragen die Verwendung von Vorlagen, entweder als

- Einzelvorlage mit allen Kategorien oder als
- Vorlagensatz (Kartensatz) mit je einer Vorlage für jede Kategorie.

Die Verwendung von Kartensätzen hat die Vorteile der leichten Variierbarkeit der Reihenfolge und der offenbar größeren Sorgfalt der Vpn bei der Entscheidung (Stroschein 1965). Vor allem werden Karten an späterer Stelle stärker beachtet als Vorgaben auf den unteren Plätzen einer Liste.

Zur Kennzeichnung der Vorgaben kann sich die Verwendung von *Symbolen* empfehlen (a, b . . .; weiß, schwarz . . .; 1, 2 . . . etc). Dadurch lassen sich Übertragungsfehler bei der Kornunikation Interviewer/Befragter vermindern. Vor allem aber wird der Befragte der Notwendigkeit enthoben, die Antwort explizit auszusprechen, was ihm bei unangenehmen Fragen peinlich sein könnte. Entgegen der landläufigen Erwartung haben sich nach Stroschein (1965) keine Antwortbevorzugungen durch bestimmte zur Kennzeichnung verwendete Symbole nachweisen lassen. Interviewereinflüsse auf die Antworten scheinen bei Verwendung von Symbolen geringer zu werden.

Als weitere Sonderform wäre das ‚semantische Differential' (Osgood et al. 1957) oder ‚Polaritätenprofil', eine Zusammenstellung von meist 18 bipolaren ‚Dimensionen' (adjektivischen Gegensatzpaaren, die jeweils durch eine siebenstufige Skala miteinander verbunden sind), anzuführen. Die Einschätzung von Begriffen in diesen 18 Polaritäten läßt sich - relativ unabhängig von den konkret verwendeten Eigenschaftspaaren - als Lokalisation dieser Begriffe in einem semantischen Raum mit den Dimensionen der Bewertung (gut/schlecht), der Aktivität (aktiv/passiv) und der Intensität (stark/schwach) interpretieren (vgl. auch Herrmann & Stäcker 1969).

Daneben werden vor allem im Bereich der kommerziellen Markt- und Meinungsforschung zahlreiche weitere, meist als ‚psychologisch' bezeichnete Techniken (von Farbwahltests bis zum Baumtest, vgl. Noelle 1963) unkritisch

und in einer Weise eingesetzt, die in erstaunlichem Kontrast zu den in anderem Zusammenhang (z.B. bei der Standardisierung der Fragenformulierungen und bei der Auswertung) erhobenen Forderungen nach Objektivität und Nachprüfbarkeit steht. Die nach Anger (1969, 583) ,,dringend benötigte Information über wichtige individuelle Merkmale und Eigenschaften" läßt sich, soweit es sich um psychische Eigenschaften handelt, nicht mittels irgendwelcher ‚Kurzverfahren' durch wenig geschultes Personal (Interviewer) nebenbei beschaffen. Den diesbezüglichen Ausführungen und Empfehlungen von Anger (1969) muß mit erheblicher Skepsis begegnet werden.

3. Fragenformulierung

Die Formulierung von Fragen wird von den meisten Autoren als ‚Kunst' betrachtet (u.v.a. Noelle 1963, Mayntz et al. 1971, Scheuch 1973), deren Grundlagen nicht am Schreibtisch, sondern nur in langer Erfahrung erworben werden könnten. Demgemäß sind auch ‚Regeln' für die Formulierung von Fragen, wie sie in der Literatur vielfach angeführt werden (z.B. Payne 1951, Edwards 1957, Noelle 1963, Maccoby & Maccoby 1972), soweit sie konkret sind, bestenfalls Ausfluß solcher Erfahrungen (oder wie im Falle von Holm 1975 b problematischer theoretischer Ansätze) und unbewiesen oder so abstrakt, daß sie zwar mit hoher Wahrscheinlichkeit nicht falsch, aber dafür auch wenig hilfreich sind (Kreutz & Titscher 1974). Die folgenden Ausführungen beinhalten weniger eine (erneute) Wiedergabe solcher ‚Regeln' als eine Beleuchtung grundsätzlicher Probleme und eine Zusammenstellung empirischer Befunde, die sich naturgemäß nur beschränkt verallgemeinern lassen.

Eine Frage stellt einerseits einen verbalen Stimulus für den Befragten, andererseits ein sprachliches Abbild eines Sachverhaltes dar (Kreutz & Titscher 1974). Bevor ein Sachverhalt sprachlich abgebildet werden kann, muß seine inhaltliche Struktur festliegen, d.h. der sprachlichen Formulierung einer Frage geht logisch die Erarbeitung einer inhaltlichen Konzeption voraus. In der Praxis allerdings lassen sich beide Schritte nicht wie hier mehr oder weniger streng trennen, mitunter ist sogar schwer zu entscheiden, ob es sich bei einem konkreten Problem um ein vorwiegend inhaltliches oder vorwiegend sprachliches handelt (vgl. z.B. die affektiv nicht neutralen Begriffe). Im Zusammenhang mit der Entwicklung der inhaltlichen Fragenkonzeption ist auch die Entscheidung über den zur Verwendung kommenden Fragentyp zu treffen. Die dabei zu beachtenden Gesichtspunkte sind bereits im vorstehenden Kapitel behandelt worden.

Wesentliche empirische Befunde zur Auswirkung der Fragenformulierung auf Antworten stammen aus der Zeit des Zweiten Weltkriegs, später traten Interviewereinflüsse in den Vordergrund des Interesses (Hartmann 1972).

3.1 Die inhaltliche Konzeption einer Frage

Ist bei der Planung einer Befragung (vgl. 1.3) die abhängige Variable (Programmfrage im Sinne von Noelle 1963) definiert worden (z.B. ‚Beurteilung der Wirtschaftspolitik der Bundesregierung durch die Bevölkerung'), so muß sie in einem weiteren Schritt operationalisiert, d.h. in eine oder mehrere Erhebungs- bzw. Testfragen (Noelle 1963) übersetzt werden.

3.1.1 Vorüberlegungen

Zunächst ist in Abhängigkeit vom Forschungsziel bzw. von der Programmfrage festzulegen, ob eine *globale* oder *differenzierte* Vorgehensweise oder eine Kombination beider erfolgen soll. Payne (1951) zeigt auf, daß z.B. bei Beurteilungen das aus mehreren Urteilen über Einzelaspekte sich ergebende Bild oft nicht mit dem einer global erfolgten Beurteilung übereinstimmt. Sodann ist zu klären, ob es sich um eine

- normative oder deskriptive,
- kognitive oder evaluative,
- allgemeine oder spezifische,
- abstrakte oder konkrete

Frage handeln soll (Karmasin & Karmasin 1977). Dabei sind allerdings die Freiheitsgrade des Untersuchers eingeschränkt, z.B. ist zu berücksichtigen, daß statt einer vorgesehenen kognitiven Beurteilung bei fehlender Informationsbasis auf Seiten der Vp leicht eine Bewertung (Evaluation) zustande kommen kann (z.B. bei einer Frage über Auswirkungen der Hochzinspolitik). Andererseits ist es sicher auch problematisch, die Menge der zulässigen Fragenkonzeptionen von vornherein stark einzuschränken, etwa auf spezifische und konkrete Ansätze (z.B. im Sinne von Payne 1951, der allgemein einen Bezug der Frage auf das ‚Wer, Wann, Warum, Wo, Wie' fordert, vgl. auch Anger 1969, Friedrichs 1973). So betonen Karmasin & Karmasin (1977), daß es zahlreiche Sachverhalte gebe, bei denen das ‚Dogma' von der konkreten und spezifischen Frage zu unsinnigen Konsequenzen führen müßte. Bei der Untersuchung von Lesegewohnheiten z.B. interessiert durchaus nicht, wie lange eine bestimmte Zeitung an einem bestimmten Tag gelesen wurde (spezifischer und konkreter Ansatz), sondern wie lange (ausführlich o.ä.) ‚in der Regel' die jeweilige Tageszeitung gelesen wird (allgemeiner, abstrakter Ansatz).

Soweit von der Vp Gedächtnisinhalte abgerufen werden müssen, ist zu überlegen, wie mit einer in der Regel identischen Frage die möglicherweise sehr unterschiedlichen Erfahrungen verschiedener Vpn aktualisiert werden können (Cannell & Kahn 1968) und wie eine für die Reproduktion bzw. das Wiedererkennen optimale, an ‚cues' reiche Situation hergestellt werden kann (Cannell et al. 1977, vgl. auch 1.2.3).

Für Beurteilungen ist festzulegen, ob sie auf dem Hintergrund eines impliziten Bezugssystems der Vp erfolgen oder - soweit ein solches nur unvollkommen ausgebildet oder interindividuell stark unterschiedlich ist - durch vergleichende Urteile vorgenommen werden sollen (vgl. Payne 1951, die ‚absolute' Beurteilung des Nährwertes von Milch führt vermutlich zu wenig brauchbaren Ergebnissen; es bietet sich an, Vergleiche mit anderen Nahrungsmitteln vornehmen zu lassen).

Damit eine Frage eindeutig ist, darf sie nur einen relevanten Gesichtspunkt (eine ‚Dimension') enthalten. Bei der Entwicklung der inhaltlichen Fragenkonzeption muß man also einerseits sicherstellen, daß der gewünschte Aspekt erfaßt ist, gleichzeitig müssen andere Aspekte ausgeschlossen sein. Das berühmte Beispiel einer vieldimensionalen Frage von Lazarsfeld (1935): ‚Warum haben Sie dieses Buch gekauft?' (Dimensionen können u.a. sein: ‚Sie' vs. andere Menschen, ‚dieses' vs. andere Bücher, ‚Buch' vs. andere Gegenstände, ‚gekauft' vs. andere Formen des Erwerbs) ist eine von der Konzeption her unangemessene Frage, da zwar die interessierende Dimension enthalten ist, andere aber nicht ausgeschlossen wurden. Neben diesen ersten allgemeinen Überlegungen zur Fragenkonzeption sind mehrere spezielle Gesichtspunkte zu berücksichtigen. Sie werden in den nachfolgenden Abschnitten behandelt.

3.1.2 Definition des Gegenstandes und Explikation eines Bezugsrahmens

In aller Regel kann nicht davon ausgegangen werden, daß die Vpn über eine einheitliche und mit der des Untersuchers übereinstimmende Vorstellung vom Befragungsgegenstand verfügen. So berichtet Noelle-Neumann (1970), daß die Frage nach dem Besitz bzw. der Verwendung einer ‚Perücke' von 1% der Befragten bejaht wurde, wenn gleichzeitig auch nach einem ‚Haarteil' gefragt wurde (15% Ja-Antworten). Wurde dagegen ohne weitere Unterscheidung nur nach einer ‚Perücke' gefragt, antworteten 8% der Befragten mit ‚Ja'. Dies läßt sich wohl nur so erklären, daß für einen Teil der Befragten auch Haarteile zu ‚Perücken' gehören, für einen anderen Teil dagegen nicht. Es ist also erforderlich, entweder eine möglichst exakte Definition für den Befragungsgegenstand vorzugeben oder aber - was im Sinne der Untersuchungsfragestellung ebenfalls interessant sein kann - die bei der Antwort zugrunde gelegte Definition von den Vpn zu erfragen (Cannell & Kahn 1968). Entsprechendes gilt für das Bezugssystem von dem ausgehend die Vpn ihre Antworten formulieren und das interindividuell sehr unterschiedlich sein kann (sicherlich haben Texaner und Bewohner Alaskas unterschiedliche Vorstellungen davon, was ein ‚warmer Sommer' ist, Cannell & Kahn 1968). Dabei muß das Bezugssystem - sofern es vorgegeben und nicht erfragt wird - für die Befragten relevant sein: Würde man eine Hausfrau nach dem jährlichen Eierverbrauch befragen, würde ihre Antwort reines Raten sein, der relevante Zeitraum hier z.B. ist die Woche

(Payne 1951). Darüber hinaus ist es u.U. erforderlich, eine Skala (Absolutwerte, Prozentangaben) für die erfragten Quantitäten zu spezifizieren und die erforderliche Genauigkeit festzulegen, um die üblicherweise beobachteten Antworthäufungen bei runden Zahlen zu vermeiden (Payne 1951). Besonderes Gewicht muß bei offenen Fragen auf die Definition von Befragungsgegenstand und Bezugssystem gelegt werden, da diese Fragen nicht durch explizite Antwortvorgaben eine weitere Einengung und Festlegung erfahren (Anger 1969).

3.1.3 Festlegung der Antwortkategorien

Soll eine Frage kategorie-neutral (Stroschein 1965, vgl. auch 2.4.1) sein, d.h. nicht schon durch die Vorgabe der Antwortkategorien bestimmte Antworten begünstigen (zu sogenannten ‚verzerrten' Fragen - Friedrichs 1973 - und ihrer legitimen Verwendung vgl. 3.1.4), so müssen die Vorgaben erschöpfend und - falls Mehrfachnennungen nicht erlaubt sind - disjunkt sein (Beispiele für Probleme bei nicht erschöpfenden Vorgaben finden sich bei Payne 1951, 87). Im Text der Frage müssen alle Vorgaben genannt werden oder es darf keine Vorgabe enthalten sein (Noelle-Neumann 1970). Zulässige Abweichungen von diesem Grundsatz beschreiben Kreutz & Titscher (1974).

Ausgewogen sind Vorgaben dann, wenn sie zu gleichen Teilen und mit gleicher Gewichtigkeit ‚positive' und ‚negative' Äußerungen dem jeweiligen Sachverhalt gegenüber beinhalten (Verstöße gegen diese Forderung und ihre Auswirkungen auf die Antworten beschreiben z.B. Payne 1951 und Rugg & Cantril 1972). Bei der Formulierung der Antwortkategorien ist auf möglichst vergleichbare soziale Wünschbarkeit zu achten (Phillips 1966) und zu bedenken, daß sich die ‚Attraktivität' erheblich durch den Aufweis von Konsequenzen der Antwort beeinflussen läßt (z.B. erfahrt eine vorgeschlagene Rentenerhöhung erheblich weniger Zustimmung, wenn die entsprechende Antwortvorgabe auch die Konsequenz einer Erhöhung der Beiträge zur Rentenversicherung deutlich macht, vgl. Karmasin & Karmasin 1977). Dies gilt verstärkt, wenn diese Konsequenzen personalisiert werden (‚. . . wenn Sie dafür einen höheren Beitrag zur Rentenversicherung zahlen müßten . . .' vs., . . . wenn dadurch die Beiträge zur Rentenversicherung steigen würden . . .'), wie überhaupt die Personalisierung von Fragen deutliche Veränderungen im Antwortverhalten zur Folge haben kann (Rugg & Cantril 1972).

Daneben spielt aber auch die ‚Extremheit' der verwendeten Vorgaben eine erhebliche Rolle: Karmasin & Karmasin (1977) zeigen am Beispiel zweier Befragungen zur gesetzlichen Regelung des Schwangerschaftsabbruchs, wie durch Hinzufügung einer extremeren Antwortkategorie (völlige Freigabe) die Befürwortung der Fristenlösung erheblich zunimmt. Beispiele für Antwortverzerrungen durch Gegenüberstellung extremer und gemäßigter Antwortka-

tegorien beschreiben Kreutz & Titscher (1974) und Payne (1951). Besondere Vorsicht ist geboten, wenn extreme Kategorien in ihrer Formulierung Existenzaussagen (‚Es gibt . . .') oder Allaussagen (‚Alle . . .', ‚immer . . .', ‚nie . . .') nahekommen (Payne 1951).

Mit der häufig beobachteten Tendenz der Befragten, extreme Antwortkategorien zu meiden (d.h. sie eher zu wählen, wenn noch extremere vorgegeben sind), hängt es auch zusammen, daß nachträgliche Kombinationen von Antwortkategorien (z.B. ‚sehr dafür' und ‚dafür') fast immer andere Ergebnisse liefern als Befragungen, in denen von vornherein eine zusammengefaßte Kategorie (‚sehr dafür oder dafür') vorgegeben war. Welche Ergebnisse die ‚richtigen' sind, ist in der Regel natürlich nicht entscheidbar (Payne 1951).

Sind bei Wissensfragen die richtigen Antworten in den Vorgaben enthalten, so ergeben sich selbst wenn sie nur dem Interviewer für Zwecke der Feldverschlüsselung vorliegen, größere Häufigkeiten richtiger Antworten als im Falle offener Fragen (Noelle-Neumann 1970).

Da nicht davon auszugehen ist, daß jede Vp zu jeder Frage eine Antwort geben kann, ist es einerseits zur Vermeidung artifizieller (zufälliger) Wahl von Antworten erforderlich, Restkategorien vorzusehen, andererseits ermöglichen solche Kategorien den Vpn ein ‚Ausweichen' (und provozieren es u.U. sogar), so daß in der Praxis häufig darauf verzichtet wird (Rugg & Cantril 1972, Kirschhofer-Bozenhardt & Kaplitza 1975).

Tatsächlich müßten zur Abdeckung aller denkbaren Fälle sogar mehrere Ausweichkategorien vorgesehen werden (Karmasin & Karmasin 1977): Galtung (1973) unterscheidet zwischen kognitiven (‚weiß nicht') und evaluativen (‚interessiert mich nicht') Nicht-Antworten, dazu müßte man noch berücksichtigen, daß ‚etwas anderes' oder ‚mehreres' für die Vp richtig sein oder sie die Frage nicht verstanden haben kann (‚nicht verstanden'). Die Hinzufügung von Nicht-Antwortkategorien zu einem Satz von Antwortvorgaben führt u.U. zu beträchtlichen Wahlhäufigkeiten für diese und entsprechenden Veränderungen für die anderen Kategorien, wobei nur im Einzelfall geklärt werden kann, ob dadurch wahre Varianz (Bequemlichkeitshypothese) oder Fehlervarianz (Hypothese des Zufallscharakters erzwungener Antworten) von den inhaltlichen Kategorien abgezogen wird.

Zahlreiche Untersuchungen zeigen auf, daß die Reihenfolge der Antwortalternativen im Fragentext in einer Listenvorgabe einen Einfluß auf die Wahlhäufigkeiten ausübt. Dabei variieren die Angaben vor allem über das genaue Ausmaß solcher ‚Positionseffekte' von Untersucher zu Untersucher beträchtlich. Wegen der zahlreichen Interaktionen mit inhaltlichen und formalen Aspekten der Fragen ist auch nicht damit zu rechnen, daß allgemein gültige Aussagen möglich sind (Kreutz & Titscher 1974). Payne (1951) spricht von Tendenzen

dahingehend, daß bei mündlichen qualitativen Vorgaben die zuletzt genannten, bei schriftlichen qualitativen Kategorien diejenigen in Extrempositionen (erste und letzte Stelle) relativ inhaltsunabhängig bevorzugt werden. Bei quantitativen Vorgaben gibt es relativ unabhängig von ihrer Höhe eine Neigung, solche in der Nähe des Mittelwertes zu wählen. In diese Richtung gehen auch die Befunde von Stroschein (1964), Belson (1966) und Ring (1974).

Zur Vermeidung systematischer Auswirkungen solcher Positionseffekte hat es sich eingebürgert, sogenannte gegabelte Befragungen (split-ballot, vgl. 3.4) durchzuführen und dabei die Reihenfolge von Vorgaben zu variieren. Wegen der Benachteiligung der Mittelpositionen ist dabei ein einfaches ‚Umdrehen' meist nicht ausreichend (‚verfeinerte' Techniken beschreibt Ring 1974). Dieses Verfahren findet allerdings dort seine Grenzen, wo die Frage unnatürlich zu wirken beginnt (Payne 1951) und befürchtet werden muß, daß dadurch zusätzliche Störeinflüsse wirksam werden (z.B. ‚Gehen Sie heute abend nicht ins Theater oder gehen Sie ins Theater?'). Da Positionseffekte sich dann nicht zeigen, wenn über jede Vorgabe durch Einzelantwort entschieden werden muß (Stroschein 1965), bietet sich dieser Fragentyp für Fälle an, in denen die Variation der Vorgabenreihenfolge nicht möglich ist. Neben der Ausgewogenheit der Kategorien hat auch die Ausgewogenheit der mit den jeweiligen Antworten verbundenen Folgefragen einen erheblichen Einfluß auf die Wahlhäufigkeit (Noelle-Neumann 1970). Vor allem scheinen Befragte (im mündlichen Interview evtl. auch Interviewer) rasch die Vermeidung bestimmter Antworten (z.B. ‚Ja') zu lernen, wenn diese regelmäßig mit einer höheren Zahl von Folgefragen verknüpft sind (Cannell & Kahn 1968).

3.1.4 Verzerrte Fragen

Verzerrte Fragen sind nach Friedrichs (1973, 198) solche, „. . . die allein durch ihre Formulierung die Verteilung der Antworten in einer bestimmten Form beeinflussen . . .". Solche Verzerrungen können einmal durch inadäquate Antwortkategorien (vgl. dazu 3.1.3) entstehen, daneben gibt es weitere Faktoren, für die verzerrende Wirkungen auf Antworten aufgezeigt worden sind.

a. *Unterstellungen* (Implikationen) führen, soweit sie zu Unrecht bestehen, häufig nicht dazu, daß sie von den Vpn zurückgewiesen werden, sondern verzerren die Antworten. Das gilt für die unterstellte Vertrautheit mit Sachverhalten und Begriffen (Payne 1951, vgl. dazu speziell 3.1.5) ebenso wie für unterstellte Voraussetzungen (‚Welcher Teil Ihrer Arbeit stört Sie am meisten?' - wer sagt, daß einer stört?) und unterstellte Konsequenzen eines zu erfragenden Sachverhaltes (‚Welcher Teil Ihrer Arbeit stört Sie am meisten, d.h. welchen schieben Sie am längsten auf?'). Letztere entstehen häufig versehentlich beim Versuch einer Konkretisierung des Befragungsgegenstandes (Payne 1951).

Aber auch wenn Unterstellungen nicht verzerrend wirken, sondern von Befragten erkannt und zurückgewiesen werden (z.B. ‚Wieviele Zigaretten rauchen Sie pro Tag?'), stellen sie einen befragungstechnischen Fehler dar. In bestimmten Fällen, z.B. um bestimmten peinlichen oder sozial unerwünschten Verhaltensweisen den Charakter des Selbstverständlichen zu verleihen (vgl. Kinsey et al. 1970),können Unterstellungen als bewußt eingesetztes methodisches Hilfsmittel gerechtfertigt sein. Auf die ansonsten erforderliche Vorschaltung einer Filterfrage wird dann zurecht verzichtet. Phillips (1966) erwähnt als Beispiel die Frage nach finanziellen Belastungen aus Ratenkäufen. Statt der üblichen Fragenfolge ‚Haben Sie regelmäßige Zahlungsverpflichtungen aus Ratenkäufen?' (= Filter), ‚wenn ja: Wie hoch sind diese pro Monat?', wäre es hier zweckmäßig, zur Verminderung des Einflusses der sozialen Erwünschtheit mit einer Unterstellung zu arbeiten und zu fragen: ‚Wie hoch sind Ihre Zahlungsverpflichtungen aus Ratenkäufen pro Monat?'.

b. Die Verknüpfung bestimmter Befragungsgegenstände mit *Persönlichkeiten, wichtigen Ereignissen* o.ä. hat - wie in verschiedenen Untersuchungen nachgewiesen - einen erheblichen Einfluß auf die Antworten (vgl. auch 2.3). So berichten Roslow et al. (1940) und Rugg & Cantril (1972) über Auswirkungen, die die Erwähnung des ‚Präsidenten' bzw. des ‚Kongresses' in Fragen zu aktuellen politischen Problemen zeigten. Dabei sind diese Auswirkungen allerdings spezifisch, d.h. sie zeigten sich nicht durchgängig bei beliebigen Befragungsgegenständen (Rugg & Cantril 1972). Mit erheblich verändertem Antwortverhalten wäre z.B. auch zu rechnen, würde man eine Frage über Sicherheitsmaßnahmen von Fluggesellschaften an einer aktuellen Flugzeugentführung ‚festmachen' (Karmasin & Karmasin 1977).

c. Affektiv getönte Begriffe vermögen Antwortverteilungen deutlich zu beeinflussen. So berichten Rugg (1941) und Rugg & Cantril (1972) niedrigere Zustimmungsraten für ‚verbieten' vs. ‚nicht erlauben' (im Zusammenhang mit ‚öffentlichen Reden gegen die demokratische Ordnung') und für ‚den Krieg erklären' vs. ‚in den Krieg eintreten' (im Zusammenhang mit dem Eintritt der USA in den Zweiten Weltkrieg). Besonders auffallend ist die Wirkung des Begriffes ‚Veränderung': Die Frage nach einer *Ergänzung* der Verfassung um eine bestimmte Vorschrift (‚hinzufügen') erhielt 36% Zustimmungen und 50% Ablehnungen, die nach einer (inhaltlich identischen) Verfassungsänderung aber nur 26% Zustimmungen und 65% Ablehnungen (Rugg & Cantril 1972, 106). Dieser und ähnliche Befunde veranlaßten Payne (1951, 183) prinzipiell alle Fragen, die explizit entweder auf den ‚Status-Quo' oder auf ‚Veränderungen' (oder auf beides) hinweisen, schon allein deshalb für verzerrt zu halten.

Daneben kann natürlich fast jeder Begriff in einem bestimmten gegebenen Befragungszusammenhang affektiv getönt sein. Falls entsprechende Befürchtungen begründet sind, empfiehlt sich im Rahmen des Pretests eine entsprechende Untersuchung z.B. unter Verwendung der Methode des se-

mantischen Differentials oder durch Bestimmung von Assoziationen (vgl. Kreutz & Tischer 1974).

d. Verzerrt kann eine Frage schließlich auch dadurch sein, daß sie durch ihren Aufbau bestimmte Antworten deutlich begünstigt (‚Lehnen Sie es ab . . .', ‚Sie sind doch auch der Meinung . . .' etc.).

Litwak (1956) hat darauf hingewiesen, daß die Verzerrtheit einer Frage nicht nur von Merkmalen dieser Frage, sondern auch von ihrem Verwendungszweck abhängt: Was im Rahmen einer demoskopischen Befragung eine ‚unerlaubte' verzerrte Frage wäre, kann innerhalb einer Einstellungsskala ein zulässiges, sogar erforderliches extremes Item sein. Auch in anderen Zusammenhängen können verzerrte Fragen (bewußt eingesetzt und die Ergebnisse entsprechend interpretiert) zu bemerkenswerteren und gültigeren Erkenntnissen führen, als ausgewogene Fragen dies tun würden (vgl. a.a.O. und Anger 1969, Kreutz & Titscher 1974). So ist es denkbar, daß bezüglich bestimmter Sachverhalte (z.B. Kernkraft) weniger die evtl. stark von Medien beeinflußten und u.U. wenig stabilen Reaktionen des ‚Durchschnittsbürgers' auf ausgewogene Fragen und eher die Sichtweisen eines durch verzerrte Fragen herausgefilterten ‚harten Kerns' von Gegnern und Befürwortern interessieren.

Außerdem ist es möglich, verzerrte Fragen bzw. Fragebogen mit verzerrten Fragen nicht zum Zwecke der Informationsgewinnung, sondern mit dem Ziel der Beeinflussung im Sinne einer Einstellungsänderung einzusetzen. Über frühe derartige Versuche (Beeinflussung interventionistischer vs. isolationistischer und gewerkschaftsfreundlicher vs. gewerkschaftsfeindlicher Einstellungen) von Roper berichten Rugg & Cantril (1972). Einstellungsänderungen ließen sich vor allem bezüglich solcher Sachverhalte erzielen, denen gegenüber die Vpn verhältnismäßig unsicher waren. Dillehay & Jernigan (1970) konnten durch einen verzerrten Fragebogen zur Behandlung von Straftätern nur Einstellungsänderungen in Richtung auf mildere, nicht aber solche in Richtung auf härtere Bestrafung erzielen. Selbstverständlich darf dabei die Verzerrung der Fragen nicht soweit gehen, daß sie von den Vpn als Beeinflussungsversuch erkannt wird (sonst wäre durchaus auch mit Bumerangeffekten zu rechnen).

Diese als Beispiele für den absichtlichen Einsatz verzerrter Fragen erwähnten Untersuchungen stellen erneut die Beeinflußbarkeit des Antwortverhaltens der Vpn durch konstruktive Merkmale der Frage unter Beweis, machen andererseits aber auch deutlich, daß dieses Problem vor allem bei unsicheren Beurteilungsgrundlagen, Einstellungen oder Meinungen besteht und daß Merkmale der Fragen an Bedeutung verlieren, wenn die zu erfragenden Inhalte deutlich ausgeprägt, stabil bzw. intensiv sind (vgl. dazu auch 3.4).

3.1.5 Uninformiertheit, Meinungslosigkeit und Urteilsausgewogenheit

Die Tatsache, daß Vpn eine Frage beantworten, kann nicht als Hinweis darauf interpretiert werden, daß sie über den erfragten Sachverhalt informiert sind oder eine Meinung dazu haben. So berichtet Payne (1951, 156) über Nonsense-Fragen (z.B. Beurteilung eines nicht existenten Gesetzentwurfs), die gleichwohl von erheblichen Anteilen der Vpn ‚beantwortet' wurden. Ein besonders eindrucksvolles Beispiel beschreibt Eysenck (1956, 156). Bei einer Umfrage in Großbritannien kurz nach dem Ende des Zweiten Weltkriegs wurde die Frage gestellt, ob man ‚König Georg von Griechenland' wieder in sein Land zurückkehren lassen sollte. 60% der Befragten bejahten diese Frage. In einer etwa gleichzeitig durchgeführten anderen Befragung gab jedoch nur ein kleiner Bruchteil der Befragten an, schon einmal etwas von König Georg von Griechenland gehört zu haben. Offenbar neigen also Befragte dazu, ihre Uninformiertheit nicht zu offenbaren (und Fragebogenkonstrukteure geben ihnen häufig auch gar keine Möglichkeit, dies zu tun). Einer Frage z.B. nach der Beurteilung eines Sachverhaltes muß deshalb entweder eine Filterfrage nach der Informiertheit, eine Unterweisungsfrage (vgl. 2.1) oder eine Erklärung vorangehen, wobei nach Noelle-Neumann (1974) die Unterweisungsfrage („... wissen Sie davon?') im Vergleich zur Erklärung der effizientere Weg ist, da sie eine aktive Auseinandersetzung mit der Information (eine Antwort) erfordert.

In jedem Fall ist dabei auf das Bedürfnis der Vpn, informiert zu erscheinen, Rücksicht zu nehmen, d.h. ein Bloßstellen der uninformierten Vpn muß vermieden werden.

Dies kann z.B. wiederum dadurch geschehen, daß in einer Filterfrage der Uninformiertheit der Charakter des Selbstverständlichen verliehen wird (vgl. Maccoby & Maccoby 1972). Statt ‚Wissen Sie, welche Länder Mitglieder der EG sind oder wissen Sie das nicht?' wäre eine Formulierung ‚Wissen Sie vielleicht, welche Länder Mitglieder der EG sind?' vorzuziehen (Karmasin & Karmasin 1977, vgl. auch Phillips 1966). Eine Erklärung im Rahmen einer Frage sollte aus den selben Gründen nicht belehrend wirken (‚Unter EG versteht man die Europäische Gemeinschaft . . .'), sondern entweder die Vermutung ihrer Entbehrlichkeit zum Ausdruck bringen (‚Wie Sie vermutlich wissen, ist die EG . . .') oder aber als Präzisierung des Befragungsgegenstandes in Erscheinung treten (‚Was halten Sie von der Europäischen Gemeinschaft, also der EG?').

Neben der gegebenen, aber durch inadäquate Fragenkonstruktion unerkannt bleibenden Uninformiertheit von Vpn ist die verbreitete Unterstellung, Befragte würden zu ausnahmslos allen Befragungsgegenständen eine Meinung haben, eine häufige Ursache für inadäquate Antworten (Kreutz & Titscher 1974). Auch hier muß die Konstruktion der Frage es der Vp in geeigneter

Weise ermöglichen, das Nichtvorhandensein einer Meinung zum Ausdruck zu bringen (zur Vorgabe entsprechender Antwortkategorien vgl. 3.1.3).

Empirische Belege sprechen außerdem dafür, daß Vpn deutlich negative Urteile über Befragungsgegenstände scheuen (vgl. Roslow et al. 1940, Rugg 1941, Phillips 1966, Kirschhofer-Bozenhardt & Kaplitza 1975), auch wenn sich das nicht durchgängig demonstrieren läßt (z.B. die Formulierung ‚X ist besser als Y' nicht generell der logisch äquivalenten ‚Y ist schlechter als X' vorgezogen wird, vgl. Adams 1956), und eher zu ausgewogenen Beurteilungen neigen. Es ist deshalb erforderlich, Befragte nicht ausschließlich zur Kritik an Befragungsgegenständen zu zwingen, sondern Gelegenheit zur Hervorhebung positiver Aspekte zu geben, auch wenn solche inhaltlich gar nicht interessieren sollten (Phillips 1966; Noelle-Neumann 1974 spricht dann von ‚Wegwerf-Fragen').

3.1.6 Antworttendenzen und vorschnelle Antworten

Bei Sachverhalten, bei denen eine wahrheitsgemäße Antwort nach Meinung der Vp gleichzeitig eine sozial unerwünschte Antwort wäre, muß durch geeignete Fragenkonstruktion der Vp z.B. die Möglichkeit gegeben werden, sich für die Antwort zu ‚entschuldigen'. Statt zu fragen ‚Besitzen Sie ein Auto oder besitzen sie ein solches nicht?' und damit der Vp evtl. das Eingeständnis ihrer ‚Armut' abzuverlangen, könnte man an folgende Konzeption denken: ‚Besitzen Sie ein Auto oder ist das für Sie im Augenblick nicht möglich oder wünschenswert?' (vgl. dazu auch Phillips 1966).

Um response sets wie der Bejahungstendenz entgegenzuwirken und außerdem die Vpn zu sorgfältiger Beantwortung der Fragen zu veranlassen, wird meist empfohlen, Items teilweise positiv (‚Sind Sie an Sport interessiert?'), teilweise negativ (‚Sind Sie an Sport uninteressiert?') zu formulieren. Allerdings ist eine solche Vorgehensweise nicht unproblematisch. Zum einen ergibt sich durch die Antwort ‚Nein' auf ein negativ formuliertes Item die Situation der doppelten Verneinung, die stets als Fehlerquelle anzusehen ist (vgl. 3.2.2). Zum andern haben Terborg & Peters (1974) gezeigt, daß die Veränderung der Formulierungsrichtung für viele Items signifikante Auswirkungen auf die Häufigkeit der Wahl von Antworten hat, d.h. die Antwort ‚Ja' auf ein positiv formuliertes Item nur logisch äquivalent der Antwort ‚Nein' auf ein negativ formuliertes (und umgekehrt) ist. Diese Beantwortungsunterschiede konnten zudem nicht ausschließlich der unterschiedlichen Wirksamkeit von Antworttendenzen ('Ja-Tendenz oder Nein-Tendenz) angelastet werden, da je nach Item Ja-Antworten bei positiver Formulierung häufiger, z.T. aber auch seltener auftraten als Nein-Antworten bei negativer Formulierung (und umgekehrt). Karmasin & Karmasin (1977) schlagen deshalb vor, auf Ja-Nein-Fragen möglichst

zu verzichten und stattdessen die Alternativen explizit zu formulieren, wobei deren Reihenfolge im Rahmen einer gegabelten Befragung (vgl. 3.4) variiert werden kann (die o.a. Frage würde dann lauten: ‚Interessieren Sie sich für Sport oder sind Sie an Sport uninteressiert?'). Prinzipiell ähnliche Überlegungen für den Fall von Persönlichkeitsfragebogen finden sich bei Ehlers (1973) und Keil (1973).

Einer zweiten Tendenz im Beantwortungsverhalten sollte ebenfalls schon durch die Konstruktion der Frage entgegengewirkt werden: der Neigung zu impliziter oder expliziter Formulierung der Antwort, bevor die gesamte Frage von der Vp zur Kenntnis genommen wurde (Tendenz zu vorschneller Antwort). Das kann z.B. dadurch erfolgen, daß die eigentliche Frage erst ganz zum Schluß, d.h. nach der genauen Definition des Gegenstandes, der Explikation des Bezugsrahmens etc. verbalisiert wird (Payne 1951). Dem zugegebenermaßen konstruierten Aufbau einer Frage

- ‚Würden Sie sagen, der Preis für Benzin ist zu hoch (*), gerade richtig oder zu niedrig (*), wenn Sie ihn mit Preisen anderer Dinge vergleichen?',

der mindestens an den mit (*) bezeichneten Stellen vorschnelle Antworten ermöglicht, wäre ein Aufbau wie etwa der folgende vorzuziehen:

- ‚Verglichen mit den Preisen anderer Dinge: Würden Sie sagen, der Preis für Benzin ist zu hoch, gerade richtig oder zu niedrig?'.

3.2 Sprachliche Formulierung der Frage

3.2.1 Kriterien für die sprachliche Formulierung

Die sprachliche Formulierung einer Frage erfolgt einerseits mit dem Ziel, den Befragten zu einer Antwort zu motivieren, andererseits muß sie erreichen, daß die Frage von der Vp richtig verstanden wird (Anger 1969). Übereinstimmend wird die sprachliche Formulierung einer Frage als ein Problem der Optimierung unter dem Kriterium der Bedeutungsäquivalenz für alle Befragten angesehen. Da die Bedeutung eines Begriffes außer von seiner Denotation (definiertem Inhalt) und den interindividuell unterschiedlichen Konnotationen (dem Bedeutungshof, der den Ort in einem semantischen Raum bestimmt, vgl. Osgood et al. 1957) auch noch von gruppenspezifischen Bedeutungsanteilen (Altersgruppen, Schichten, regionalen Gruppierungen; Karmasin & Karmasin 1977) und der Verwendung in unterschiedlichen Lebensbereichen (Arbeitswelt, Privatleben etc.; vgl. Scheuch 1973) geprägt wird, ist es grundsätzlich ausgeschlossen, das absolute Optimum der Bedeutungsäquivalenz (inhaltliche Standardisierung) durch identische sprachliche Formulierung für alle Vpn (formale Standardisierung im Sinne von Stroschein 1965) zu erreichen. Vielmehr würde dazu eine für jede Vp unterschiedliche sprachliche Formulierung erforderlich sein. Ausgehend von der Erfahrung, daß in der alltäglichen Kommuni-

kation wechselseitiges Verstehen mehr oder weniger möglich ist, wird im freien mündlichen Interview die Aufgabe, eine inhaltliche Standardisierung in dem genannten Sinne zu erreichen, der Intuition des Interviewers übertragen. Der dagegen erhobene Einwand, daß hierbei (von einzelnen hochqualifizierten und ‚begabten' Interviewern vielleicht abgesehen) zu der Störvariable ‚Verständnisunterschiede durch die Vpn' nur noch weitere hinzugefügt würden (wie etwa die Ausdrucksfähigkeit des Interviewers und seine Vorstellung von dem, was die Vp wie zu verstehen habe: Wottawa 1980), läßt sich jedoch kaum entkräften.

Nach dem Prinzip der ,,maximalen Übelminimierung" (Wottawa 1980, 209) werden bei der formalen Standardisierung (Identität der Fragen auf verbaler Ebene) die interindividuellen sprachlichen Unterschiede vernachlässigt. Ziel bleibt auch hier die Bedeutungsäquivalenz, nur liegt diesem Vorgehen die Annahme zugrunde, daß diese durch identische sprachliche Formulierungen besser als durch unkontrollierte freie Formulierungen zu erreichen sei (Mayntz et al. 1971, Wottawa 1980). Dies wird in der Regel auch für Instruktionen im Rahmen von Leistungstests oder von psychologischen Experimenten angenommen. Da es andererseits offensichtlich unsinnig ist, eine Frage beantworten zu lassen, die nicht verstanden wurde, ergibt sich fast zwangsläufig die Forderung, sich bei der sprachlichen Formulierung an der untersten Grenze der Zielgruppe zu orientieren (Payne 1951). Allerdings genügen solche Formulierungen mindestens für sprachlich differenziertere Vpn nicht mehr dem Kriterium der Motivierung von Antworten (Erdos 1970), da durch Übersimplifizierungen Zweifel an der Seriosität der Befragung ausgelöst werden können (Kreutz & Titscher 1974). Deshalb wird als Kompromiß heute eher eine Orientierung an der Alltagssprache (Umgangssprache) des durchschnittlichen Mitgliedes der Zielpopulation vorgeschlagen (Karmasin & Karmasin 1977).

Teilweise wird versucht, gruppenspezifische Bedeutungsunterschiede durch unterschiedliche sprachliche Formulierungen zu berücksichtigen. Dies gilt vor allem für regionale Unterschiede. Noelle-Neumann (1963, 1974) etwa schlägt vor, den Interviewer durch eine sogenannte informelle Ermittlung (eine Frage ohne festgelegten Wortlaut, z.B. nach der Gebräuchlichkeit der Bezeichnungen ‚Samstag' oder ‚Sonnabend') die Zugehörigkeit des Befragten zu einem bestimmten Sprachraum feststellen zu lassen und in Abhängigkeit davon zu verschieden formulierten Fragen zu verzweigen. Im übrigen muß aber in der Regel (mit Karmasin & Karmasin 1977, 176) festgestellt werden: ,,. . . über die Bedeutungsverschiebungen von einzelnen Begriffen bzw. über die jeweils relevanten wörtlichen Bezeichnungen von Sachverhalten in den üblichen Sprachrepertoires von Jugendlichen gegenüber Erwachsenen, Männern gegenüber Frauen, Unterschicht gegenüber Oberschicht ist jedoch im Augenblick aus dem deutschen Sprachraum noch zu wenig bekannt, so daß auch hier nur der Ausweg bleibt, wörtliche Äquivalenz zu wahren und bei allen Begriffen und

Formulierungen, bei denen wechselnde Bezugsrahmen vermutet werden können, den Bezugsrahmen mit anzugeben, unter dem der Forscher den Begriff einzuordnen wünscht“ (vgl. auch Cannell & Kahn 1968). Mangels gesicherten Wissens über sprachliche Unterschiede ist häufig nicht ohne weiteres zu entscheiden, ob Beantwortungsunterschiede z.B. zwischen Angehörigen verschiedener Schichten auf unterschiedliches Fragenverständnis oder auf Unterschiede des erfragten Sachverhaltes zurückgehen (Kreutz & Titscher 1974). Mindestens empfiehlt es sich in solchen Fällen, einen Sachverhalt durch mehrere Fragen unterschiedlicher Formulierung zu erfassen bzw. Verständnis-Kontrollfragen einzubauen.

3.2.2 Anforderungen an die sprachliche Formulierung

Wegen der starken Kontextabhängigkeit und der Vielfalt der Interaktionen mit inhaltlichen Aspekten scheint es von vornherein verfehlt, nach besonders geeigneten Standardformulierungen zu suchen (Raab 1974). Stattdessen wird man die Argumente für oder gegen bestimmte Vorgehensweisen im Einzelfall gegeneinander abwägen müssen. Folgt man dem Prinzip, sich bei der sprachlichen Formulierung an der Umgangssprache des Durchschnitts der Zielpopulation zu orientieren (Karmasin & Karmasin 1977, vgl. auch 3.2.1), so muß man sich zunächst die hauptsächlichen Kennzeichen dieser Sprache vergegenwärtigen (eine entsprechende Zusammenstellung unter Berücksichtigung der Ergebnisse von Lesbarkeitsuntersuchungen findet sich - allerdings für den angloamerikanischen Sprachbereich - z.B. bei Wright & Barnard 1975).

Dazu gehört - mindestens im Falle von Wohnbevölkerungen als Zielgruppen - die Verwendung kurzer Wörter. Payne (1951) berichtet, daß Fragen, die unter dem Kriterium geringer Beeinflußbarkeit der Antworthäufigkeiten durch Variation der Vorgabenreihenfolge als ‚klar‘ klassifiziert worden waren, zu ca. 8%, ‚unklare‘ Fragen dagegen zu 12,5% zwei- oder mehrsilbige Wörter enthielten. Bei klaren Fragen waren 30%, bei unklaren 40% aller Silben Vor- oder Nachsilben. Sodann ist für die Alltagssprache der Gebrauch solcher Wörter charakteristisch, die in der Sprache häufig vorkommen. Dementsprechend schließt sich auch Friedrichs (1973) im Zusammenhang mit der Fragenformulierung der Empfehlung einer Beschränkung auf die 1000 gebräuchlichsten Wörter (mit Vorbehalten) an. Das Kriterium der Worthäufigkeiten ist indessen recht oberflächlich, da eigentlich relevanter die Gebräuchlichkeit des Wortes in einem gegebenen Zusammenhang ist. Fremdwörter und Abstrakta jedenfalls sollten soweit als möglich vermieden werden.

Weniger eine Frage der Orientierung an der Alltagssprache als eine Notwendigkeit im Hinblick auf die Approximation der Bedeutungsäquivalenz für alle Befragten ist die Notwendigkeit einer Beschränkung auf klare Begriffe, das sind solche, deren denotative Bedeutungen prägnant und die arm an konno-

tativen Bedeutungen sind (Rohrmann 1978), was gegebenenfalls durch Analysen z.B. unter Verwendung des semantischen Differentials nachzuprüfen wäre (Friedrichs 1973). Im Interesse dieser Klarheit empfiehlt es sich auch nicht, Synonyme im Wechsel zu verwenden (Anger 1969).

Zwar herrschen in der Alltagssprache personalisierte Formulierungen vor (Karmasin & Karmasin 1977), doch muß hier die Entscheidung in Abhängigkeit von der Fragestellung erfolgen, da mit spezifischen Einflüssen der Personalisierung einer Frage auf die Antworten zu rechnen ist (vgl. 3.1.3 und 3.3).

Die grammatikalische Satzkonstruktion der Alltagssprache ist durch geringe ‚Satztiefe' (geringen Komplexitätsgrad der syntaktischen Struktur, vgl. Karmasin & Karmasin 1977) ausgezeichnet. Ob man deshalb mit Kreutz & Titscher (1974) für Fragen eine Beschränkung auf Hauptsätze fordern muß, ist zweifelhaft. Immerhin sollten

- ungewöhnliche Tempora,
- komplizierte Nebensatzkonstruktionen (Schachtelsätze),
- adverbiale Konstruktionen und
- passivische Formulierungen

möglichst vermieden werden (Wright & Barnard 1975, Karmasin & Karmasin 1977). Dasselbe gilt für doppelte Verneinungen, die entweder im Fragentext selbst liegen (‚Sind Sie dagegen, daß der 17. Juni als Feiertag abgeschafft wird oder sind Sie nicht dagegen?') oder durch eine verneinende Antwort auf eine negativ formulierte Frage entstehen können (‚Soll der 17. Juni in Zukunft kein Feiertag mehr sein? ja/nein', vgl. auch 3.1.6). In vielen Fällen genügt es allerdings nicht, eine grammatikalisch richtige Fragenkonstruktion zu verwenden, zusätzlich muß auch sichergestellt sein, daß der Bezug der Antwort auf die Frage unmittelbar evident ist. So ist nach Payne (1951, 69) bei einer Frage des Typs ‚Ist Ihr Gesundheitszustand heute besser oder schlechter als vor einem Jahr?' trotz grammatikalischer Eindeutigkeit vielen Vpn nicht klar, ob sich die Antwort ‚besser' auf ‚heute' oder auf das vergangene Jahr bezieht. In solchen Fällen ist es unabdingbar, daß die Alternativen explizit formuliert werden (‚Ist Ihr Gesundheitszustand heute besser oder war er vor einem Jahr besser?').

Vielfach findet sich in der Literatur die Empfehlung, Fragen möglichst kurz zu fassen (z.B. Holm 1974b, Kreutz & Titscher 1974, Wright & Barnard 1975, Karmasin & Karmasin 1977). Oppenheim (1966) empfiehlt 20 Wörter als Obergrenze, Payne (1951) berichtet für die nach seinem Kriterium ‚klaren' Fragen (S.O.) eine durchschnittliche Länge von 22, für ‚unklare' eine von 31 Wörtern. Schneider-Düker & Schneider (1977) fanden bei ihren Versuchen zur freien Reproduktion von Fragebogenitems Korrelationen von 0,54 bzw. 0,71 zwischen Anzahl von Intrusionen (Umformungen bzw. Einfügungen von Wörtern) und Itemlänge (Wortanzahl bzw. Silbenanzahl).

Für offene Fragen und Listenfragen mit Mehrfachnennungen in mündlichen standardisierten Interviews widersprechen Cannell et al. (1977) der Forderung nach möglichst kurzen Items. Sie stellen der dieser Forderung zugrundeliegenden Hypothese von der ‚Verwirrung' der Vp durch eine lange Frage eine Hypothese der ‚Vorbildwirkung' des Interviewers gegenüber und vermuten, daß die Vp ihr Engagement und ihre Ausführlichkeit bei der Beantwortung an den cues orientiert, die sie aus dem Verhalten des Interviewers entnimmt. Beim Vergleich der Antworten auf Items in Kurzform (durchschnittlich 14 Wörter) und Langform (= Kurzform + Redundanz, durchschnittlich 38 Wörter) zeigte sich zwar kein Unterschied in den Antwortlängen, dafür enthielten freie Antworten und Antworten in Listenfragen im Falle der Langform mehr Information. Außerdem waren auch die Antworten auf kurze Fragen informationshaltiger, wenn der Fragebogen teils kurze, teils lange Fragen enthielt. Soweit es in den Fragen um Reproduktion von Gedächtnisinhalten geht, ist auch zu bedenken, daß eine lange Frage der Vp mehr Zeit läßt und u.U. relevante cues mehrfach wiederholt darbietet (Cannell et al. 1977). Allerdings wirkt sich die Fragenlänge möglicherweise nicht auf alle Vpn gleichmäßig aus. Koomen & Dijkstra (1975) z.B. fanden (anders als Cannell et al. 1977) einen Anstieg der Antwortlängen in Abhängigkeit von der Länge der Fragen, allerdings nur für solche Vpn, die bei kurzen Fragen zu ausgesprochen kurzen Antworten neigten (vgl. auch Sudman & Bradburn 1974).

Auch zur Fragenlänge lassen sich keine unbeschränkt gültigen Aussagen machen. Sie ist unter Berücksichtigung von Frageninhalt, Untersuchungsziel und Verständlichkeit im Einzelfall zu optimieren.

3.3 Spezielle Gesichtspunkte der Formulierung von Items für diagnostische Fragebogen

Für diagnostische Fragebogen, die auf der Grundlage eines streng empirischen Validitätskonzeptes erstellt wurden, ist die Kriteriumskorrelation der Prüfstein für die adäquate Formulierung eines Items. Theoretisch ist es hier sogar zulässig, vollkommen unverständliche Items zu konstruieren, solange nur die ‚Art des Unverständnisses' (z.B. der dann die Antwort determinierende response set, vgl. 1.2.1 und 1.2.2) gültige Vorhersagen ermöglicht.

Wird von inhaltlichen Überlegungen ausgegangen, so muß sich letztlich in Itemanalysen (entweder im Rahmen eines klassischen oder eines probabilistischen Meßmodelles) die inhaltliche und formale Brauchbarkeit der Items erweisen. Die Wahrscheinlichkeit dafür, daß dies gelingt, dürfte bei Berücksichtigung der in diesem Kapitel aufgezeigten Zusammenhänge erhöht sein.

Im Hinblick auf die Motivation der Probanden müssen mehr noch als bei demoskopischen Fragebogen Überlegungen zur face-validity der Items ange-

stellt werden. Auch bei den demoskopischen Fragebogen spielt dieser Aspekt eine Rolle: Nicht jedem Untersucher wird von den Befragten jede Frage ‚zugestanden' (vgl. Richter 1969). Für rein empirische Fragebogenkonstruktionen ist face-validity im Interesse der Undurchschaubarkeit der Items geradezu unerwünscht und oft auch tatsächlich nicht gegeben, was solche Verfahren bei Betroffenen und in der Öffentlichkeit häufig in Mißkredit bringt. Empfehlungen für die Formulierung von Items in diagnostischen Fragebogen und von Aufgaben in Leistungstests finden sich in der entsprechenden Literatur, z.B. bei Lienert (1969, 62ff) und Wottawa (1980, 212ff).

3.4 Die Kontrolle von Formulierungseinflüssen

Während im Zusammenhang mit Einstellungsskalen (vgl. z.B. Suchman & Guttman 1947) und diagnostischen Verfahren (z.B. soweit sie auf der Basis eines probabilistischen Meßmodelles, vgl. Fischer 1974, Wottawa 1980, konstruiert sind) auch grundlegend andere Strategien verfolgt werden, versucht man im Fall demoskopischer Fragebogen die vielfaltigen Einflüsse der Fragenformulierung auf Antworten durch ‚Mittelungsprozeduren' zu eliminieren. Schon in den 30er Jahren (vgl. z.B. Roslow et al. 1940, Rugg 1941) wurde damit begonnen, innerhalb einer Befragung unterschiedliche Fragenformulierungen zu verwenden (gegabelte Befragung, split-ballot-verfahren) und als Ergebnis einen Mittelwert aus den in der Regel differierenden Antworten zu verwenden. Heute ist diese Vorgehensweise weithin üblich (vgl. z.B. Payne 1951, Stroschein 1965, Noelle-Neumann 1963, 1970, Rugg & Cantril 1972). Karmasin & Karmasin (1977) referieren eine Untersuchung, in der mit 12 verschiedenen Varianten des Fragebogens (variierten Reihenfolgen von Antwortkategorien) gearbeitet worden ist. Zweifel am Sinn dieses Verfahrens äußert allerdings bereits Noelle-Neumann (1970). Es fragt sich, was ein auf diese Weise zustandegekommenes ‚mittleres' Ergebnis eigentlich bedeutet. Es wäre sinnvoll nur interpretierbar, handelte es sich bei den Formulierungseffekten um ‚Zufallsfehler' mit einem Erwartungswert von Null, tatsächlich aber muß angenommen werden, daß mit unterschiedlich formulierten Fragen Unterschiedliches gemessen wird (Raab 1974), sonst dürften die Antwortunterschiede in Abhängigkeit von der Formulierung nicht (wie in diesem Kapitel oft berichtet) signifikant bzw. konsistent und stabil sein.

In mehreren Untersuchungen ist aufgezeigt worden, daß Antwortunterschiede in Abhängigkeit von der Fragenformulierung vor allem auftreten, wenn die Vpn von einem Sachverhalt nicht betroffen, an ihm nicht interessiert oder über ihn nicht informiert sind (vgl. vor allem Payne 1951, Noelle-Neumann 1970, Rugg & Cantril 1972). Sicherlich ist es in solchen Fällen unsinnig, aus Antworten, die weitestgehend methodenbedingt sind, einen Inhalt herausmitteln zu wollen. Angemessener wäre es wohl zu folgern, daß es fast ausschließlich von der Formulierung der Frage abhängt, was die Vpn antworten.

Ausgehend von den Überlegungen von Campbell & Fiske (1959) sollte man gewissermaßen im Rahmen eines ‚Multi-Content-Multi-Question-Ansatzes' für Fragen konvergente und diskriminante Validität fordern und Antworten erst dann interpretieren, wenn diese Forderungen erfüllt sind. Konkret würde das bedeuten, daß unterschiedlich formulierte Fragen zu einem bestimmten Inhalt zu ähnlichen Antworten führen müssen (Methodenkonvergenz), mindestens aber zu ähnlicheren Antworten als vergleichbar formulierte Fragen zu verschiedenen Inhalten (diskriminante Validität). Hält man sich vor Augen, daß allein durch Formulierungsähnlichkeit etwa auf dem Wege über die Wirkung von response sets hohe Korrelationen zwischen Antworten auf verschiedene Fragen zustandekommen und sich z.B. in einer Faktorenanalyse als ‚Formulierungsfaktor' niederschlagen können (Holm 1974a, b), wird man auch diskriminante Validität nicht mehr einfach unterstellen können, wie das heute noch vielfach geschieht.

4. Reihenfolge der Fragen und Umfang des Fragebogens

In stärkerem Maße noch als das bei der Formulierung von Fragen der Fall ist, stützt sich der Fragebogenaufbau üblicherweise auf Vermutungen und unsystematische Erfahrungen von Praktikern (Bradburn & Mason 1964). Die relativ wenigen empirischen Untersuchungen über Auswirkungen der Fragenreihenfolge und des Fragebogenumfangs können nur begrenzte Gültigkeit beanspruchen, so daß die Feststellung von Kreutz & Titscher (1974, 40), derzufolge „...über den Aufbau des Fragebogens sehr wenig gesichertes Wissen vorhanden ist", auch heute noch zutreffen dürfte.

4.1 Ziele beim Aufbau eines Fragebogens

Aus prinzipiell den gleichen Gründen, wie sie für die Festlegung von Fragenformulierungen angeführt wurden (vgl. 3.1.1), scheint in vielen Fällen auch die Standardisierung der Fragenfolge der mit dem geringeren Risiko für Verzerrungen behaftete Weg zu sein. Allerdings schließt dieser Weg naturgemäß die im freien mündlichen Interview gegebene Möglichkeit der Anpassung der Fragenfolge an die Erfordernisse der jeweiligen Befragungssituation durch den Interviewer aus. Damit sind prinzipiell Gefahren für die Motivation des Befragten verbunden, z.B. wenn ihm eine an früherer Stelle unaufgefordert bereits beantwortete Frage entsprechend ihrer Position im Fragebogen später erneut gestellt wird (Noelle 1963). Um diese und ähnliche Schwierigkeiten möglichst zu vermeiden, formulieren Karmasin & Karmasin (1977, 197) als Leitlinie für den Aufbau eines Fragebogens, diesen „. . . so zu gestalten, daß für den Befragten der Charakter eines Gesprächs, einer Konversation simuliert wird". Ähnlich äußern sich auch Kirschhofer-Bozenhardt & Kaplitza (1975). Praktisch bedeutet dies u.a., daß Fragen, die zusammenhängen, auch im Zu-

sammenhang zu stellen sind, zumal auf seiten der Vpn ein ausgeprägtes Bedürfnis zu bestehen scheint, Zusammenhänge zwischen Fragen bzw. Frageninhalten herzustellen (Karmasin & Karmasin 1977). Die Forderung nach Gruppierung des Zusammengehörigen kollidiert möglicherweise mit der den gebräuchlichen Meßmodellen zugrundeliegenden Annahme stochastischer Unabhängigkeit der Antworten auf verschiedene Fragen (Wottawa 1980), d.h. alle Arten von Reihenfolgeeffekten (seien sie unbeabsichtigt oder, wie bei Fragen, die auf einen bestimmten Sachverhalt hinführen sollen, bewußt eingesetzt), stellen einen Verstoß gegen Grundannahmen der Meßmodelle (Unkorreliertheit von Zufallsfehlern bzw. Abhängigkeit der Antworten nur von Item- und Personenparametern) dar. Aus diesem Grund ist - neben der ‚Natürlichkeit' des Gesprächsverlaufs - die Ausschaltung von Reihenfolgeeffekten (d.h. von Einflüssen auf die Antworten, die sich allein aus der Reihenfolge der Fragen ergeben) ein mit dem erstgenannten nur mehr oder weniger zu vereinbarendes Ziel des Fragebogenaufbaus.

Noelle-Neumann (1974) führt daneben die Motivierung der Befragten (d.h. das Bemühen, Interesse für die Befragung bzw. die Frageninhalte zu wecken) und die ‚Optimierung der Auskunftsfähigkeit' (d.h. die Steigerung bzw. Aufrechterhaltung der Aufmerksamkeit über den Befragungsverlauf) als wichtige Ziele an, die bei der Festlegung der Reihenfolge von Fragen im Auge zu behalten sind. Die Orientierung am natürlichen Verlauf eines Gesprächs kann in dieser Richtung wirken, ist mit diesen Zielen jedoch nicht identisch. Im Interesse der Verminderung interindividueller Beantwortungsunterschiede (Fehlervarianz im Falle der demoskopischen Befragung) und der Interpretierbarkeit von Subgruppenergebnissen müssen außerdem eine ‚Vergleichbarkeit des Befragungsablaufs' (z.B. im Falle von Verzweigungen) für alle Vpn angestrebt und bei der Festlegung der Fragenfolge auch die spätere Auswertbarkeit (Belange der Datenerfassung) im Auge behalten werden (Noelle-Neumann 1974).

Kontrovers behandelt wird die Frage, ob dem ‚natürlichen' bzw. ‚logischen' Aufbau des Fragebogens (der Fragenfolge) ein Wert an sich beizumessen (bzw. er infolge Strukturierungsbedürfnisses auf seiten der Vpn unvermeidbar) sei (z.B. Phillips 1966, Cannell et al. 1977, Karmasin & Karmasin 1977) oder ob ein ‚logischer' Fragebogenaufbau allenfalls ein denkbares Mittel unter vielen auf dem Weg zur Motivierung von Vpn, Verbesserung ihrer Auskunftsfähigkeit, Vermeidung von Reihenfolgeeffekten und Sicherstellung der Vergleichbarkeit des Befragungsablaufes darstelle (Stroschein 1965, Noelle-Neumann 1974). Entsprechend unterscheiden sich die Autoren auch darin, welchen Stellenwert sie dem ‚Themenwechsel' im Aufbau eines Fragebogens einräumen.

Bei der Abfolge von Fragen muß grundsätzlich unterschieden werden zwischen der Abfolge von Frageninhalten (‚Themendisposition' im Sinne von Stroschein 1965) und von Fragentypen (‚Fragendisposition', vgl. Stroschein 1965).

4.2 Motivation der Befragten und Steigerung der Antwortfähigkeit

Einerseits können sich durch Veränderungen der Motivation der Befragten im Verlaufe des Interviews Einflüsse auf Antworten in Abhängigkeit von der Position der Frage ergeben (vgl. 4.3.2, 4.4 und 4.5), insofern stellt die Motivation der Befragten eine mögliche Ursache beobachtbarer Reihenfolgeeffekte dar (Anger 1969). Andererseits ist die Gestaltung der Fragenfolge aber ein Mittel, angemessene Motivation der Befragten zu erreichen bzw. zu erhalten (vgl. besonders Perreault 1975). Hierzu schlägt Noelle-Neumann (1974, vgl. auch Noelle 1963) vor, besonderes Augenmerk den einleitenden Fragen zu schenken und diese (notfalls als ‚Wegwerf-Fragen') zu Kontakt- bzw. ‚Eisbrecher'-Fragen zu machen, die insbesondere mißtrauischen und unsicheren Vpn (z.B. älteren Menschen, Angehörigen der Unterschicht, Hausfrauen) ‚Sicherheit' vermitteln sollen. Als geeignete Themen gelten z.B. die erwartete Entwicklung der Preise, der Einfluß des Wetters auf die Befindlichkeit u.ä., wobei die Fragen zwar leicht zu beantworten und nicht kontrovers (Goode & Hatt 1972), andererseits aber auch nicht banal sein sollten. Um die Vpn ,,ins Gespräch zu ziehen" (Noelle-Neumann 1974, 244), empfiehlt sich evtl. eine offene Frage.

Während des Interviews sollen sowohl Motivation (Antwortbereitschaft) als auch Antwortfähigkeit (die sich vermutlich nicht streng trennen lassen) durch Wechsel der Themen, Wechsel der Inhalte (Wissen, Fakten, Meinungen, Verhalten) und Wechsel der Fragentypen (geschlossene, offene Fragen, Listen- oder Kartenvorlagen, wechselnde Formate und Farben des Vorlagematerials) aufrechterhalten bzw. gesteigert werden (Stroschein 1965). Noelle-Neumann (1963, 1974) schlägt sogar einen eigenen Typ instrumenteller Fragen, die sogenannten ‚Spielfragen' (Beurteilung von Frisuren, Kleidern, Farbwahlen etc.) nur zur Beeinflussung von Motivation bzw. Aufmerksamkeit vor. Lange Serien geschlossener Fragen gelten als frustrierend und monotonieförderndd, Serien offener Fragen als anstrengend und dadurch ermüdend (Noelle-Neumann 1974).

Die behaupteten Wirkungen spezieller Einleitungsfragen und der verschiedenen Techniken zur Beeinflussung von Aufmerksamkeit und Motivation sind empirisch allerdings nicht abgesichert (Kreutz & Titscher 1974).

Im Zusammenhang mit der Forderung nach häufigem, durchaus auch sprunghaftem (Noelle 1963) Themenwechsel werden von Autoren aus dem Bereich der kommerziellen Markt- und Meinungsforschung die sogenannten Mehrthemenumfragen (Omnibus-Befragungen) als methodisch besonders vorteilhaft hervorgehoben. Es stellt sich allerdings die Frage, ob hier nicht eine Not zur Tugend gemacht werden soll. Immerhin betont Richter (1969), daß jeder The-

menwechsel mit einer besonderen Anstrengung für den Befragten (Umorientierung) verbunden sei, und fordert für unpersönlich-schriftliche (postalische) Befragungen im Interesse eines hohen Rücklaufs eine Zusammenstellung von Fragen nach Maßgabe der Befragungsthemen zu sogenannten ‚assoziativen Blöcken'. Eine möglichst sinnvolle Ordnung von Fragen wird von Autoren wie Phillips (1966), Goode & Hatt (1972), (mit Einschränkungen) Holm (1974b), Cannell et al. (1977), Karmasin & Karmasin (1977) gefordert (vgl. auch 4.1). Auch Anger (1969) warnt vor zu starkem Themenwechsel, von dem er Gefahren für die erlebte Seriosität der Befragung ausgehen sieht. Empirische Untersuchungen, in denen Themenwechsel und logischer Fragebogenaufbau verglichen worden wären, scheinen nicht zu existieren (vgl. Kreutz & Titscher 1974), allerdings ist auch zweifelhaft, ob sie generalisierbare Befunde zutage fördern könnten. Vermutlich gibt es nur den Weg, unter Berücksichtigung von möglichen Monotonieeffekten einerseits und Belastungen durch inhaltliche Umorientierung sowie des Bedürfnisses der Vpn nach sinnvollem Zusammenhang der Fragen andererseits, die günstigste Themen- und Fragendispostition im Einzelfall durch Pretest empirisch zu bestimmen.

4.3 Reihenfolgeeffekte

Einflüsse der Stellung einer Frage innerhalb eines Fragebogens auf die Antworten können einmal durch die Inhalte vorangegangener Fragen, dann aber auch unabhängig von diesen Inhalten dadurch zustande kommen, daß die Frage früher oder später im Verlauf einer Befragung gestellt wird und das Antwortverhalten der Vpn sich über die Dauer der Befragung verändert (Bradburn & Mason 1964). Da vorausgehende Fragen immer Inhalte haben und andererseits das Vorausgehen einer Frage bestimmten Inhaltes die zu betrachtende Frage notwendig an eine spätere Stelle verschiebt, ist es prinzipiell nicht möglich, diese Effekte völlig voneinander zu isolieren. Im Interesse theoretischer Klarheit werden dennoch im folgenden als ‚Kontexteffekte' Einflüsse des Inhaltes vorangehender Fragen und als ‚Positionseffekte' Einflüsse der relativen Position auf die Antworten zu einer gegebenen Frage unterschieden.

4.3.1 Kontexteffekte

Diese oft auch als Ausstrahlungseffekte angesprochenen Einflüsse vorangegangener auf nachfolgende Fragen lassen sich nach einem auf Bradburn & Mason (1964) zurückgehenden Vorschlag in

- Aktualisierungs- (Präsenz-, saliency-) Effekte oder allgemeiner Lerneffekte (Anger 1969),
- Konsistenzeffekte und
- Redundanzeffekte

einteilen.

Aktualisierungseffekte kommen dadurch zustande, daß eine vorausgegangene Frage die Antworten auf eine nachfolgende beeinflußt, indem sie bestimmten Sachverhalten oder bestimmten Bezugsrahmen im Bewußtsein der Befragten höheres Gewicht verleiht. Hierfür finden sich in der Literatur mehrfach Beispiele bzw. empirische Belege. Stellt man z.B. zunächst eine Frage nach Erwartungen zur Preisentwicklung und danach eine solche nach den wichtigsten Fragen, mit denen Politiker sich in nächster Zeit beschäftigen sollten, ist zu erwarten, daß die Preisstabilität erheblich häufiger genannt wird, als sie ohne eine derartige vorausgegangene Frage genannt worden wäre (Noelle-Neumann 1974), einfach weil die Preisstabilität als politisches Thema einen höheren Grad der Bewußtheit erhalten hat. Durch Aktualisierung eines geeigneten Bezugsrahmens läßt sich erklären, daß ‚Kartoffeln' in einer Untersuchung, die Noelle-Neumann (1970) referiert, von 30% der Befragten die Eigenschaft eines ‚deutschen' Nahrungsmittels zugesprochen wurde, wenn nach ihnen vor, von 48%, wenn nach ihnen h i n t e r dem Nahrungsmittel ‚Reis' gefragt wurde (ähnliche Reihenfolgeeffekte gab es auch für Reis und Nudeln). Willick & Ashley (1971) befragten College-Studenten, welche politischen Parteien sie und welche ihre Eltern bevorzugen würden, und erhielten signifikant mehr übereinstimmende Angaben (für die eigene Bevorzugung und die der Eltern), wenn zuerst nach der Bevorzugung des Studenten und danach nach der der Eltern gefragt wurde. Sie erklären diesen Befund mit dem Bemühen der Studenten, Unabhängigkeit von den Ansichten ihrer Eltern zu demonstrieren. Dies war dann nicht ohne weiteres möglich, wenn die Studenten zum Zeitpunkt der Antwort betreffend ihre eigene Meinung nicht wußten, daß sie auch nach der Haltung ihrer Eltern (die für sie in der Regel anschaulich festlag) befragt werden würden. Weitere Beispiele beziehen sich auf das Recht des Eintritts für Amerikaner in die deutsche bzw. englische oder französische Armee während des Zweiten Weltkrieges (Rugg & Cantril 1972) und auf die Haltung von Einwohnern der BRD gegenüber den USA bzw. der UdSSR (Noelle-Neumann 1970).

Aktualisierung kann je nach Befragungszielen ein unerwünschter, evtl. aber auch ein erwünschter Reihenfolgeeffekt sein. Besteht das Ziel der Befragung darin, Beurteilungen oder Bewertungen von Sachverhalten möglichst unbeeinflußt von Aktualisierungen zu erhalten, wählt man häufig eine als ‚Trichter' bezeichnete, vom Allgemeineren zum Spezielleren fortschreitende Reihenfolge (vgl. Maccoby & Maccoby 1972, Hennig 1975, Karmasin & Karmasin 1977). Friedrichs (1973) beschreibt die von Gallup verwendete Standard-Fragenfolge eines Trichters:

- Vertrautheit mit dem Sachverhalt
  (offene Wissensfrage, z.B. ‚Was verstehen Sie unter . . .'),
- unbeeinflußte Einstellung
  (offene Einstellungsfrage, z.B. ‚Was sollte X für . . . tun?'),

- Reaktion auf spezifische vorgegebene Einstellungen (geschlossene Fragen, z.B. ‚Manche sagen . . . andere sagen . . . was meinen Sie ist richtig?'),
- Begründung der Reaktion auf vorgegebene Einstellungen (offene Warumfrage),
- Intensität der Einstellung (Skalafrage).

In Fällen, in denen die aktualisierende Wirkung vorausgegangener Fragen befragungstaktisch erwünscht bzw. erforderlich ist, bedient man sich gelegentlich auch der Technik des umgekehrten Trichterns, d.h. des Fortschreitens vom Speziellen, Konkreten zum Allgemeinen, Abstrakten (Maccoby & Maccoby 1972). Hennig (1975) führt z.B. aus, daß eine Frage an Arbeiter nach der vorausgesehenen Organisation von Produktionsabläufen in zehn Jahren kaum zu verwertbaren Antworten führen dürfte, wenn sie ‚unvermittelt' gestellt wird. Erfolgversprechender ist hier eine Fragenfolge nach Art eines umgekehrten Trichters, z.B.

- ‚Sind in Ihrem Betrieb in der nächsten Zeit Neuerungen im Produktionsablauf geplant? Wenn ja: welche?',
- ‚Was glauben Sie, wie in zehn Jahren der Produktionsablauf aussehen wird?'.

Befragungstaktisch beabsichtigt und gezielt eingesetzt werden Aktualisierungen durch geeignete Fragenfolge auch, wenn seitens der Vp eine Reproduktion von Gedächtnisinhalten erforderlich ist. Cannell & Kahn (1968) schlagen in solchen Fällen vor, die Vpn z.B. durch chronologisch geordnete Fragen auf den thematischen Sachverhalt hinzuführen (vgl. auch Mauldin & Marks 1950, Phillips 1966).

Von *Konsistenzeffekten* der Fragenfolge spricht man, wenn die Vp eine Frage nicht ‚zutreffend' sondern so beantwortet, daß sie zu ihren Antworten auf vorangegangene Fragen nicht in Widerspruch gerät. Noelle-Neumann (1974) nennt als Beispiel Aussagen von Befragten über Aufwendungen für ‚Luxusartikel' (z.B. Blumen), die dann niedriger angegeben werden, wenn die Befragten in einer vorangegangenen Frage ‚sparsame Lebensführung' für sich in Anspruch genommen haben. Auch Holm (1974 b) weist auf solche Gefahren hin, insbesondere dann, wenn die Fragen zu einer bestimmten ‚Zieldimension' gruppiert (zusammengestellt) sind. Bradburn & Mason (1964) gelang es andererseits nicht, in ihren Untersuchungen Anhaltspunkte für derartige Reihenfolgeeffekte zu finden: Antworten auf eine Frage nach ‚globaler' Zufriedenheit wurden hier nicht davon beeinflußt, ob Fragen zur Zufriedenheit mit speziellen Aspekten vorausgegangen waren oder nicht.

Ist mit ausgeprägten Konsistenzneigungen der Vpn zu rechnen, so empfiehlt sich nach Noelle-Neumann (1974) eine Fragenfolge nach Art des umgekehrten

Trichters: Spezifische Angaben (z.B. ‚Ausgaben für Luxusartikel') würden dann als weniger widersprüchlich mit allgemeinen (‚prinzipielle Sparsamkeit') empfunden, wenn sie diesen vorangehen. Mit ähnlicher Begründung empfehlen Tittle & Hill (1967) erst nach *Verhalten* und dann nach *Einstellungen zu* fragen.

Als *Redundanzeffekt* bezeichnen Bradburn & Mason (1964) das Ausbleiben bestimmter Antworten auf Fragen dadurch, daß diese Antworten bereits auf frühere Fragen gegeben wurden und die Vpn sich nicht wiederholen wollen. Sie berichten von geringeren Häufigkeiten für die Nennung bestimmter (z.B. Partner-)Probleme in einer offenen Frage nach ‚Sorgen', wenn diese Probleme bereits Gegenstand vorangegangener Fragen waren. Auch Noelle-Neumann (1974) betont, daß man in solchen Fällen die Antworten auf die spätere Frage nicht unabhängig von denen auf die vorangegangenen Fragen betrachten und behandeln dürfe.

In gewissem Sinne liegen Reihenfolgeeffekte nach Art von Kontexteffekten auch vor, wenn durch (erfolgreiche) Verwendung von Puffer- oder Ablenkungsfragen (vgl. 2.1) Auswirkungen früherer auf spätere Fragen *vermieden* werden: Auch hier lauteten die Antworten anders, würden diese instrumentellen Fragen der thematischen Frage nicht vorangehen. Darüber hinaus lassen sich durch die Schwierigkeitsabstufung von Fragen Reihenfolgeeffekte erzeugen, etwa im Sinne einer Erleichterung besonders schwieriger Fragen durch langsamen Schwierigkeitsanstieg oder im Sinne einer Überwindung von ‚Antworthemmungen' durch starke Schwierigkeitsunterschiede (vgl. 4.4).

Daß im Falle unpersönlicher schriftlicher (postalischer) Befragungen infolge Nichtkontrollierbarkeit der Reihenfolge der Bearbeitung von Fragen auch Auswirkungen nachfolgender auf vorangehende Fragen möglich sind, sei der Vollständigkeit halber erwähnt. Entsprechendes gilt, da der Interviewer ja den ganzen Fragebogen kennt und die registrierten Antworten auf verschiedenen Wegen mitbeeinflußt, übrigens auch für mündliche Interviews.

4.3.2 Positionseffekte

Nach Richter (1969) und Goode & Hart (1972), die dafür allerdings empirische Belege nicht vorlegen, nimmt die Wahrscheinlichkeit für den Abbruch eines Interviews vom Anfang zum Ende des Interviews hin ab. Karmasin & Karmasin (1977) führen dies (im Falle von mündlichen Interviews) darauf zurück, daß mit der Interaktionshäufigkeit auch die Sympathie zwischen Interviewer und Befragtem (man muß wohl ergänzen: in der Regel) wachse, und leiten daraus auch eine Tendenz *zu weniger negativen* Urteilen an späterer Stelle im Interview ab. Kraut et al. (1975) konnten empirisch allerdings nur leichte Tendenzen in Richtung auf *weniger extreme* Urteile und mehr Auslassungen

bzw. Nichtbeantwortungen gegen Ende einer persönlichen schriftlichen Befragung nachweisen. Von abnehmender ‚Sorgfalt' im Verlauf des Interviews berichtet auch Stroschein (1965). Johnson et al. (1974) kamen in entsprechenden Untersuchungen für eine offene Frage zu dem Ergebnis, daß diese insgesamt am meisten Information lieferte, wenn sie einmal am Anfang und dann erneut am Ende einer Serie von 18 bzw. 62 geschlossenen Fragen gestellt wurde. Besteht (wie im Regelfall) nur die Möglichkeit, die Frage einmal zu stellen, so liefert sie am Anfang des Fragebogens mehr Information als am Ende, d.h. der mögliche Zugewinn an Aspekten durch Lernprozesse während der Befragung wird durch Effekte verringerter Motivation bzw. Aufmerksamkeit überkompensiert.

Zusammenfassend können die vorliegenden Befunde wohl als Hinweise darauf gelten, daß insbesondere schwierige und anstrengende Fragen nicht zu spät, wegen der Abbruchgefahr aber auch nicht zu früh im Fragebogen auftauchen sollten. Beurteilungen sind in ihren extremen Ausprägungen nur vergleichbar, wenn die Fragen etwa gleiche Positionen im Fragebogen hatten, betrachtet man allerdings nicht die Extremkategorien (z.B. ‚sehr dafür'), sondern die Mittelwerte der Einstufungen, so sind (mindestens nach den Befunden von Kraut et al. 1975) Positionseffekte kaum noch zu befürchten.

4.4 Unangenehme und heikle Fragen

übereinstimmend findet sich in der Literatur die Empfehlung, unangenehme bzw. heikle Fragen (z.B. solche nach Einkommen, Kindererziehung, Allgemeinbildung, Sexualität, Familienverhältnissen, körperlicher Sauberkeit, vgl. 2.3) erst in der zweiten Hälfte des Fragebogens zu stellen, um einerseits das verringerte Risiko für Abbrüche, andererseits das angewachsene ‚Vertrauen' der Befragten zu nutzen (Kreutz & Titscher 1974, Karmasin & Karmasin 1977). Darüber hinaus gibt es jedoch spezielle Techniken der Berücksichtigung heikler Fragen im Fragebogenaufbau. Im Sinne einer Erfahrungsregel schlägt Noelle-Neumann (1974) z.B. vor, solche Fragen besonders einfach zu formulieren und sie nach betont schwierigen Fragen (z.B. offenen Wissensfragen mit schwierigen Inhalten) zu plazieren. Durch einen Kontrasteffekt werde die kritische Frage dann als besonders leicht erlebt und gewissermaßen nach Art einer Selbstüberrumpelung beantwortet, bevor auf Seiten der Vp mögliche Antworthemmungen überhaupt zur Wirkung kommen könnten. Koolwijk (1968) fand in seinen Untersuchungen Angleichungs- und Kontrasteffekte für die Unangenehmheit von Fragen: Ließ er eine unangenehme Frage auf eine neutrale folgen, so verstärkte sich die Unangenehmheit (Kontrast), ging umgekehrt die unangenehme Frage der neutralen voran, so wurde tendentiell auch letztere als unangenehm erlebt (Angleichung, halo). Aus diesen Befunden leitet er die Forderung ab, einer inhaltlich interessierenden unangenehmen Frage

zur ‚Einstimmung' und zur Vermeidung des Kontrasteffektes eine ebenfalls unangenehme (evtl. Wegwerf-Frage) voranzuschicken und nachfolgende neutrale Fragen durch Pufferfragen gegen vorausgegangene unangenehme Fragen abzuschirmen.

Schließlich haben Goode & Hatt (1972) darauf hingewiesen, daß Unangenehmheit nicht nur eine Einzelfrage, sondern auch eine Fragenfolge kennzeichnen kann, d.h. sie fordern für den Fragebogenaufbau die Vermeidung von Fragenfolgen, die für bestimmte Vpn peinlich werden könnten. Der Fragenfolge 1. ‚Haben Sie Kinder?', 2. ‚Sind Sie verheiratet?' wäre unter diesem Kriterium die umgekehrte Folge mit Filterung ‚Sind Sie verheiratet?, wenn ja: Haben Sie Kinder?' vorzuziehen.

4.5 Fragen zur Person

Die unsicheren Grundlagen einer an Erfahrungsregeln statt an empirischen Untersuchungsergebnissen orientierten Fragebogenkonstruktion werden im Zusammenhang mit den Empfehlungen deutlich, die verschiedene Autoren für die Position von Angaben zur Person (biographischen Angaben bzw. demographischen Fragen) geben. So stellt Noelle-Neumann (1974, 244) nachdrücklich fest: ,,Personenstandsdaten gehören nicht an den Anfang des Interviews, sondern an das Ende; an den Anfang gesetzt geben Sie dem Interview den Charakter eines Verhörs". Eine ähnliche Position vertritt auch Stollberger (1966). Dem steht die Auffassung von Kreutz & Titscher (1974) entgegen, die Fragen zur Person an den Anfang des Fragebogens gestellt wissen möchten, weil ihnen einerseits die Forderung nach der Schlußposition für solche Fragen empirisch nicht begründet zu sein scheint und sie andererseits sorgfältigere und damit gültigere Antworten erwarten, wenn die Vp gleich zu Beginn der Befragung (bei den Fragen zur Person) feststellt, daß sich der Untersucher für sie als Individuum interessiert. Auch hier handelt es sich allerdings um eine Spekulation, die man im Vergleich zur Gegenposition für plausibler halten kann oder auch nicht.

4.6 Filterfragen und Verzweigungsfragen

Ablauf-Ordnungsfragen (vgl. 2.1) wie Filter- und Gabelungs- bzw. Verzweigungsfragen stellen im Interesse der Vermeidung von ‚Unterstellungen' (vgl. 3.1.4) bzw. der Nichtbelastung von Vpn mit unzutreffenden Fragen ein häufig unverzichtbares Mittel bei der Festlegung einer angemessenen Fragenfolge innerhalb eines Fragebogens dar. Andererseits sollte sich der Fragebogenkonstrukteur vergegenwärtigen, daß mit dem Einbau von Filterungen und Verzweigungen in einen Fragebogen die Fehlerhäufigkeit unweigerlich ansteigt.

So berichten Cannell et al. (1977) für mündliche standardisierte Interviews, daß

- von Fragen, die allen Vpn gestellt werden sollten, nur 1,5% bei mehr als 10% der Vpn ausgelassen wurden, aber
- von Fragen, die infolge Filterung bzw. Verzweigung nur für Subgruppen vorgesehen waren, 54% bei mehr als 10% der Vpn nicht oder nicht richtig gestellt wurden.

Richter (1969) hat für den Fall unpersönlich-schriftlicher (postalischer) Befragung die Beachtung von Filteranweisungen untersucht und je nach Bildungsstand, Einkommen, Beruf etc. Nichtbeachtungsanteile bis in die Größenordnung von 30% der Befragten gefunden, so daß die Verwendung von Filterungen und Gabelungen beim Aufbau von Fragebogen für unpersönlich-schriftliche Befragung möglichst zu vermeiden ist (Wieken 1974).

4.7 Spezielle Gesichtspunkte für die Itemreihenfolge diagnostischer Fragebogen

Für diagnostische Fragebogen stellt sich insbesondere die Frage, ob Items nach ihrer ‚Zieldimension' gruppiert bzw. ob sie in Zufallsfolge vorgegeben werden sollten. Die Gruppierung von Items nach ihrem Inhalt erhöht tendenziell die Durchschaubarkeit und wird daher für empirische Konstruktionen (vgl. 1.1.2) von vornherein nicht in Betracht gezogen. Hier liegen prinzipiell zufällige Itemfolgen vor, die allenfalls im Interesse leichterer Auswertbarkeit durch für die Vpn nicht ersichtliche systematische Anordnungen durchbrochen werden (vgl. als Beispiel den MMPI, die ‚Lügenitems' sind hier systematisch angeordnet, Hathaway & Mc Kinley 1963).

Sieht man von der höheren Durchschaubarkeit und damit Verfälschbarkeit und der Störung der stochastischen Unabhängigkeit der Einzelantworten ab, so kann für Fragebogen mit inhaltlichem Validitätsanspruch eine inhaltliche Gruppierung von Items zu einer ‚Sensibilisierung' der Vp in dem Sinne führen, daß sie durch die Zusammenstellung der Items ihre ‚Lage' auf der Zieldimension besser bestimmen kann, als sie das anhand verstreuter Items tun könnte. Dies würde zu valideren Antworten führen. Andererseits kann die Zusammenstellung von Items die Vp aber auch zu inadäquat konsistentem Antwortverhalten veranlassen und dadurch die Validität beeinträchtigen. Es ist nur im Einzelfall zu klären, ob vorwiegend die (erwünschte) Sensibilisierungstendenz' oder die (unerwünschte) ‚Konsistenztendenz' (Holm 1974 b) durch eine inhaltliche Gruppierung der Items begünstigt wird. Daß eine solche Gruppierung nicht notwendig zu artifizieller Konsistenz im Antwortverhalten führen muß, haben Metzner & Mann (1953) gezeigt: Sie fanden keine systematische Erhöhung der Interkorrelationen von Items durch Gruppierung; für Einzelfäl-

le berichten sie sogar erhebliche Senkungen der Korrelationen zwischen Items. Sie begründen dies einleuchtend mit itemspezifischen Kontexteffekten, d.h. die Bedeutung eines Items kann sich durch direkte Nachbarschaft zu anderen Items ändern und zwar nicht nur (wie unter der Konsistenz-Hypothese erwartet) in Richtung auf höhere (Un-)Ähnlichkeit zu den anderen Items der entsprechenden Dimension.

Darüber hinaus sind bei diagnostischen Fragebogen die Effekte der Reihenfolge der Items abhängig von der ‚Sicherheit' bzw. ‚Zugänglichkeit' des zu erfassenden Sachverhaltes bzw. der Lage einer Vp auf dem interessierenden latenten Kontinuum. Hayes (1964) konstruierte je eine Guttman-Skala aus ‚Angst-Items' und ‚Mathematik-Aufgaben' und stellte fest, daß zwar bei Angst-Items, nicht aber bei Mathematik-Aufgaben die (in aufsteigender Schwierigkeit) geordnete Vorgabe der Items zu signifikant anderen Antworten und damit (geringeren) Angst-Werten führte als die ungeordnete Vorgabe. Eingeschobene ‚irrelevante' Items blieben bei beiden Skalen ohne Auswirkungen.

4.8 Überlegungen zur Vermeidung unerwünschter Reihenfolgeeffekte

Für die Elimination von Reihenfolgeeffekten aus den Ergebnissen für Vpn-Gruppen gelten prinzipiell diejenigen Überlegungen, die im Zusammenhang mit der Kontrolle von Einflüssen der Fragenformulierung in 3.4 angestellt worden sind. Wie dort, so kann auch bezüglich der Fragenreihenfolge im Rahmen einer gegabelten Befragung (Split-ballot-verfahren) mit verschiedenen Fragebogenvarianten gearbeitet und die Absicht verfolgt werden, die Reihenfolgeeinflüsse ‚herauszumitteln' (vgl. zu dieser Vorgehensweise die Ausführungen über die Variation der Reihenfolge von Antwortvorgaben in 3.1.3). Wie im Falle der Formulierungseffekte muß aber auch hier gefragt werden, ob solche ‚mittlere' Antworten eine inhaltliche Bedeutung haben oder ob die Existenz von Reihenfolgeeffekten ein Hinweis darauf ist, daß die Erfassung des Inhaltes mit der verwendeten Methode nicht oder nur unzureichend gelingt.

Hält man die Variation der Fragenreihenfolge für ein angemessenes Verfahren, so ist es prinzipiell günstig, mit möglichst vielen Varianten des Fragebogens zu arbeiten und im Extremfall für jeden Befragten einen eigenen Fragebogen zu erstellen. Die damit verbundenen Probleme und die Möglichkeiten, die der Einsatz elektronischer Datenverarbeitungsanlagen zur Erstellung auswertbarer individualisierter Fragebogen bietet, diskutiert Perreault (1976; für den Fall des semantischen Differentials vgl. auch Kane 1969). Cataldo et al. (1970) schlagen nicht zuletzt wegen leichterer Variierbarkeit der Reihenfolge den verstärkten Einsatz von card-sorting-Techniken im Rahmen von (persönlichen) Befra-

gungen vor: Statt für die einzelnen Statements Fragen nach dem Grad ihres Zutreffens beantworten zu lassen, werden dabei die Statements auf Karten geschrieben und den Vpn zur Einordnung in bestimmte Antwortkategorien (z.B. ‚sehr dafür', ‚dafür', ‚dagegen', ‚sehr dagegen') übergeben.

Soweit man die Variation der Reihenfolge nicht für ein angemessenes Vorgehen hält, mit der Existenz von Reihenfolgeeffekten aber rechnen muß, ist entsprechend den Ergebnissen von Kraut et al. (1975; vgl. auch 4.3.2) zu berücksichtigen, daß streng vergleichbar nur Antworten auf Fragen sind, die in Fragebogen an der gleichen Position (und im gleichen Kontext) verwendet wurden, und daß die Antworten außer inhaltlichen auch Reihenfolgeeffekte wiederspiegeln. Je intensiver, klarer, sicherer die zu erfassenden Sachverhalte für die Vpn sind, *desto* geringer sind ceteris paribus die Einflüsse der Fragenfolge auf die Antworten (vgl. Bradburn & Mason 1964, Hayes 1965, Willick & Ashley 1971).

4.9 Fragebogenumfang

Zwar findet sich in fast jeder Darstellung von Befragungsmethoden auch eine Angabe bzw. Empfehlung bezüglich des akzeptablen Fragebogenumfanges bzw. der akzeptablen Interview- bzw. Bearbeitungsdauer, doch handelt es sich dabei stets nur um Erfahrungswerte bzw. common-sense-Angaben. Typisch dafür ist z.B. Noelle (1963), die als Richtwert für die Dauer eines mündlichen Interviews 30 Minuten nennt (ähnliche Werte finden sich z.B. auch bei Kirschhofer-Bozenhardt & Kaplitza 1975, Karmasin & Karmasin 1977), bei ‚gutem Aufbau' aber auch mehr als eine Stunde für möglich hält. Als Indikator für die Einhaltung bzw. Überschreitung der akzeptablen Dauer schlägt sie vor, die Befragten am Ende des Interviews diese Dauer schätzen zu lassen: Werde sie unterschätzt, sei das Interview nicht zu lang gewesen.

Empirische Untersuchungen der Wirkung unterschiedlicher Fragebogenumfänge wurden - soweit bekannt - nur im Zusammenhang mit unpersönlich-schriftlichen (postalischen) Befragungen durchgeführt, wobei abhängige Variable stets allein die Rücklaufquote war. Berdie (1973), der auch ältere Untersuchungen referiert und die prinzipiell plausible negative Korrelation zwischen Fragebogenumfang und Rücklaufquote, von der viele Autoren berichten, als tradiertes ‚Einvernehmen' ohne nennenswerte empirische Basis entlarvt, fand zwar unterschiedliche Rücklaufquoten von 64%, 56% und 42% für Fragebogen mit einer Seite (10 Fragen), 2 Seiten (20 Fragen) und 4 Seiten (40 Fragen), doch waren diese Unterschiede (bei 108 Vpn) statistisch nicht bedeutsam. Sheth & Roscoe (1975) verglichen die Rücklaufquoten für einen vierseitigen (23 Items, 10 min Bearbeitungszeit) und einen sechsseitigen (49 Items, 18 min Bearbeitungszeit) Fragebogen und fanden keinen Unterschied, wobei aller-

dings zu bedenken ist, daß die Fragebogenumfange nur wenig differierten und der längere Fragebogen sich vom kürzeren auch inhaltlich systematisch unterschied. Unterhalb des von ihm untersuchten Maximums von 4 Seiten und einer Bearbeitungszeit von 15-20 min fand auch Richter (1969) keinen Zusammenhang zwischen Fragebogenumfang und Rücklaufquote, und Perreault (1975) berichtet von sehr hohen Rücklaufquoten sogar für einen 9 Seiten umfassenden Fragebogen, der allerdings ‚personalisiert' (mit persönlich wirkendem Anschreiben etc. versehen) war (vgl. auch Erdos 1970, Linsky 1975).

Natürlich ist mit diesen Untersuchungen nicht bewiesen, daß es keinen Einfluß des Fragebogenumfangs auf Rücklaufquoten gibt, nur ist (möglicherweise durch die Anlage der Untersuchungen) der Nachweis für einen solchen Zusammenhang noch nicht eindeutig erbracht worden. Auswirkungen des Fragebogenumfangs auf andere Variablen (insbesondere die Qualität der Antworten) und für andersartige Befragungstechniken (z.B. persönliche Befragung) wurden erst gar nicht untersucht.

Solche Untersuchungen müßten berücksichtigen, daß der Fragebogenumfang d r e i (n i c h t unabhängige, aber unterscheidbare) Aspekte, die Item-Anzahl, die Seitenzahl und die Bearbeitungsdauer, aufweist und daß der von der Vp erlebte Umfang nicht notwendig mit dem ‚objektiven' Umfang identisch sein muß (Richter 1969 und Erdos 1970 betonen z.B. die Wichtigkeit der Gliederung von Fragenserien; vgl. auch die Ausführungen über die äußere Gestaltung des Fragebogens, 5.). Besondere Schwierigkeiten für die Untersuchung der Auswirkungen des Fragebogenumfangs ergeben sich einmal aus den zu erwartenden Interaktionen mit anderen Merkmalen, dann aber auch aus der Tatsache, daß der Fragebogenumfang nicht ohne gleichzeitige Veränderung entweder des Fragebogeninhalts (bei einer Vermehrung der Zahl von Fragen) oder der Fragebogengestaltung (bei einer Verteilung der Fragen auf mehrere Seiten) vergrößert werden kann. Vielleicht liegt in diesen methodischen Schwierigkeiten eine Erklärung für den bemerkenswerten Mangel an empirischen Untersuchungen zur Rolle des Fragebogenumfangs.

5. Äußere Gestaltung (Layout) des Fragebogens

Fragen der typographischen und farblichen Gestaltung und des Layouts von Fragebogen sind kaum empirisch untersucht worden. Entsprechend gibt es sowohl was schriftlich zu bearbeitende Fragebogen (vgl. Hartley et al. 1977), als auch was Interviewer-Fragebogen für mündliche Befragungen betrifft (vgl. Haase 1978) wenig gesicherte Erkenntnisse. Andererseits berichtet Gray (1975) für unpersönlich-schriftliche (postalische) Befragungen durch Verbesserung der graphischen Gestaltung des Fragebogens im Vergleich zu einer maschinenschriftlichen ersten Version Rücklaufsteigerungen von ca. 30%, Ver-

minderungen der Bearbeitungszeiten von ca. 40 auf 20 Minuten und der Ablochzeiten um etwa 1/3. Rücklaufsteigerungen um immerhin noch 8% durch veränderte graphische Gestaltung fand auch Richter (1969). Es scheint also, daß gerade in der äußeren Aufmachung von Fragebogen erhebliche Möglichkeiten für eine Optimierung unter dem Kriterium der Qualität der Antworten und/oder unter ökonomischen Aspekten liegen.

Die *Verwendung von* Farben kann innerhalb eines Fragebogens unterschiedlichen Zielen dienen. Bei unpersönlich-schriftlichen (postalischen) Befragungen wird mitunter versucht, durch Wahl einer ansprechenden Papierfarbe den Rücklauf günstig zu beeinflussen, zumal damit keine ins Gewicht fallenden zusätzlichen Kosten verbunden sind. Sharma & Singh (1967) konnten - allerdings bei hochmotivierten und akademisch gebildeten Vpn mit einem Gesamtrücklauf von 87,7% - keinerlei Einfluß der Papierfarbe (weiß, rosa, gelb) auf den Rücklauf feststellen, ebensowenig gelang dies Gullahorn & Gullahorn *(1963)* für die Farben weiß und grün. Das bedeutet natürlich nicht, daß für andere Farben, in anderen Populationen und bei Fragebogen mit anderen formalen und inhaltlichen Merkmalen solche Einflüsse ebenfalls ausgeschlossen wären. Wegen der dadurch entstehenden Ähnlichkeit mit Werbedrucksachen warnt Erdos (1970) vor mehrfarbigen Fragebogen für unpersönlich-schriftliche (postalische) Befragungen.

Mit Vorteil läßt sich Farbe zur Kennzeichnung von Gabelungen und Verzweigungen in Fragebogen im Interesse einer besseren Handhabbarkeit durch den Interviewer einsetzen. Derartigen Farbkodierungen sind in der Praxis allerdings durch die hohen Kosten mehrfarbigen Druckes enge Grenzen gesetzt (Noelle 1963).

Ausgiebiger Gebrauch wird vom Medium Farbe bei Listen- und Kartenvorlagen gemacht, einmal im Interesse der Abwechslung für den Befragten (vgl. 4.2), aber auch zum Zwecke besserer Unterscheidbarkeit und eindeutiger Zuordnung zu den betreffenden Fragen für den Interviewer. Unproblematisch ist dies jedoch nur, wenn man davon ausgehen kann, daß die verwendete Farbe die Verteilung der Antworten (gewählten Karten) nicht beeinflußt. Ring (1969) versuchte dies für rote und graue Kartenvorlagen zu klären. Er kam zu dem Ergebnis, daß weder Zahl noch Art der Antworten durch die Hintergrundfarbe der Kartenvorlage beeinflußt wurden (die vereinzelt und unsystematisch aufgetretenen Antwortunterschiede in Abhängigkeit von der Farbe des Kartensatzes lassen sich als Produkte des Zufalls betrachten).

Obwohl es für die Motivation der Befragten vorteilhaft sein dürfte, wenn der Fragebogen durch Bedrucken der Vorder- und Rückseiten kurz erscheint (Erdos 1970), ist von diesem Vorgehen abzuraten: Fragen auf der Rückseite werden zu häufig übersehen (Kirschhofer-Bozenhardt & Kaplitza 1975).

Im Hinblick auf das Layout im engeren Sinne fordern Karmasin & Karmasin (1977) bei Fragebogen für mündliche Befragungen vor allem eine deutliche optische Unterscheidung (z.B. durch verschiedene Schrifttypen, Umrahmungen u.ä.) zwischen Anweisungen an den Interviewer, eigentlichem Fragentext und Antwortvorgaben. Nach Richter (1969) sollte bei schriftlich zu bearbeitenden Fragebogen (besonders im Falle unpersönlicher Befragung) auf Seiten der Vp der Eindruck vermieden werden, es handele sich um lange Fragenserien oder um viele einzelne Fragen, damit Ermüdungs-, Sättigungs- und Monotonieerlebnisse bei der Beantwortung möglichst gering gehalten werden können. Dazu schlägt er vor, einerseits der Einzelfrage nicht zuviel optisches Gewicht zu geben, sondern sie in einen Fragenblock einzugliedern, andererseits aber diese Fragenblöcke auch nicht zu umfangreich zu gestalten und sie durch Überschriften etc. voneinander abzuheben. Eine Beschreibung und Diskussion verschiedener Schrifttypen, Typengrößen, Satz- und Drucktechniken findet sich z.B. bei Erdos (1970), Gray (1975) und Wright & Barnard (1975).

Bei geschlossenen Fragen sind Kästchen oder Kreise vorzusehen, die der Vp anzeigen, wo sie ihre Markierung anzubringen hat. Handelt es sich um Fragenserien, sollten diese Kästchen bzw. Kreise eine klare graphische Anordnung (z.B. in einer Reihe untereinander) erhalten (Richter 1969) und zur Vermeidung von Verwechslungen nicht zu weit vom Text der Antworten entfernt sein (Wright & Barnard 1975). Richter (1969) fordert darüber hinaus, bei der Wahl der Größe für die Kästchen bzw. Kreise auf Besonderheiten der jeweiligen Zielpopulation Rücksicht zu nehmen. So seien bei älteren Menschen größere Kästchen bzw. Kreise erforderlich, aber auch z.B. Architekten unterschieden sich von z.B. Elektroingenieuren erheblich in der Größe der Kreuze, was bei ersteren im Interesse der Unmißverständlichkeit der Markierungen größere Kästchen bzw. Kreise erforderlich mache.

Hartley et al. (1977) untersuchten den Einfluß der Reihenfolge und genauen Anordnung von Antworttext, Kästchen und Codeziffern experimentell und stellten für die vier von ihnen verwendeten Varianten keine Auswirkungen auf das Antwortverhalten fest. Gewisse Unterschiede ergaben sich beim Zeitbedarf für die Erstellung des Fragebogenentwurfs, bei den Kosten für den Drucksatz und bei den Ablochkosten.

Zur Veranschaulichung und Verdeutlichung von Situationen bzw. Zusammenhängen, für die Beurteilungen oder Bewertungen erfragt werden sollen, z.T. aber auch im Interesse des Abwechslungsreichtums (vgl. 4.2), wird die Verwendung bildlicher Vorlagen empfohlen (z.B. Noelle 1963). Karmasin & Karmasin (1977) stellen jedoch fest, daß dabei mit subtilen, im einzelnen überwiegend nicht bekannten Einflüssen auf die Antworten zu rechnen sei. So werde z.B. ein Mann mit Hut innerhalb einer solchen bildlichen Vorlage als konservativer und besser situiert eingeschätzt als ein Mann ohne Hut, eine Hausfrau

mit sehr langem Haar gelte als weniger kompetent im Vergleich zu einer Frau mit kurzem Haar. In Abhängigkeit davon, wie bestimmte Rollenträger bzw. die Vertreter bestimmter Meinungen in den bildlichen Vorlagen dargestellt werden, sind dadurch Einflüsse auf die Antworten zu erwarten. Ring (1975) konnte Veränderungen in der Wahlhäufigkeit für bestimmte Statements in einer Größenordnung von 5% - 10% nachweisen, wenn die Zuordnung der Statements zu (stilisierten) Personen in einer bildlichen Vorlage vertauscht wurde. Nach Anlage dieser Untersuchung kann allerdings nicht entschieden werden, ob es sich dabei um Positionseffekte oder Einflüsse der Darstellungsweise der Personen handelt.

Bei der Erarbeitung des Fragebogen-Layouts muß auch festgelegt werden, wieviel *Antwortraum* bei offenen Fragen für die Eintragung der Antworten vorgesehen werden soll. Payne (1951) und Goode & Hatt (1972) berichten - gestützt auf entsprechende Erfahrungen - einen Anstieg der Antwortlänge mit Vergrößerung des für die Antworten vorgegebenen Raumes. Dies gelte einmal für schriftliche Befragungen, bei denen die Vp sieht, wieviel von ihr als Antwort erwartet wird, zum anderen aber auch für mündliche Befragungen, wobei ungeklärt sei, ob der Interviewer die Antworten ausführlicher protokolliere oder die Vp z.B. durch stärkeres Insistieren des Interviewers tatsächlich ausführlicher antworte. Diese Frage griff Haase (1978) auf. Er ließ die Antworten der Vpn auf Tonband aufnehmen und stellte fest, daß - gemessen an der Zahl der Wörter - die Antworten bei Vergrößerung des für die Eintragung vorgesehenen Raumes tatsächlich länger wurden. Vom Antwortinhalt (Zahl der enthaltenen Antwortkategorien) her war die Ausführlichkeit der Antworten jedoch nicht unterschiedlich. Außerdem bestand eine Abhängigkeit vom Frageninhalt: Ein Anstieg der Antwortlänge (Wortanzahl) durch Vergrößerung des Antwortraumes konnte nur für Fragen nach Merkmalen einer kurzzeitig dargebotenen Anzeige und nach ‚Gefühlen', die die Vpn mit dieser Anzeige verbinden, nicht aber für eine Frage nach bekannten Markennamen für ein bestimmtes Produkt aufgezeigt werden.

Bei schriftlichen Befragungen fand Tränkle (1974) Hinweise darauf, daß Antworten auch auf inhaltlicher Ebene (Zahl enthaltener Kategorien) ausführlicher waren, wenn 8 statt nur 3 Zeilen für die Eintragung der Antwort vorgesehen waren. Einflüsse des für Antworten vorgesehenen Raumes auf Antworten zu offenen Fragen scheinen also tatsächlich zu existieren, allerdings nur für bestimmte Frageninhalte, für einen bestimmten Variationsbereich des Antwortraumes (Haase 1978) und möglicherweise eher für die Form als für den Inhalt der Antworten.

Eine neuere, erst durch Einsatz von Textverarbeitungsanlagen realisierbar gewordene Entwicklung im Bereich der Fragebogengestaltung ist die *Individualisierung und Personalisierung* von Fragebogen. Für unpersönlich-schriftliche (postalische) Befragungen berichtet Perreault (1975) günstige Einflüsse auf den

Rücklauf, wenn der Fragebogen (scheinbar) individuell maschinenschriftlich erstellt, evtl. mit Namen und Adresse der Vp und mit dem Hinweis versehen ist, daß er in dieser Form nur an sie verschickt worden sei. Eingehender wird der Einsatz der Personalisierung im Interesse der Rücklaufsteigerung bei Erdos (1970) behandelt. In Abhängigkeit vom jeweiligen Befragungsgegenstand ist allerdings auch zu bedenken, daß die Personalisierung, wenn sie ihr Ziel erreicht, die extremste Form der Nicht-Anonymität und insofern ein ‚zweischneidiges Schwert' (Linsky 1975) ist.

6. Weitere Aspekte für die Konstruktion von Fragebogen

6.1 Anonymität des Befragten und Vertraulichkeit der Antworten

Als ‚anonym' wird eine Befragung dann bezeichnet, wenn es prinzipiell nicht möglich ist, ausgehend vom Fragebogen den Befragten zu identifizieren. Ist eine Befragung nicht anonym, aber vertraulich, so ist die Identität des Befragten zwar bekannt, wird aber gegenüber Dritten geheimgehalten (Dickson et al. 1977). Während die Zusicherung der Vertraulichkeit im Zusammenhang mit einer Befragung fast eine Selbstverständlichkeit zu sein scheint, wird die Notwendigkeit der Anonymität unterschiedlich beurteilt. Für persönlich mündliche Befragungen wird sie höchstens für einzelne Fragen angestrebt (vgl. z.B. die in 2.3 erwähnte ‚Urnentechnik'), obschon sie objektiv auch für das ganze Interview dadurch gewährleistet werden könnte, daß ein Interviewer zahlreiche Interviews durchführt und auf die Kennzeichnung der Fragebogen verzichtet wird. Die für das Antwortverhalten der Vp einzig maßgebliche *erlebte* Anonymität wird allerdings für persönlich-mündliche Befragungen kaum zu erreichen sein. Demzufolge beschränkt sich die Diskussion auch auf schriftliche, insbesondere unpersönlich-schriftliche, z.B. postalische Befragungen.

Einige Autoren berichten für diese Befragungsform bedeutsame Beantwortungsunterschiede in Abhängigkeit von der Anonymität bzw. Nicht-Anonymität der Befragten. Knudsen et al. (1967) fanden, daß in persönlichen Befragungen restriktivere Normen betreffend den vorehelichen Geschlechtsverkehr vertreten wurden als in unpersönlichen und anonymen Befragungen. Auch Fuller (1974) und Bradburn & Sudman (1979) berichten Antwortverzerrungen nach Maßgabe sozialer Erwünschtheit bei nicht-anonymer im Vergleich zu anonymer Befragung. Als Hinweise in dieser Richtung könnten auch die Ergebnisse von Taietz (1972) angesehen werden, der bei der Befragung älterer Menschen nach ihren Lebensverhältnissen erhebliche Verschiebungen in den Antworten dann erhielt, wenn eine dritte Person beim Interview anwesend war (vgl. auch Bradburn & Sudman 1979).

Verzerrungen in genau entgegengesetzte Richtung fanden Epperson & Peck (1977) in einer Untersuchung zur Evaluation von Driver-Improvement-Programmen (vgl. dazu Spoerer 1979). Hier fanden sich signifikant mehr negative Kommentare der Teilnehmer, wenn die Befragung nicht anonym durchgeführt wurde. Insgesamt sind die Nachweise bedeutsamer Antwortunterschiede in Abhängigkeit von der Anonymität jedoch eher spärlich. Kepes & True (1967) und auch Fuller (1974) kommen bei der Sichtung empirischer Befunde zu dem Ergebnis, daß der Einfluß der Nicht-Anonymität auf Antworten eher nur befürchtet als real sei. Die erstgenannten Autoren sehen ihn - außer in einigen ziemlich speziellen Situationen - vor allem dann, wenn die Vp eigens und explizit auf die Namentlichkeit hingewiesen wird, wozu in der Regel aber keine Notwendigkeit besteht. Daß den Vpn in vielen Fällen das Nicht-Vorliegen von Anonymität bzw. Vertraulichkeit gar nicht bewußt ist und der Einfluß von Anonymität bzw. Vertraulichkeit zutreffend nur nach entsprechendem explizitem Hinweis abgeschätzt werden kann, betonen auch Futrell & Swan (1977). Sie fanden zwischen anonymer und nicht-anonymer, aber vertraulicher postalischer Befragung keinerlei Unterschiede und sehen in der Anonymität keine Vorteile, wenn Untersucher und Auftraggeber nicht identisch sind und den Vpn Vertraulichkeit zugesichert werden kann, Butler (1973) ließ Fragen, die sich unter anderem auf Drogenkonsum bezogen, von Experten hinsichtlich ihrer Anfälligkeit für Antwortverzerrungen bei Verwendung in nicht-anonymer Befragung skalieren, bevor er sie Kadetten einer Militärakademie teils anonym, teils nicht-anonym zur Beantwortung vorlegte. Er fand bei Fragen, die die Experten als ‚unempfindlich' eingestuft hatten, erwartungsgemäß keine Beantwortungsunterschiede, wider Erwarten unterschieden sich anonyme und nicht-anonyme Antworten aber auch bei den als ‚empfindlich' klassifizierten Fragen nicht. Neben anderen möglichen Erklärungen könnte auch hier die den nicht-anonym antwortenden Vpn gegebene Vertraulichkeitszusage zur Vermeidung von Verzerrungen ausgereicht haben. Keine Unterschiede zwischen anonymen und nicht-anonymen Antworten von Lehrern auf Fragen zur Beurteilung der Notwendigkeit gewerkschaftlicher Organisierung und der eigenen Streikbereitschaft fand auch Wildman (1977). Andererseits berichtet er aber, daß im Rahmen dieser postalischen Befragung 12% der nicht-anonymen Vpn die Identifikationsnummern auf ihren Antwortbogen vor der Rücksendung unkenntlich gemacht hatten, was sich wohl nur auf ein Bedürfnis nach Anonymität zurückführen läßt.

Einflüsse der Anonymität auf den Rücklauf in unpersönlich-schriftlichen (postalischen) Befragungen sind nach Richter (1969) zwar je nach Zielpopulation unterschiedlich, insgesamt aber nicht ‚durchschlagend'. Bei Bradburn & Sudman (1979) fanden sich geringe, bei Wildman (1977) keinerlei Unterschiede im Rücklauf zwischen anonym und nicht-anonym befragten Vpn. Fuller (1974) berichtet sogar - abweichend von der landläufigen Erwartung - bei Nicht-Anonymität einen höheren Rücklauf (evtl. ein Personalisierungseffekt, vgl. 5.).

Während bei mündlichen und persönlichen schriftlichen Befragungen auch andere Gründe (z.B. die Notwendigkeit der Kontrolle von Interviewern) für den Verzicht auf Anonymität in Betracht kommen, ist das Interesse an der Identifizierbarkeit der Befragten in unpersönlich-schriftlichen (postalischen) Befragungen weitgehend in dem Wunsch nach Kontrollierbarkeit und gezielter Beeinflussung des Rücklaufs begründet. Einerseits sprechen Kostenerwägungen, andererseits aber auch die Gefahr von Doppel-Beantwortungen dagegen, die Fragebogen mehrfach an alle Vpn zu verschicken. Die Möglichkeit einer gezielten Erinnerung der Nicht-Beantworter besteht aber natürlich nur bei Identifizierbarkeit der Rückläufe. Um dennoch die vermuteten Vorteile der Anonymität nutzen zu können, hat in den USA die unsichtbare Kennzeichnung der Fragebogen eine weite Verbreitung gefunden, eine Praxis, die aus ethischen und juristischen Gründen zweifellos abzulehnen ist (vgl. Dickson et al. 1977). Alternativen, die eine Kontrolle der Rückläufe trotz strikter Anonymität gestatten, beschreibt z.B. Wieken (1974). So kann man dem Fragebogen eine mit der Adresse der Vp als Absender versehene frankierte Postkarte beifügen und die Vp bitten, diese gleichzeitig mit, aber getrennt von dem nicht gekennzeichneten Fragebogen zurückzuschicken, damit der Untersucher weiß, daß, aber nicht was sie geantwortet hat (Linsky 1975).

6.2 Spezielle Probleme bei unpersönlich-schriftlichen Befragungen

Unpersönlich-schriftliche Befragungen sind dadurch gekennzeichnet, daß ein Fragebogen in Abwesenheit des Interviewers bearbeitet wird. Der Fragebogen kann dem Befragten persönlich übergeben oder z.B. mit der Post zugeschickt worden sein. Diese letztgenannte Form der Befragung, die sogenannte postalische Befragung, erfreut sich aus mehreren Gründen vergleichsweise großer Beliebtheit, deren wichtigster die relativ geringen Kosten sein dürften (Stroschein 1965, Richter 1969, Goode & Hatt 1972, Wieken 1974). Allerdings sind zum Zwecke einigermaßen akzeptabler Stichprobenausschöpfungen fast immer mehrere Befragungswellen oder Erinnerungsschreiben notwendig, so daß den niedrigen Kosten ein vergleichsweise hoher Zeitbedarf (selten weniger als 6-8 Wochen, vgl. z.B. Buchner 1968) für die Datenerhebung gegenübersteht. Einerseits eignet sich die Methode damit für die Gewinnung aktueller Daten prinzipiell nicht, andererseits ist sie besonders anfällig gegenüber unvorhergesehenen, während der langen Erhebungsphase wirksam werdenden Einflüssen (z.B. Veröffentlichungen zum Thema).

Sachliche Vorteile der postalischen Befragung liegen für bestimmte Fragestellungen (wie bei allen schriftlichen Befragungsformen) in ihrem unpersönlichen und gegebenenfalls anonymen Charakter, der Antwortverzerrungen etwa nach Maßgabe sozialer Erwünschtheit weniger wahrscheinlich macht (Tränkle

1974). So berichtet etwa Friedrich (1970), daß Antworten in schriftlichen Befragungen weniger stark gesellschaftlichen Normen entsprechen als in mündlichen Interviews. Metzner & Mann (1952) erhielten für die Zufriedenheit von Arbeitnehmern mit ihren Vorgesetzten in schriftlichen verglichen mit mündlichen Befragungen ungünstigere Antworten. Auch Linsky (1975) sieht in der geringeren sozialen Kontrolle, der das Antwortverhalten unterliegt, einen wesentlichen Vorteil der schriftlichen Befragung.

Weitere Vorteile können - je nach Fragestellung und Zielpopulation - auch darin liegen, daß zur Beantwortung der Fragen Unterlagen herangezogen werden und an der Beantwortung mehrere Personen mitwirken können (Linsky 1975). In den meisten Fällen aber wird die bei der unpersönlichen Befragung prinzipiell fehlende Möglichkeit der Kontrolle von Beantwortungsperson, Beantwortungssituation, Beantwortungszeitpunkt und Reihenfolge der Beantwortung der Fragen als Nachteil betrachtet werden müssen.

Ein noch gravierenderer Nachteil der unpersönlichen, meist postalisch durchgeführten Befragung liegt in den relativ hohen Anforderungen, die sie an die Befragten stellt und die mindestens für den Teil der Bevölkerung, den Scheuch (1973) den ‚funktionellen Analphabeten' zuordnet, zu hoch sein dürften. Auch Kreutz & Titscher (1974, 60) stellen fest, daß „. . . in weiten Kreisen der Bevölkerung Angst vor Rechtschreibfehlern und Schwierigkeiten bei der schriftlichen Formulierung bestehen . . .", und halten mindestens offene Fragen in unpersönlich-schriftlichen Befragungen dann für kontraindiziert, wenn die Zielpopulation nicht z.B. durch Bildung bzw. Beruf sprachlich besonders geübt ist.

Da der Rücklauf in postalischen Befragungen positiv mit der sozialen Schicht, dem Bildungs-, Berufs- und Einkommensniveau korreliert (z.B. Richter 1967, 1969, Goode & Hatt 1972, Wieken 1974, Binder et al. 1979), ist bei inhomogenen Stichproben (z.B. Bevölkerungsstichproben) mit systematischen Verzerrungen dadurch zu rechnen, daß Angehörige unterer sozialer Schichten mit niedrigem Bildungs-, Berufs- und Einkommensniveau in der Gruppe der antwortenden Vpn unterrepräsentiert sind. Unpersönlich-schriftliche (postalische) Befragungen sollten deshalb nur für homogene lese- und schreibgewandte Populationen in Betracht gezogen werden. Nach Kish & Barnes (1973) eignet sich die postalische Befragung außerdem nicht für Befragungsinhalte, die in der Zielpopulation kontrovers eingeschätzt werden: Der Rücklauf erwies sich als umgekehrt proportional der Strittigkeit der Inhalte.

Auch wenn z.B. Mc Donagh & Rosenblum (1965) in einer mündlichen Nachbefragung von Antwortern und Nicht-Antwortern einer vorangegangenen postalischen Befragung keinerlei Beantwortungsunterschiede feststellen konnten und deshalb annehmen, das Problem der Irrepräsentativität der Antworter für die gesamte Population werde üblicherweise überschätzt, läßt sich die

Gefahr der Irrepräsentativität natürlich nie prinzipiell ausschließen, Binder et al. (1979) etwa fanden beträchtliche Unterschiede zwischen Antwortern und Nicht-Antwortern in demographischen und Persönlichkeitsmerkmalen. Es ist deshalb von besonderer Wichtigkeit, durch geeignete Anlage der Untersuchung (z.B. Vorkontakten der Vpn; Gestaltung von Begleitschreiben, Fragebogen; Rückumschlag; mehrfache Erinnerungsschreiben) für eine möglichst hohe Stichprobenausschöpfung zu sorgen. Entsprechende Hinweise und empirische Ergebnisse finden sich z.B. bei Richter (1969), Alutto (1970), Erdos (1970), Hendrick et al. (1972), Kish & Barnes (1973), Wieken (1974) und Sieber (1979). Linsky (1975) hat eine ausführliche Zusammenstellung empirischer Befunde zur Frage der Beeinflußbarkeit des Rücklaufs erarbeitet, die u.a. erkennen läßt, daß es kaum Befunde betreffend den Einfluß der Gestaltung des Fragebogens auf den Rücklauf gibt (vgl. 5.).

Neben und evtl. zusätzlich zu dem Bemühen um einen möglichst hohen Rücklauf werden gelegentlich auch Korrekturen der Ergebnisse zum Ausgleich etwa bestehender Irrepräsentativität vorgenommen. Solche Korrekturen beruhen natürlich auf Annahmen betreffend das potentielle Antwortverhalten der Nicht-Antwortenden. Meist gehen sie davon aus, daß die Vpn, die zuletzt geantwortet haben, ‚Beinahe-Nicht-Antworter' sind, und verwenden deren Antwortverhalten zur Schätzung des Verhaltens der Nicht-Antworter. Evtl. wird auch versucht, einen Trend, der sich im Antwortverhalten von den frühesten zu den letzten Rückläufen hin zeigt, auf die Nicht-Antworter zu extrapolieren. Überlegungen zur Repräsentativitätskorrektur finden sich u.a. bei Richter (1967, 1969), Buchner (1968), Erdos (1970) und Wieken (1974).

6.3 Erprobung und Überarbeitung des Fragebogenentwurfs

Itemanalysen, Itemselektion und Konstruktion abgeleiteter Variabler (z.B. Summierung von Antworten) erfolgen bei diagnostischen Fragebogen wie bei jedem Test auf der Grundlage eines bestimmten Meßmodells und Validitätskonzeptes (vgl. z.B. Lienert 1969, Fischer 1974, Wottawa 1980). Darüber hinaus sind jedoch auch fragebogenspezifische Gütekriterien (Verstehbarkeit, Ambiguität, soziale Erwünschtheit der Items) und Eigentümlichkeiten (z.B. veränderte Bedeutung der Itemschwierigkeit) zu beachten (Janke 1973). Hierauf soll an dieser Stelle nicht näher eingegangen werden.

Bei Fragebogen mit sozialwissenschaftlicher bzw. demoskopischer Zielsetzung wird die Notwendigkeit des Pretests (Karmasin & Karmasin 1977 fordern dafür sogar 100 Vpn) und der Revision des Fragebogens zwar allgemein anerkannt bzw. hervorgehoben, doch charakterisieren Cannell et al. (1977, 27) die in der Praxis übliche Vorgehensweise (in der Übersetzung des Verfassers) folgendermaßen:

,,Normalerweise erstellt man den Fragebogenentwurf am Schreibtisch und schickt dann eine Gruppe von Interviewern damit ins Feld. Danach gibt es eine Konferenz (oder eine Serie von Konferenzen), auf der Forscher und Interviewer den Fragebogen diskutieren. Man hört dabei Aussagen wie ‚. . . diese Frage scheint gut zu funktionieren . . .' oder der Interviewer sagt: ‚. . . Ich glaube nicht, daß die Befragten diese Frage wirklich verstanden haben . . .'. Auf der Grundlage derart subjektiver Bewertungen werden Fragebogen üblicherweise entwickelt." Selbstverständlich können die Erfahrungen der Interviewer einen wichtigen Beitrag zur Revision des Fragebogens leisten, nur sollte es nicht deren einzige Grundlage sein. Guski et al. (1978) beschreiben beispielhaft die Konstruktion eines sozialwissenschaftlichen Fragebogens zur Erfassung von Auswirkungen des Umweltlärms. Ausgehend von den Ergebnissen einer Vorstudie (Explorationen mit 30 Vpn), von einer Inhaltsanalyse der Beschwerden über Lärmbelästigung, die bei Behörden eingegangen waren, und von bereits existierenden Fragebogen zum Thema wurde ein Fragebogenentwurf erstellt und einem Pretest an 40 Vpn unterworfen. Statistische Itemanalysen (dazu können wie bei der Konstruktion diagnostischer Fragebogen u.a. Verteilungs-, Schwierigkeits-, Trennschärfe- und Interkorrelationsanalysen gehören; vgl. Berk & Griesemer 1976) und Interviewererfahrungen bildeten die Grundlage einer Revision des Fragebogens für die Hauptuntersuchung an über 600 Vpn. Damit waren die methodologischen Bemühungen um den Fragebogen allerdings nicht abgeschlossen, vielmehr wurden die Definitionen abgeleiteter Variabler (z.B. die Summierungen von Reaktionen auf verschiedene Items) mit den Daten der Hauptuntersuchung (zur Prüfung der Stabilität gegenüber einer Variation der Stichprobe meist getrennt für zwei Zufallshälften der Stichprobe) jeweils empirisch abgesichert, vor allem mittels Cluster- und Faktorenanalysen. Wegen der hierfür erforderlichen hohen Vpn-Zahl wäre es unrealistisch, solche Analysen schon im Stadium des Pretests zu verlangen, vielmehr wird man die Fragebogenkonstruktion und -Überprüfung als einen Prozeß auffassen müssen, der nie abgeschlossen, sondern höchstens abgebrochen werden kann.

Die Aufgaben, die der Pretest erfüllen kann und erfüllen muß, nämlich die Überprüfung von Fragenformulierungen, Fragebogenaufbau und -gestaltung, werden um so wichtiger, je weniger Kompensationsmöglichkeiten für Mängel des Fragebogens in der Befragungssituation selbst vorhanden sind. Von besonderer Bedeutung ist die Erprobung des Fragebogens demnach für alle nichtpersönlichen Befragungen, also besonders für die postalische Befragung. Richter (1969) spricht in diesem Zusammenhang von der Notwendigkeit, den Fragebogen im ‚Putzfrauentest', d.h. durch Anwendung bei den sprachlich und intellektuell am wenigsten differenzierten Vpn der Zielpopulation zu erproben.

Für diagnostische Fragebogen mit längerer Lebensdauer sind - wie für jeden Test - kontinuierliche Kontrolluntersuchungen erforderlich (Lennertz 1973).

So weist z.B. Strong (1962) auf die Notwendigkeit hin, veraltende Inhalte von Items (Persönlichkeiten, Buch- und Filmtitel etc.) entweder grundsätzlich zu meiden oder aber häufigere Revisionen und Aktualisierungen der Fragebogen durchzuführen. Ash & Edgell (1975) demonstrierten die Nichtübereinstimmung des sprachlichen Niveaus des Position-Analysis-Questionnaire (PAQ) von Mc Cormick (vgl. Mc Cormick et al. 1965) mit demjenigen der tatsächlichen Anwender und machten deutlich, daß auch Änderungen der Zielpopulation in Rechnung zu stellen sind.

Die Kontrolle des Fragebogenentwurfs muß sich sodann natürlich auf Fragen der Reliabilität und Validität (bzw. Generalisierbarkeit im Sinne von Cronbach et al. 1972) erstrecken. Wie für alle Tests so ist auch für diagnostische Fragebogen unbestritten, daß Aussagen über ihre Güte nur in bezug auf ein bestimmtes Meßmodell, auf ein bestimmtes Validitätskonzept und evtl. auf eine bestimmte Population möglich bzw. sinnvoll sind. Ebenso ist es für sozialwissenschaftliche bzw. demoskopische Anwendungen abwegig, die Qualität von Fragebogenerhebung bzw. Interview allgemein feststellen zu wollen, wie dies etwa Friedrich (1963, 1966) und Förster (1967) für schriftliche und Fisseni (1974) für mündliche Befragungen zu tun versuchen (vgl. auch Sieber 1979). Aussagen sind auch hier nur möglich für die Methode bezogen auf einen Gegenstand und eine bestimmte Population. Bei mündlichen und persönlich-schriftlichen Befragungen sind außer dem Fragebogen die Interviewer, bei unpersönlich-schriftlichen Befragungen die Techniken und der Grad der Stichprobenausschöpfung zentrale Bestandteile der Methode. Nachweise hoher Objektivität, Reliabilität und Validität stellen hier bestenfalls Existenzbeweise dar.

7. Zukünftige Entwicklung im Bereich der Fragebogenkonstruktion

Um je nach gewähltem Validitätskonzept und Meßmodell (vgl. 1.1.2) die Konzeption einer Frage auf empirisch gesicherter Basis entwickeln zu können, ist es erforderlich, das Wissen über den Beantwortungsprozeß und die Determinanten der Antwort (vgl. 1.2) zu erweitern. Erhebliche Wissenslücken bestehen sodann im Bereich der sprachlichen Formulierung der Frage (3.2), der Fragenreihenfolge (4.) und vor allem der Auswirkungen der äußeren Gestaltung des Fragebogens (5.) auf das Beantwortungsverhalten.

Für *demoskopische (sozialwissenschaftliche)* Fragebogen zeichnet sich durch die leichtere Verfügbarkeit elektronischer Datenverarbeitungsanlagen ein Verschwinden des für alle Vpn einheitlichen Fragebogens zugunsten einer größeren Zahl von Fragebogenvarianten mit variierter Reihenfolge der Fragen, variierten Frageformulierungen, variierter äußerer Gestaltung bis hin zum indivi-

dualisierten und möglicherweise personalisierten Fragebogen ab (Perreault 1975). Dabei ist es durchaus auch möglich, unter Verwendung von Vorinformationen über den Befragten eine ganz spezielle Fragenzusammenstellung zu konzipieren und damit Filterungen und Verzweigungen, wie sie in traditionellen Fragebogen erforderlich sind, entbehrlich zu machen, was besonders für unpersönlich-schriftliche (postalische) Befragungen die möglichen Befragungsinhalte erheblich ausweiten dürfte. Darüber hinaus lassen sich unter Nutzung elektronischer Datenverarbeitungsanlagen Fragenpools aufbauen, die eine rasche Ad-hoc-Konstruktion von Fragebogen für bestimmte Anwendungen ermöglichen (Doyle & Wattawa 1977).

Inwieweit neue elektronische Medien, wie Videotext und Telekommunikation, auch die Durchführung von Befragungen nachhaltig verändern werden, ist derzeit nicht abzuschätzen. In Anlehnung an die Erfahrungen mit Telefon-Interviews ist jedoch zu vermuten, daß es das Bildschirm-Interview für bestimmte Untersuchungen geben wird, daß es die traditionellen Befragungsformen jedoch nicht wird verdrängen können.

Die statistische Auswertung von Fragebogendaten, die heute noch überwiegend einzelfragenorientiert erfolgt, wird sich zunehmend der angemesseneren multivariaten Analyse- und Testverfahren bedienen (vgl. Whitney & Feldt 1973).

Im Bereich *diagnostischer Fragebogen* werden sich die test- und meßtheoretischen Grundlagen weiterentwickeln. Dabei dürfte einerseits dem ordinalen Charakter von Fragebogendaten stärker Rechnung getragen, andererseits dürften aber auch Versuche unternommen werden, die Datenqualität in Richtung auf metrische Eigenschaften zu verbessern. Dabei haben mehrkategorielle probabilistische Modelle gerade für Fragebogen große Bedeutung.

Erhebliche Möglichkeiten scheinen auch in der Anwendung der Methoden individualisierten (antwortabhängigen) Testens im Falle von Fragebogen zu liegen; Versuche in dieser Richtung beschreibt z.B. Hornke (1979). In gewisser Hinsicht handelt es sich dabei um die Realisierung der auch für demoskopische Fragebogen gebräuchlichen Techniken der Filterung und Verzweigung: Während in einem herkömmlichen diagnostischen Fragebogen jeder Proband alle Fragen zu bearbeiten hat, werden beim antwortabhängigen Test diejenigen Items nicht dargeboten, die zur (zuverlässigen) Schätzung des Ortes des Probanden auf der interessierenden Dimension nichts Wesentliches beitragen.

Außer zur Datengewinnung im Bereich sozialwissenschaftlicher Fragestellungen und zu diagnostischen Zwecken sind Fragebogen auch mit dem Ziel der Änderung von Einstellungen (Dillehay & Jernigan 1970) und mit therapeutischer Intention als Hilfsmittel bei der Selbsterfahrung (Hendrix 1978) eingesetzt worden. Es ist schwer abzuschätzen, ob sich diese Anwendungen bewäh-

ren und vermehren und ob sich weitere Einsatzmöglichkeiten eröffnen werden.

Umgekehrt dürften Fragebogen überall dort ihre Berechtigung verlieren, wo direktere und objektivere Methoden der Datengewinnung verfügbar werden. So können bestimmte Daten aus diagnostischen Fragebogen möglicherweise durch physiologische Messungen ersetzt werden: Statt nach Schlafqualität zu fragen, kann man sie u.U. dem EEG entnehmen. *Demoskopische bzw. sozialwissenschaftliche* Fragebogen dürften in den Bereichen entbehrlich werden, in denen vorhandene Dateien abgefragt werden können (z.B. muß der Führerscheinbesitz z.Z. noch durch Befragung erhoben werden, nach Aufbau einer entsprechenden Datei würde diese Notwendigkeit entfallen).

Für die im Bereich der Diagnostik wie der sozialwissenschaftlichen Datenerhebung wichtigen Beurteilungen und Bewertungen durch Personen sind zwar Alternativen zur hergebrachten Form des Fragebogens, nicht aber zur Methode der Befragung erkennbar.

Literatur

Adams, J. S. 1956. An experiment on question and response bias. Public Opinion Quarterly, 20, 593-598.

Alutto, J. A. 1970. Some dynamics of questionnaire completion and return among professional and managerial personnel. Journal of Applied Psychology, 54, 430-432.

Anastasi, A. 1968. Psychological testing. London: Macmillan.

Andersen, E. B. 1973. Conditional inference for multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 26, 31-44.

Anger, H. 1969. Befragung und Erhebung. In: Graumann, C. F. (Hrsg.): Handbuch der Psychologie, Bd. 7: Sozialpsychologie, 1. Halbbd. Göttingen: Hogrefe.

Ash, R. & Edgell, S. L. 1975. A note on the readability of the position analysis questionnaire (PAQ). Journal of Applied Psychology, 60, 775-776.

Atteslander, P. 1971. Methoden der empirischen Sozialforschung. Berlin: De Gruyter.

Barton, A. H. 1958. Asking the embarrassing question. Public Opinion Quarterly, 22, 67-68.

Behrens, K. C. (Hrsg.) 1974. Handbuch der Marktforschung. Wiesbaden: Gabler.

Belson, W. A. 1966. The effect of reversing the presentation order of verbal rating scales. Journal of Advertising Research, 6, 30-37.

Berdie, D. R. 1973. Questionnaire length and response rate. Journal of Applied Psychology, 58, 278-280.

Berg, I. A. 1967. Response set in personality assessment. Chicago: Aldine.

Berk, R. A. & Griesemer, H. A. 1976. Iteman: An item analysis program for tests, questionnaires and scales. Educational and Psychological Measurement, 36, **189-191.**

Binder, J., Sieber, M. & Angst, J. 1979. Verzerrungen bei postalischer Befragung: das Problem der Nichtantworter. Zeitschrift für experimentelle und angewandte Psychologie, 24, 53-71.

Block, J. 1965. The challenge of response sets. New York: Appleton Century Crofts.

Bradburn, N. M. & Mason, W. M. 1964. The effect of question order on response. Journal of Marketing Research, 1, 57-61.

Bradburn, N. M. & Sudman, S. 1979. Improving interview method and questionnaire design. London: Jossey Bass.

Buchner, D. 1968. Probleme und Antwortmuster bei postalischen Ärztebefragungen. Der Marktforscher, 7, 178-181.

Burisch, M. 1976. Konstruktionsstrategien für multidimensionale Persönlichkeitsfragebögen. Hamburg: Phil. Diss.

Butler, R. P. 1973. Effects of signed and unsigned questionnaires for both sensitive and nonsensitive items. Journal of Applied Psychology, 57, 348-349.

Cahalan, D., Tamulonis, V. & Verner, H. W. 1947. Interviewer bias involved in certain types of opinion survey questions. International Journal of Opinion and Attitude Research, 1, 63-77.

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

Cannell, C. F. & Kahn, R. L. 1968. Interviewing. In: Lindzey, G. & Aronson, E. (eds.): Handbook of Social Psychology, Vol. 2. Reading: Addison-Wesley.

Cannell, C. F., Marquis, K. H. & Laurent, A. 1977. A summary of studies of interviewing methodology. Rockville: U.S. Department of Health, Education and Welfare.

Carl, W. 1968. Eine Untersuchung zur Faktorenstruktur von Antworttendenzen bei Antwortskalen unterschiedlicher Stufenzahl. Zeitschrift für experimentelle und angewandte Psychologie, 15, 419-434.

Cataldo, E. F., Johnson, R. M., Kellstedt, L. A. & Milbrath, L. W. 1970. Card sorting as a technique for survey interviewing. Public Opinion Quarterly, 34, 202-215.

Cattell, R. B. 1974. How good is the modern questionnaire? General principles for evaluation. Journal of Personality Assessment, 38, 115-129.

Cattell, R. B., Eber, W. H. & Tatsuoka, M. M. 1970. Handbook for sixteen personality factor questionnaire. Champaign: Institute for Personality and Ability Testing.

Clauss, G. 1968. Zur Methodik von Schätzskalen in der empirischen Forschung. Probleme und Ergebnisse der Psychologie, 26, 7-53.

Cliff, N. 1977. Further study of cognitive processing models for inventory response. Applied Psychological Measurement, 1, 41-49.

Cliff, N., Bradley, P. & Girard, R. 1973. The investigation of cognitive models for inventory response. Multivariate Behavioral Research, 8, 407-425.

Coan, R. W. 1964. Facts, factors and artifacts: The quest for psychological meaning. Psychological Review, 71, 123-140.

Cohen, R. & Carl, W. 1964. Beantwortungsstereotypen (response sets) im Polaritätsprofil und ihre Beziehung zum Neurotizismus. Diagnostica, 10, 133-144.

Cronbach, L. J. 1970. Essentials of psychological testing. New York: Harper & Row.

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

Cronbach, L.J., Gleser, G. C., Nanda, H. & Rajaratnam, N. 1972. The dependability of behavioral measurement. New York: Wiley.

Crutchfield, R. S. & Gordon, D. A. 1947. Variations in respondent interpretations of an opinion poll question. International Journal of Opinion and Attitude Research, **1, 1-12.**

Damarin, F. 1970. A latent structure model for answering personal questions. Psychological Bulletin, 73, 23-40.

Dickson, J. P., Casey, M., Wyckoff, D. & Wynd, W. 1977. Invisible coding of survey questionnaires. Public Opinion Quarterly, 41, 100-106.

Dillehay, R. & Jernigan, L. R. 1970. The biased questionnaire as an instrument of opinion Change. Journal of Personality and Social Psychology, 15, 144-150.

Doyle, K. C. & Wattawa, S. 1977. Programs for the construction and analysis of custom questionnaires and rating scales. Educational and Psychological Measurement, 37, 237-239.

Edwards, A. L. 1957. Techniques of attitude scale construction. New York: Appleton Century Crofts.

Edwards, A. L. 1970. The measurement of personality traits by scales and inventories. New York: Holt, Rinehart & Winston.

Ehlers, T. 1973. Zur Effektivität der Kontrollen von Reaktionseinstellungen. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.

Ellis, A. 1947. A comparison of the use of direct and indirect phrasing in personality questionnaires. Psychological Monographs, 61, whole No. 284.

Epperson, W. V. & Peck, R. C. 1977. Questionnaires response bias as a function of respondent anonymity. Accident Analysis and Prevention, 9, 249-256.

Erdos, P. L. 1970. Professional mail surveys. New York: McGraw Hill.

Eysenck, H. J. 1953. Fragebogen als Meßmittel der Persönlichkeit. Zeitschrift für experimentelle und angewandte Psychologie, 1, 291-335.

Eysenck, H. J. 1956. Wege und Abwege der Psychologie. Reinbek: Rowohlt.

Falthzik, A. M. & Carroll, S. J. 1971. Rate of return for closed and opened questions in a mail questionnaire survey of industrial Organisation. Psychological Reports, 29, **1121-1122.**

Feger, H. 1974. Die Erfassung individueller Einstellungsstrukturen. Zeitschrift für Sozialpsychologie, 5, 242-254.

Fischer, G. 1974. Einführung in die Theorie psychologischer Tests. Bern: Huber.

Fiske, D. W. 1978. Strategies for personality research. San Francisco: Jossey Bass.

Fisseni, H. J. 1974. Zur Zuverlässigkeit von Interviews. Archiv für Psychologie, 126, 71-84.

Förster, P. 1967. Zu einigen methodischen Problemen der schriftlichen Befragung. Jugendforschung, 1/2, 39-67.

Friedrich, W. 1963. Die Befragungsmethode: ein notwendiges Arbeitsmittel der marxistischen Jugendforschung. Deutsche Zeitschrift für Philosophie, 10, **1230-1247.**

Friedrich, W. 1966. Zur Reliabilität von schriftlichen Befragungen. Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig, 15, 805-808.

Friedrich, W. (Hrsg.) 1971. Methoden der marxistisch-leninistischen Sozialforschung. Berlin: VEB Deutscher Verlag der Wissenschaften.

Friedrich, W. & Hennig, W. (Hrsg.) 1975. Der sozialwissenschaftliche Forschungsprozeß. Berlin: VEB Deutscher Verlag der Wissenschaften.

Friedrichs, J. 1973. Methoden empirischer Sozialforschung. Reinbek: Rowohlt.

Frisbie, B. & Sudman, S. 1968. The use of computers in coding free responses. Public Opinion Quarterly, 32, 216-232.

Fürntratt, E. 1969. Antworttendenzen in Fragebogen 1: Bejahungs- und Varianztendenzen. Psychologische Rundschau, 20, 1-18.

Fuller, C. 1974. Effect of anonymity on return rate and response bias in a mail survey. Journal of Applied Psychology, 59, 292-296.

Futrell, C. M. & Swan, J. E. 1977. Anonymity and response by salespeople to a mail questionnaire. Journal of Marketing Research, 14, 611-616.

Galtung, J. 1973. Theory and methods of social research. London: Allen & Unwin.

Getzels, J. W. 1954. The question-answer process: a conceptualisation and some derived hypotheses for empirical examination. Public Opinion Quarterly, 18, 80-91.

Goode, W. J. & Hatt, P. K. 1972. Die schriftliche Befragung. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.

Gray, A. 1975. Questionnaire typography and production. Applied Ergonomics, 6, 81-89.

Guilford, J. P. 1954. Psychometric methods. New York: McGraw Hill.

Guilford, J. P. 1965. Persönlichkeit. Weinheim: Beltz.

Gullahorn, J. E. & Gullahorn, J. T. 1963. An investigation of the effects of three factors on response to mail questionnaires. Public Opinion Quarterly, 27, **294-296.**

Gulliksen, H. 1950. Theory of mental tests. New York: Wiley.

Guski, R., Wichmann, U., Rohrmann, B. & Finke, H. O. 1978. Konstruktion und Anwendung eines Fragebogens zur sozialwissenschaftlichen Untersuchung der Auswirkungen von Umweltlärm. Zeitschrift für Sozialpsychologie, 9, 50-65.

Haase, H. 1978. Zum Einfluß des Fragebogen-Layouts auf Befragungsergebnisse. In: Hartmann, K. D. & Koeppler, K. (Hrsg.): Fortschritte der Marktpsychologie, Bd. 1. Frankfurt: Fachbuchhandlung für Psychologie.

Häcker, H., Schwenkmezger, P. & Utz, H. 1979. über die Verfälschbarkeit von Persönlichkeitsfragebogen und objektiven Persönlichkeitstests unter SD-Instruktionen und in einer Auslesesituation. Diagnostica, 25, 7-23.

Hampel, R. & Klinkhammer, F. 1978. Verfälschungstendenzen beim Freiburger Persönlichkeitsinventar in einer Bewerbungssituation. Psychologie und Praxis, 22, 58-69.

Hartley, J., Lindsey, D. & Burnhill, P. 1977. Alternatives in the typographic design of questionnaires. Journal of Occupational Psychology, 50, 299-304.

Hartmann, H. 1972. Empirische Sozialforschung. München: Juventa.

Hase, H. & Goldberg, R. 1967. Comparative validity of different strategies of constructing personality inventory scales. Psychological Bulletin, 67, 231-248.

Hathaway, S. R. & Mc Kinley, J. C. 1963. MMPI Saarbrücken, Handbuch. Bern: Huber.

Hayes, D. P. 1964. Item order on Guttman scales. American Journal of Sociology, 70, **51-58.**

Heller, D. & Krüger, P. 1976. Analyse dreistufig zu beantwortender Fragebogenitems. Psychologische Beiträge, 18, 431-442.

Hendrick, C., Borden, R., Giesen, M., Murray, E. J. & Seyfried, B. A. 1972. Effectiveness of ingratiation tactics in a cover letter on mail questionnaire response. Psychonomic Science, 26, 349-351.

Hendrix, L. 1978. Studying ourselves: the questionnaire as a teaching tool. Family Coordinator, 27, 47-54.

Hennig, W. 1971. Einige Fragen des Aufbaus von Interviewfragebogen und der Interviewerausbildung. In: Friedrich, W. (Hrsg.): Methoden der marxistisch-leninistischen Sozialforschung. Berlin: VEB Deutscher Verlag der Wissenschaften.

Herrmann, T. & Stäcker, K. H. 1969. Sprachpsychologische Beiträge zur Sozialpsychologie. In: Graumann, C. F. (Hrsg.): Handbuch der Psychologie, Bd. 7: Sozialpsychologie, 1. Halbbd. Göttingen: Hogrefe.

Hoeth, F. 1980. Antworttendenzen und ihre methodische Bedeutung für Befragungsverfahren. In: Hartmann, K. D. & Koeppler, K. (Hrsg.): Fortschritte der Marktpsychologie, Bd. 2. Frankfurt: Fachbuchhandlung für Psychologie.

Hoeth, F. & Köbler, V. 1967. Zusatzinstruktionen gegen Verfälschungstendenzen bei der Beantwortung von Persönlichkeitsfragebogen. Diagnostica, 13, 117-130.

Holm, K. 1974a. Theorie der Frage. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 26, 91-114.

Holm, **K.** 1974b. Theorie der Fragenbatterie. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 26, 316-341.

Holm, K. (Hrsg.) 1975a. Die Befragung, Bd. 1. München: Francke.

Holm, K. 1975b. Die Frage. In: Holm, K. (Hrsg.): Die Befragung, Bd. 1. München: Francke.

Hornick, C. W., James, L. R. & Jones, A. P. 1977. Empirical item keying versus a rational approach to analyzing psychological climate questionnaire. Applied Psychological Measurement, 1, 489-500.

Hornke, L. 1979. Konstruktion eines adaptiv-antwortabhängigen Fragebogens zur Erfassung der Prüfungsangst. Diagnostica, 25, 208-218.

Janke, W. 1973. Das Dilemma von Persönlichkeitsfragebogen. Einleitung des Symposiums über Konstruktion von Fragebogen. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.

Jannsen, J. P. 1978. Zur Validität und Reliabilität von Persönlichkeitsfragebogen in Ernstsituationen und beim Rollenspiel. Köln: TÜV Rheinland.

Jetzschmann, H., Kallabis, H., Schulz, R. & Taubert, H. (Hrsg.) 1966. Einführung in die soziologische Forschung. Berlin: VEB Deutscher Verlag der Wissenschaften.

Johnson, W. R., Sieveking, N. A. & Clanton, E. S. 1974. Effects of alternative positioning of open-ended questions in multiple-choice questionnaires. Journal of Applied Psychology, 59, 776-778.

Jonsson, C. O. 1957. Questionnaires and interviews. Stockholm: Almqvist.

Kahn, R. L. & Cannell, C. L. 1957. The dynamics of interviewing. New York: Wiley.

Kalinowsky-Czech, M. 1979. Assoziationen und Entscheidungsprozesse bei der Beantwortung von Persönlichkeitsfragebogenitems. Bonn: Unveröff. Dipl. Arbeit.

Kane, R. B. 1969. Computer generation of semantic-differential questionnaires. Educational and Psychological Measurement, **29, 191-192.**

Karmasin, F. & Karmasin, H. 1977. Einführung in die Methoden und Probleme der Umfrageforschung. Wien: Böhlau.

Keil, W. 1973. Reaktionseinstellungen und Fragebogenkonstruktion. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.

Kepes, S. Y. & True, J. E. 1967. Anonymity and attitudes toward work. Psychological Reports, 21, 353-356.

Kinsey, A. C., Pomeroy, J. E. & Martin, C. E. 1970. Das sexuelle Verhalten des Mannes. Frankfurt: Fischer Bücherei.

Kirschhofer-Bozenhardt, A. von & Kaplitza, G. 1975. Der Fragebogen. In: Holm, K. (Hrsg.): Die Befragung, Bd. 1. München: Francke.

Kish, G. B. & Barnes, J. 1973. Variables that effect return rate of mailed questionnaires. Journal of Clinical Psychology, 29, 98-100.

Knudsen, D. D., Pope, H. & Irish, D. P. 1967. Response differences to questions on sexual Standards: An interviewer - questionnaire comparison. Public Opinion Quarterly, 31, 290-297.

König, R. (Hrsg.) 1972. Das Interview. Köln: Kiepenheuer & Witsch.

König, R. (Hrsg.) 1974. Handbuch der empirischen Sozialforschung. Stuttgart: Enke.

Koolwijk, J. von 1968. Fragebogenprofile. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 20, 780-791.

Koolwijk, J. von 1969. ‘Unangenehme Fragen‘. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 21, 864-875.

Koolwijk, J. von & Wieken-Mayser, M. (Hrsg.) 1974. Techniken der empirischen Sozialforschung, Bd. 4: Erhebungsmethoden: Die Befragung. München/Wien: Oldenbourg.

Koomen, W. & Dijkstra, W. 1975. Effects of question length on verbal behavior in a bias-reduced interview Situation. European Journal of Social Psychology, 5, 399-403.

Kraut, A. I., Wolfson, A. D. & Rothenberg, A. 1975. Some effects of position on opinion survey items. Journal of Applied Psychology, 60, 774-776.

Kreutz, H. & Titscher, S. 1974. Die Konstruktion von Fragebögen. In: Koolwijk, J. von & Wieken-Mayser, M. (Hrsg.): Techniken der empirischen Sozialforschung, Bd. 4: Erhebungsmethoden: Die Befragung. München/Wien: Oldenbourg.

Kuncel, R. B. 1973. Response processes and relative location of subject and item. Educational and Psychological Measurement, 34, 743-755.

Lansing, J. B., Ginsberg, G. P. & Braaten, K. 1961. An investigation of response error. Urbana: University of Illinois.

Lantermann, E. D. & Gehlen, H. 1977. Skalierung von Items und Individuen unter Beachtung individueller Urteilsstrukturen. Zeitschrift für Sozialpsychologie, 8, 242-246.

Lazarsfeld, P. F. 1935. The art of asking ‚why‘. National Marketing Review 1. Zitiert nach: Maccoby, E. E. & Maccoby, N. 1972. Das Interview: Ein Werkzeug der Sozialforschung. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.

Lazarsfeld, P. F. & Barton, A. H. 1955. Some general principles of questionnaire classification. In: Lazarsfeld, P. F. & Rosenberg, M. (eds.): The language of social research. Glencoe: The Free Press.

Lennertz, E. 1973. Thesen zur Itemsammlung bei Persönlichkeitsfragebogen. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.

Lienert, G. A. 1969. Testaufbau und Testanalyse. Weinheim: Beltz.

Linsky, A. S. 1975. Stimulating responses to mailed questionnaires: A review. Public Opinion Quarterly, 39, 82-101.

Litwak, E. 1956. A classification of biased questions. American Journal of Sociology, **62, 182-186.**

Maccoby, E. E. & Maccoby, N. 1972. Das Interview: ein Werkzeug der Sozialforschung. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.

Magnussen, D. 1966. Introduction to test theory. Reading: Addison-Wesley.

Mauldin, W. P. & Marks, E. S. 1950. Problems of response in enumerative surveys. American Sociological Review, 15, 649-657.

Mayntz, R., Holm, K. & Hübner, P. 1971. Einführung in die Methoden der empirischen Soziologie. Opladen: Westdeutscher Verlag.

Mc Cormick, E. J., Jeanneret, P. R. & Mecham, R. C. 1969. The development and background of the position analysis questionnaire. Occupational Research Center Report No. 5. Lafayette: Purdue University Press.

Mc Donagh, E. C. & Rosenblum, A. L. 1965. A comparison of mail questionnaires and subsequent structured interviews. Public Opinion Quarterly, 29, 131-136.

Mc Kelvie, S. J. 1978. Graphic rating scales: how many categories? British Journal of Psychology, 69, 185-202.

Metzner, H. & Mann, F. 1952. A limited comparison of two methods of data collection: The fixed alternative questionnaire and the open-ended interview. American Sociological Review, 17, 486-491.

Metzner, H. & Mann, F. 1953. Effects of grouping related questions in questionnaires. Public Opinion Quarterly, 17, 136-141.

Mittenecker, E. 1971. Subjektive Tests zur Messung der Persönlichkeit. In: Heiss, R., Groffmann, K. & Michel, L. (Hrsg.): Handbuch der Psychologie, Bd. 6: Psychologische Diagnostik. Göttingen: Hogrefe.

Mucchielli, R. 1973. Die Befragung in der Sozialpsychologie. Salzburg: Müller.

Münch, W. 1971. Datensammlung in den Sozialwissenschaften. Stuttgart: Kohlhammer.

Narayana, C. L. 1977. Graphic positioning scale: an economical instrument for surveys. Journal of Marketing, 14, 118-122.

Noelle, E. 1963. Umfragen in der Massengesellschaft. Reinbek: Rowohlt.

Noelle-Neumann, E. 1970. Wanted: Rules for wording structured questionnaires. Public Opinion Quarterly, 34, 191-201.

Noelle-Neumann, E. 1974. Probleme des Fragebogenaufbaus. In: Behrens, K. C. (Hrsg.): Handbuch der Marktforschung. Wiesbaden: Gabler.

Nowakowska, M. 1971. A model for answering a questionnaire item. Polish Psychological Bulletin, 2, 37-45.

Oppenheim, A. N. 1966. Questionnaire design and attitude measurement. New York: Basic Books.

Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. 1957. The measurement of meaning. Urbana: University of Illinois.

Payne, S. L. 1951. The art of asking questions. Princeton: Princeton University Press.

Perreault, W. D. 1975. Controlling order-effect bias. Public Opinion Quarterly, 39, 544-551.

Phillips, B. S. 1966. Social research: strategy and tactics. New York: Macmillan.

Phillips, B. S. 1970. Empirische Sozialforschung. Wien: Springer.

Raab, E. 1974. Probleme der Frageformulierung. In: Behrens, K. C. (Hrsg.): Handbuch der Marktforschung. Wiesbaden: Gabler.

Richardson, S. A., Dohrenwend, B. S. & Klein, D. 1965. Interviewing: its forms and functions. New York: Basic Books.

Richter, H. J. 1967. Ist die schriftliche Befragung eine brauchbare Methode in der empirischen Sozialforschung? Der Marktforscher, 8, 234-235.

Richter, H. J. 1969. Grundlagen schriftlicher Massenbefragungen: ein verhaltenstheoretischer Beitrag zur Methodenkritik. München: Phil. Diss.

Ring, E. 1969. Haben Hintergrundfarben des Testmaterials Einfluß auf die Ergebnisse? Psychologie und Praxis, 13, 82-87.

Ring, E. 1974. Wie man bei Listenfragen Einflüsse der Reihenfolge ausschalten kann. Psychologie und Praxis, 18, 105-113.

Ring, E. 1975. Eine Fehlerquelle bei Bildern als Testvorlage. Zeitschrift für experimentelle und angewandte Psychologie, 22, 89-93.

Rogers, T. B. 1974a. An analysis of the stages underlying the process of responding to personality items. Acta Psychologica, 38, 205-213.

Rogers, T. B. 1974b. An analysis of two central stages underlying responding to personality items: the selfreferent decision and response selection. Journal of Research in Personality, 8, 128-138.

Rohrmann, B. 1978. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie, 9, 222-245.

Rorer, L. G. 1965. The great response-style myth. Psychological Bulletin, 63, 129-156.

Roslow, S., Wulfeck, H. & Corby, G. 1940. Consumer and opinion research: Experimental studies on the form of questions. Journal of Applied Psychology, 24, **334-346.**

Rugg, D. 1941. Experiments in wording questions. Public Opinion Quarterly, 5, **91-92.**

Rugg, D. & Cantril, H. 1972. Die Formulierung von Fragen. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.

Scheier, I. H. & Cattell, R. B. 1965. Bestätigung von objektiven Testfaktoren und Beurteilung ihrer Beziehung zu Fragebogenfaktoren. Diagnostica, **11, 95-120.**

Scheuch, E. K. 1962. Skalierungsverfahren in der Sozialforschung. In: König, R. (Hrsg.): Handbuch der empirischen Sozialforschung. Stuttgart: Enke.

Scheuch, E. K. 1973. Das Interview in der Sozialforschung. In: König, R. (Hrsg.): Handbuch der empirischen Sozialforschung. Stuttgart: dtv.

Scheuch, E. K. & Zehnpfennig, H. 1974. Skalierungsverfahren in der Sozialforschung. In: König, R. (Hrsg.): Handbuch der empirischen Sozialforschung. Stuttgart: Enke.

Schneider, J. 1972. Versuchsleitereinfluß in Abhängigkeit von Merkmalen der Versuchsperson und dem Aussehen des Versuchsleiters. Saarbrücken: Phil. Diss.

Schneider-Düker, M. & Schneider, J. 1977. Untersuchungen zum Beantwortungsprozeß bei psychodiagnostischen Fragebogen. Zeitschrift für experimentelle und angewandte Psychologie, 24, 282-302.

Schreiber, K. 1974. Standardisierte und nichtstandardisierte Interviews. In: Behrens, K. C. (Hrsg.): Handbuch der Marktforschung. Wiesbaden: Gabler.

Schriesheim, C. & Schriesheim, J. 1974. Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. Educational and Psychological Measurement, 34, 877-884.

Schyberger, B. W. 1966. A case against direct questions on reading habits. Journal of Advertising Research, 6, 25-30.

Sharma, S. N. & Singh, Y.P. 1967. Does the colour pull response? Manas: A Journal of Scientific Psychology, 14, 77-79.

Sheatsley, P. B. 1972. Die Kunst des Interviewens. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.

Sheth, J. & Roscoe, A. 1975. Impact of questionnaire length, follow-up methods, and geographical location on response rate to a mail survey. Journal of Applied Psychology, 60, 252-254.

Sieber, M. 1979a. Zur Zuverlässigkeit von Eigenangaben bei einer Fragebogenuntersuchung. Zeitschrift für experimentelle und angewandte Psychologie, 26, 157-167.

Sieber, M. 1979b. Zur Erhöhung der Rücksendequote bei einer postalischen Befragung. Zeitschrift für experimentelle und angewandte Psychologie, 26, 334-340.

Simpson, R. H. 1944. The specific meaning of certain terms indicating different degrees of frequency. Quarterly Journal of Speech, 30, 328-330.

Sixtl, F. 1972. Gedanken über die Verzahnung von allgemeiner und differentieller Psychologie. Archiv für Psychologie, 124, 145-157.

Spoerer, E. 1979. Einführung in die Verkehrspsychologie. Darmstadt: Wissenschaftliche Buchgesellschaft.

Steward, C. J. & Cash, W. B. 1978. Interviewing: principles and practics. Dubuque: W. C. Brown.

Stollberger, R. 1966. Die Befragung. In: Jetzschmann, H., Kallabis, H., Schulz, R. & Taubert, H. (Hrsg.): Einführung in die soziologische Forschung. Berlin: VEB Deutscher Verlag der Wissenschaften.

Strahan, R. & Gerbasi, K. C. 1973. Semantic style variance in personality questionnaires. Journal of Psychology, 85, 109-118.

Strong, E. R. 1962. Good and poor interest items. Journal of Applied Psychology, 46, 269-275.

Stroschein, F. R. 1965. Die Befragungstaktik in der Marktforschung. Wiesbaden: Gabler.

Suchman, E. A. & Guttman, L. 1947. A Solution to the problem of question bias. Public Opinion Quarterly, 11, 445-455.

Sudman, S. & Bradburn, N. M. 1974. Response effects in surveys: a review and Synthesis. Chicago: Aldine.

Süllwold, F. 1969. Theorie und Methodik der Einstellungsmessung. In: Graumann, C. F. (Hrsg.): Handbuch der Psychologie, Bd. 7: Sozialpsychologie, 1. Halbbd. Göttingen: Hogrefe.

Taietz, P. 1972. Conflicting group norms and the ‚third' person in the interview. American Journal of Sociology, 68, 97-104.

Terborg, J. R. & Peters, L. H. 1974. Some observations on wording of item stems for attitude questionnaires. Psychological Reports, 35, 463-466.

Tholey, V. 1976. Die ‚social desirability'-Variable bei der Beantwortung von Persönlichkeitsfragebogen. Darmstadt: Phil. Diss.

Tittle, C. R. & Hill, R. J. 1967. The accuracy of self-reported data and prediction of political activity. Public Opinion Quarterly, 31, 103-106.

Tränkle, U. 1974. Empirische Untersuchungen zum Einfluß der Befragungsmethode auf Befragungsergebnisse: standardisiertes Interview und schriftliche Befragung. Frankfurt: Unveröff. Jahresarbeit.

Turner, C. & Fiske, D. W. 1968. Item quality and appropriateness of response processes. Educational and Psychological Measurement, 28, 297-315.

Whitney, D. R. & Feldt, L. S. 1973. Analyzing questionnaire results: multiple tests of hypothese and multivariate hypotheses. Educational and Psychological Measurement, 33, 365-380.

Wieken, K. 1974. Die schriftliche Befragung. In: Koolwijk, J. von & Wieken-Mayser, M. (Hrsg.): Techniken der empirischen Sozialforschung, Bd. 4: Erhebungsmethoden: Die Befragung. München/Wien: Oldenbourg.

Wildman, R. C. 1977. Effects of anonymity and social setting on survey responses. Public Opinion Quarterly, 41, 74-79.

Wilk, G. 1974. Psychologische Probleme der Interviewsituation. In: Behrens, K. C. (Hrsg.): Handbuch der Marktforschung. Wiesbaden: Gabler.

Willick, D. H. & Ashley, R. K. 1971. Survey question order and the political party preference of College students and their parents. Public Opinion Quarterly, 35, **189-199.**

Wottawa, H. 1980. Grundriß der Testtheorie. München: Juventa.

Wright, P. & Barnard, P. 1975. Just fill in this form: a review for designers. Applied Ergonomics, 6, 213-220.