ZUMA-Arbeitsbericht Nr. 96/05

Verfahren zur Evaluation von Survey - Fragen: Ein Überblick

Peter Prüfer und Margrit Rexroth

Mai 1996

ZUMA
Quadrat B2,1
Postfach 12 21 55
D-68072 Mannheim

Telefon: (0621) 12 46 - 227 und 230
Telefax: (0621) 12 46 - 100
E-mail: pruefer@ZUMA-Mannheim.de
rexroth@ZUMA-Mannheim.de

INHALT

Seite

1. Einleitung: Grundlagen

Wer Daten mittels Umfragen erhebt, kennt das Problem: Werden die Fragen des Fragebogens „gute“ Daten liefern, d.h. werden sie zuverlässig das messen, was sie messen sollen und damit reliable und valide Antworten liefern?
Was man nicht erst seit heute weiß: Keinesfalls genügt eine Evaluation des Fragebogens am Schreibtisch durch Experten, um gute Daten sicherzustellen. Beispielsweise betonen Sudman/Bradburn (1982):

„Even after years of experience, no expert can write a perfect questionnaire.“

Dafür gibt es gute Gründe: Das wesentliche Problem, mit dem sich der Fragenkonstrukteur in der Praxis konfrontiert sieht, ist der offensichtliche Mangel an empirisch fundierten, konkreten "Konstruktionsrichtlinien".
Zwar gibt es zur Konstruktion von Fragen in der methodologischen Literatur sowohl eine Reihe von ad-hoc Regeln, Rezepten und Empfehlungen (z.B. Payne 1951; Belson 1981 und 1986; Sheatsley 1983; Sudman/Bradburn 1982) als auch etliche experimentell gewonnene Einzelergebnisse über die Auswirkungen unterschiedlicher Frageformulierungen (vgl. z.B. Schuman/Presser 1981), sowie Ergebnisse aus dem Bereich der kognitionspsychologischen Forschung (vgl. z.B. Schwarz/Sudman 1992; Schwarz/Sudman 1994; Sudman/Bradburn/Schwarz 1996), jedoch erweisen sich diese vorhandenen Informationen in der Praxis lediglich als hilfreich, wenn es darum geht, grobe Fehler zu vermeiden.
Daraus ergibt sich die unbefriedigende Situation, daß trotz Befolgung aller vorhandenen Regeln und Informationen bei der Konstruktion von Fragen ein nur schwer kalkulierbares Restrisiko verbleibt, das auch durch noch so große "Erfahrung" des Fragenkonstrukteurs nicht gänzlich vermieden werden kann:
Als Konsequenz daraus kann nur die dringende Empfehlung ausgesprochen werden, möglichst jede Frage eines Fragebogens vor deren endgültigem Einsatz einem bzw. einer Reihe von Evaluationsverfahren zu unterziehen, um mit möglichst hoher Wahrscheinlichkeit die Existenz von Mängeln auszuschließen.

Unmißverständlich wird dies z.B. von Sudman/Bradburn (1982) zum Ausdruck gebracht:

„If you don't have the resources to pilot test your questionnaire, don't do the study."

Bis ca. Mitte der 80-iger Jahre wurde als Evaluationsverfahren - mit wenigen Ausnahmen - nur der sog. Pretest eingesetzt, ein Verfahren, bei dem üblicherweise der Fragebogen an einer Mini-Stichprobe unter Feldbedingungen erhoben bzw. getestet wird (field-pretest). Zum Einsatz kamen dabei eine Reihe unterschiedlichster Pretest-Varianten[1], was zum Teil auch die unsystematische und verwirrende Vielzahl der Begriffe erklärt, mit der Autoren die von ihnen beschriebene bzw. empfohlene Vorgehensweise bezeichneten.
Kennzeichnend für die damalige Situation war ganz offensichtlich eine fehlende Sensibilisierung innerhalb der gesamten Profession für die grundlegende Problematik, was Converse/Presser (1986) zum Anlaß nehmen, folgende Situationsbeschreibung abzugeben:

„There are no general principles of good pretesting, no systematization of practice, no consensus about expectations, and we rarely leave records for each other."

In der zweiten Hälfte der 80-iger Jahre verändert sich die gesamte Situation durch die Entwicklung sog. „kognitiver Labormethoden" schnell. Bei diesen kognitiven Labormethoden handelt es sich um kognitionspsychologische Techniken, die in den USA zur Erforschung des Frage-Antwortprozesses eingesetzt wurden und schließlich von amerikanischen Umfrageforschern dazu verwendet wurden, die Qualität von Fragen zu evaluieren.

[1] siehe Kapitel 2.1.1

2. Welche Verfahren stehen zur Verfügung?

2.1 Testerhebungen im Feld

2.1.1 Der Standard - Pretest

Der Begriff „Standard-Pretest" wird in der Literatur erstmals von Oksenberg/Cannell/Kalton (1991) erwähnt. Presser/Blair (1994) verwenden den synonym zu verstehenden Begriff „Conventional Pretest". Mitunter wird auch innerhalb der Umfrageforschung die Bezeichnung „klassischer Pretest" oder „Beobachtungspretest" verwendet. In jüngster Zeit findet sich in der Literatur auch der Begriff „Old-style-pretest" (Fowler 1992).
Die Begriffe lassen zurecht vermuten, daß es sich bei dieser Methode um ein „etabliertes", häufig angewandtes Verfahren handelt, das seit Beginn der Umfrageforschung eingesetzt wurde und allgemein als „Pretest" bezeichnet wird. Erstaunlicherweise existieren selbst für die Durchführung eines Standard-Pretests keine verbindlichen bzw. allgemein akzeptierten Regeln. In fast jedem sozialwissenschaftlichen Methodenlehrbuch findet sich zwar ein Abschnitt zum Thema „Pretest", die Autoren geben jedoch - wenn überhaupt - höchst unterschiedliche und zum Teil auch widersprüchliche Empfehlungen bezüglich dessen Durchführung.

Im folgenden sind die unterschiedlichen Empfehlungen verschiedener Autoren bezüglich der wichtigsten Pretest-Elemente aufgeführt.

- **Stichprobe:**

  Die empfohlene Fallzahl, d.h. die Anzahl der zu befragenden Personen variiert von N = 10 bis N = 200 (vgl. z.B. Backstrom/Hursh 1963; Boyd/Westfall 1964; Schrader 1971; Elliott/Christopher 1973; Friedrichs 1973; Warwick/Lininger 1975; Karmasin/Karmasin 1977; Williamson/Karp/Dalphin 1977; Wellenreuther 1982; Sheatsley 1983; Fowler 1984; Converse/Presser 1986; Schnell/Hill/Esser 1995).

- **Einsatz der Interviewer:**

  Während einige Autoren empfehlen, bei der Durchführung von (Standard-) Pretests ausschließlich mit erfahrenen oder speziell ausgebildeten Interviewern zu arbeiten (vgl. z.B. Atteslander 1984; Schrader 1971; Elliott/ Christopher 1973; Converse/Presser 1986), plädieren andere dafür, einen „Querschnitt" aller bei der Haupterhebung beteiligten Interviewer einzusetzen (vgl. z.B. Backstrom/Hursh 1963; Karmasin/Karmasin 1977; DeMaio 1983).
  Reine Experten als Pretestinterviewer werden z.B. von Noelle-Neumann (1971) und Kidder (1981) empfohlen.
  Daß die Forscher bzw. die Mitglieder von Projektgruppen auch selbst Pretest-Interviews durchführen sollen, findet in der Literatur breite Zustimmung (vgl. z.B. Backstrom/Hursh 1963; Oppenheim 1966; Williamson/Karp/Dalphin 1977; Sudman/Bradburn 1982; Porst 1985).

- **Informiertheit der Befragten:**

  Einerseits besteht die Möglichkeit, Befragte über den Testcharakter der Befragung zu informieren, andererseits kann ein (Standard-) Pretest unter möglichst realistischen Bedingungen durchgeführt werden, ohne die Befragten über den Testcharakter zu informieren. Converse/Presser (1986) führen für die beiden Varianten die Begriffe „participating pretest" und „undeclared pretest" ein.

- **Informationsbeschaffung:**

  Es existieren unterschiedliche Möglichkeiten darüber, wie der Interviewer seine jeweiligen Pretest-"Erkenntnisse" an den Forscher weiterleitet bzw. meldet:
  Report des Interviewers, und zwar entweder schriftlich in Form eines sog. „Erfahrungsberichts" bzw. „Pretest-Reports" (meist über jedes durchgeführte Interview) oder mündlich als sog. „Debriefing", und zwar entweder in Einzel-Sitzungen oder alle am Standard-Pretest beteiligten Interviewer berichten in einer gemeinsamen Sitzung über ihre Interview-Erfahrungen.
  Grundsätzlich besteht weiterhin die Möglichkeit, auch Befragte einem Debriefing zu unterziehen.

Trotz fehlender empirisch fundierter Regeln gibt es zumindest eine Art Übereinstimmung darüber, wie das Grundgerüst eines Standard-Pretests beschaffen ist bzw. sein sollte. Dafür sprechen neben der wissenschaftlichen Literatur einerseits eigene langjährige Erfahrungen der Autoren bei der Durchführung von Standard-Pretests sowie andererseits Informationen darüber, wie kommerzielle Umfrageinstitute die Durchführung dieses Verfahrens handhaben.

Danach zeichnet sich ein Standard-Pretest durch folgende Merkmale aus:

- **Einmalige Erhebung eines Fragebogens unter möglichst realistischen Hauptstudie - Bedingungen**

- **Durchführung von 20 bis 50 Interviews (Quota oder Random)**

- **Interviewer haben die Aufgabe, Probleme und Auffälligkeiten bei der Durchführung der Interviews zu beobachten und zu berichten.**

- **In der Regel handelt es sich um ein passives Verfahren, d.h. der Interviewer beobachtet nur (deshalb auch „Beobachtungspretest"), ohne aktiv zu hinterfragen.**

Bei dieser Vorgehensweise liegt die Strategie bzw. das Prinzip zugrunde, aus der Reaktion bzw. Antwort der Befragten Rückschlüsse auf das Fragenverständnis zu ziehen. Dabei werden ganz allgemein Probleme der Befragten bei bzw. mit einer Frage auf Konstruktionsmängel der jeweiligen Frage zurückgeführt, während Fragen, bei denen Befragte formal „korrekt" antworten, als „gut konstruiert" angesehen werden.

Stärken des Verfahrens

- Ein Standard-Pretest ist in der Regel relativ schnell und problemlos durchführbar. Der organisatorische Aufwand ist eher niedrig. Dies ist besonders dann der Fall, wenn die Befragten nach einem Quotenverfahren ausgewählt werden.

- Die Kosten sind relativ niedrig. Auch hier gilt: Der Einsatz des Quotenverfahrens wirkt sich kostenmindernd aus.
- Eine annähernd realistische Schätzung der Befragungsdauer ist möglich.

Schwächen des Verfahrens

- Der diesem Pretest-Prinzip zugrundeliegende Schluß, Fragen, die Befragte formal „korrekt" beantworten, könnten als „gut konstruiert" angesehen werden, ist grundsätzlich unzulässig. Belson (1981 und 1986) kann überzeugend nachweisen, daß trotz - formal - korrekter Antwort ein falsches Fragenverständnis zu Grunde liegen kann[2].
- Die Instruktion an die Interviewer, was sie beobachten und berichten sollen, ist meist wenig präzise. Das Produkt - die Berichte - sind dementsprechend auch sowohl inhaltlich als auch formal wenig systematisch.
- Interviewer berichten trotz intensiver Schulung bei weitem nicht alle im Standard-Pretest aufgetretenen „Probleme" (Kreiselmaier/Prüfer/Rexroth 1989).
- Insgesamt gesehen handelt es sich beim Standard-Pretest - allein schon wegen der geringen Fallzahl - um ein sehr „grobes" Verfahren.

2.1.2 Behaviour Coding

Dieses Evaluationsverfahren hat in den letzten Jahren einen wichtigen Platz eingenommen. Cannell/Oksenberg/Kalton/Bischoping/Fowler (1989) bezeichnen sie als sinnvolle und nützliche Technik zur Identifizierung von Fragenmängeln.

Das grundlegende Prinzip dieser Technik basiert auf der Klassifizierung von Verhalten. Mit Hilfe eines mehr oder weniger detaillierten Codesystems wird das Verhalten von befragter Person und Interviewer bewertet und analysiert.

Ursprünglich wurde diese Technik eingesetzt, um ausschließlich Interviewerverhalten zu klassifizieren und zu bewerten[3]. In späteren Arbeiten wurde dann auch das Befragtenverhalten vercodet (vgl. hierzu z.B. Morton-Williams 1984, Prüfer/Rexroth

[2] siehe Kapitel 2.1.4

1985, Oksenberg/Cannell/Kalton 1991). Die von den einzelnen Forschern konzipierten Codesysteme zur Bewertung der Verhaltensweisen bzw. der verbalen Aktivitäten im Interview unterscheiden sich etwas in Aufbau und Detailliertheit, für alle gilt jedoch das grundsätzliche Prinzip: Das Verhalten von befragter Person und Interviewer wird mit Hilfe des Codesystems systematisch registriert. Durch diese Eigenschaft der Technik lassen sich Rückschlüsse auf die Qualität einer Frage ziehen. Damit ist die Behaviour-Coding-Technik eine ernstzunehmende Alternative, wenn es im Pretest darum geht, zur Qualitätsbestimmung von Fragen und Instrumenten nutzbringende Techniken einzusetzen.
Wie arbeitet diese Methode und was leistet sie? Im folgenden soll darauf kurz eingegangen werden.

2.1.2.1 Traditionnelle Vorgehensweise beim Einsatz der Behaviour-Coding-Technik

Bei dem vornehmlich in der englischsprachigen Literatur beschriebenen Behaviour Coding bewerten sogenannte „Coder“ das auf Tonband aufgezeichnete Interview, d.h. sie bewerten das Interviewer - und Befragtenverhalten mittels eines Codesystems, das mehr oder weniger umfangreich sein kann, und damit mehr oder weniger differenziert Verhalten erfaßt.

[3] siehe hierzu die grundlegende Arbeit von Cannell, Lawson & Hausser 1975

Als Beispiel für ein Codesystem sei dasjenige aufgeführt, das in der Studie von Oksenberg/Cannell/Kalton (1991) verwendet wurde. Dieses Schema sieht 3 Kategorien zur Bewertung von Interviewerverhalten beim Vorlesen des Fragetextes vor (Codes E, S, M) und 7 Kategorien zur Erfassung von Befragtenverhalten (Codeziffern ❶ bis ❼):

Codesystem für Behaviour Coding

Code	Beschreibung	
	Interviewer	
E	**Exact**	**Interviewer liest Frage exakt**
S	**Slight change**	**leichte Änderungen**
M	**Major change**	**starke Änderungen**
	Befragte(r)	
❶	**Interruption**	**Befragte(r) antwortet vorzeitig**
	Clarification	**Befragte(r) will Wiederholung der Frage oder Klärung der Frage oder macht Bemerkung, die auf Verständnisproblem schließen läßt**
	Adequate answer	**Befragte(r) antwortet adäquat**
	Qualified answer	**Antwort ist adäquat, zusätzliche Bemerkung läßt jedoch auf Unsicherheit schließen**
	Inadequate answer	**Inadäquate Antwort**
	Don`t know	**Weiß nicht**
❼	**Refusal to answer**	**Befragte(r) verweigert Beantwortung der Frage**

Vor der Vercodung sollte genau festgelegt werden, welches Verhalten (besonders das des/der Befragten) überhaupt berücksichtigt werden soll bzw. welche Vercodungsregeln zugrunde gelegt werden.
Oksenberg/Cannell/Kalton (1991) berichten von drei möglichen Varianten:

1. Vercodet wird nur die erste Reaktion des/der Befragten nach der Präsentation des Fragestimulus.
   Diese Variante hat den Nachteil, daß Befragte manchmal zunächst adäquat antworten und erst danach „problematisch" reagieren.
2. Vercodet wird das gesamte Befragtenverhalten, d.h. es können mehrere Codes pro Frage - auch mehrfach - vergeben werden.
   Diese Variante hat zum einen den Nachteil, daß das mehrfache Auftreten eines Codes nur wenig Zusatzinformation bringt, zum anderen könnten einzelne Extremfälle die Gesamthäufigkeiten verschiedener Codes pro Frage in verzerrender Weise beeinflussen.
3. Vercodet wird das gesamte Befragtenverhalten, im Unterschied zur zweiten Variante werden identische Codes nur einmal vergeben, auch wenn sie mehrmals auftreten sollten.

In der Studie von Oksenberg/Cannell/Kalton (1991) kam die dritte Variante zur Anwendung. Praktisch erhält man damit pro Frage eine Häufigkeitsverteilung über alle bei der Frage vergebenen Codes. Dabei wird die Häufigkeit der durch die Codes repräsentierten Verhaltensweisen von Interviewer und Befragten, die bei der Beantwortung einer Frage auftreten, als Qualitätsindikator dieser Frage gewertet.

Im folgenden soll unter Verwendung des oben dargestellten Codeschemas von Oksenberg/Cannell/Kalton (1991) die Vercodung einer Frage beispielhaft demonstriert werden, wobei lediglich das Befragtenverhalten, nicht das Interviewerverhalten, anhand von 10 Interviews vercodet wurde:

Übersicht 1: Vercodung von Befragtenverhalten bei einer Frage für 10 Interviews

Code-ziffer	ID 1	ID 2	ID 3	ID 4	ID 5	ID 6	ID 7	ID 8	ID 9	ID 10	N
1		x						x			2
2	x				x						2
3			x	x	x			x	x	x	6
4	x			x	x						3
5	x	x					x				3
6		x					x				2
7						x					1
											19

Als erster Indikator zur Qualitätsbestimmung dieser Frage kann die Anzahl der insgesamt vergebenen 19 Codes gewertet werden.
Wäre die Frage in allen 10 Fällen spontan und adäquat, also formal korrekt beantwortet worden, so wären insgesamt 10 Codes - und zwar 10 mal die Codeziffer 3 für adäquate Beantwortung - vergeben worden. Ein solches Ergebnis spräche für eine gute Qualität der Frage, da man auf Grund spontaner und adäquater Beantwortung einer Frage bei dieser Art der Pretestbeobachtung davon ausgeht, daß keine Verständnisprobleme vorliegen.

Eine Anzahl von insgesamt 19 vergebenen Codes läßt auf zusätzliche unerwünschte Befragtenreaktionen schließen und ist somit ein Indikator dafür, daß die Befragten irgendwelche Probleme bei der Beantwortung der Frage hatten.
Die einzelnen konkreten Probleme können jedoch allein auf Grund der Verteilung der Codeziffern nicht erkannt werden. Die in der Matrix gekennzeichneten Codezif-

fern zeigen lediglich, daß in 40% der Fälle (ID 1, 2, 6 und 7) die Antwort inadäquat gegeben wurde, in 30% zwar adäquat, jedoch mit zusätzlichen unerwünschten Reaktionen (ID 4, 5 und 8) und nur in 30% der Fälle antworteten die Befragten formal korrekt ohne offensichtlichen Hinweis auf Probleme (ID 3, 9 und 10).
Das Vercodungsergebnis in unserem Beispiel weist auf mangelnde Qualität bei der Frage hin.
Vergleicht man die Leistungsfähigkeit der Technik mit der anderer Pretesttechniken, so ist sie mit Abstand diejenige Technik, deren Pretesterkenntnisse am reliabelsten sind (vgl. hierzu z.B. Presser/Blair 1994). Gerade im Vergleich zum Standard-Pretest, bei dem die Pretesterkenntnisse oftmals von der subjektiven Wahrnehmung des einzelnen Interviewers geprägt sind, besticht die Behaviour-Coding-Technik durch ihre objektive und systematische Vorgehensweise.

Der wesentliche Nachteil der Technik liegt allerdings darin, daß sie Hinweise auf mögliche Ursachen für inadäquates Verhalten nicht miterfaßt. Diese Tatsache ist umso schwerwiegender, als es ja gerade das Ziel einer Evaluation ist, ganz konkret die Schwächen bei einer Frage zu erkennen, um sie dann für den Hauptfragebogen zu eliminieren.
Da das Erhebungsverfahren beim Behaviour Coding im Grunde demjenigen des Standard-Pretests entspricht (d.h. aus der Beobachtung des Befragtenverhaltens werden Rückschlüsse auf das Fragenverständnis gezogen), ist ein weiterer Nachteil darin zu sehen, daß trotz formal korrekter Antwort ein falsches Fragenverständnis zu Grunde liegen kann[4].
Schließlich besteht ein Nachteil des Behaviour Coding darin, daß die Interviews auf Band aufgezeichnet werden müssen.

2.1.2.2 Eine Variante des Behaviour Coding: Problem Coding

In der Zuma - Feldabteilung wird das Behaviour Coding seit Jahren modifiziert eingesetzt. Die Modifikation besteht in einer Verbindung zwischen Standard-Pretest und traditionnellem Behaviour Coding. Die Autoren nennen dieses Verfahren „Problem Coding“ .

[4] siehe Kapitel 2.1.4

Wesentlich für das Problem Coding ist, daß die Bewertung der Verhaltensweisen des/der Befragten nicht vom Coder nach dem Interview, sondern vom Interviewer selbst während des Interviews vorgenommen wird. Dabei ist für den Interviewer die Anwendung eines ausführlichen, detaillierten Codesystems nicht möglich. Es würde den Interviewer in Verbindung mit seinen eigentlichen Aufgaben, nämlich der korrekten Durchführung des Interviews und der Registrierung der Antworten unter Berücksichtigung der eingeübten Regeln zur Durchführung des standardisierten Interviews stark überfordern. Die Voraussetzung zur Bewältigung der „Vercodungsarbeit" während des Interviews ist der Einsatz eines auf das äußerste reduzierten Codesystems, das - spontane - Befragtenverhalten nur noch im Hinblick darauf, ob es im Sinne der Fragestellung adäquat oder nicht adäquat ist, mittels einer Codeziffer im Fragebogen bewertet. Dabei bieten sich die beiden Ziffern „0" für „adäquates Verhalten" und „1" für „nicht adäquates Verhalten" an.
Eine weiteres Kennzeichen des Problem Coding liegt darin, daß der Interviewer im Unterschied zum traditionnellen Behaviour Coding in einem zweiten Schritt zusätzlich bei inadäquater Verhaltensweise des/der Befragten in einem schriftlichen Interviewererfahrungsbericht nach dem Interview möglichst detailliert dieses Befragtenverhalten beschreibt. Damit erhält der Forscher Hinweise auf mögliche Ursachen für den Mangel bei einer Frage.
An einer Frage aus einer Pilotstudie zur Europawahl soll beispielhaft demonstriert werden, welche Informationen ein Forscher durch den Einsatz des Problem Codings bei einer Frage erhalten kann. Der Fragebogen zu dieser Pilotstudie wurde unter Anwendung des Problem Coding von der ZUMA-Feldabteilung 1993 an 500 Fällen erhoben.

Fragetext:
Nehmen wir einmal an, bei einer Diskussion stünden sich zwei Meinungen gegenüber. Die eine Seite vertritt die Ansicht, daß es eine Selbstverständlichkeit sei, sich an Wahlen zu beteiligen. Die andere Seite vertritt die Ansicht, man solle sich grundsätzlich nicht an Wahlen beteiligen.
Sagen Sie mir bitte Ihre Meinung mit Hilfe dieser Liste. Der Wert 1 bedeutet, man solle sich grundsätzlich nicht an Wahlen beteiligen. Der Wert 11 bedeutet, es sei eine Selbstverständlichkeit, sich an Wahlen zu beteiligen. Mit den Werten dazwischen

können Sie Ihre Meinung abstufen. (Listenvorlage mit 11 Punkte-Skala, wobei die Endpunkte der Skala verbalisiert waren).

Rechts neben den Fragetext waren die Codeziffern „0“ und „1“ in den Fragebogen gedruckt.
Nannte der/die Befragte ohne erkennbare Probleme spontan einen Skalenwert, so hatte der Interviewer die Aufgabe, zusätzlich zur formalen Registrierung des genannten inhaltlichen Wertes die Codeziffer „0“ für adäquates Antwortverhalten einzukreisen. In allen anderen nicht adäquaten Befragungsabläufen (z.B. bei Rückfragen zum Verständnis, Antwortformulierungen in eigenen Worten u.s.w.) kreiste der Interviewer die Codeziffer „1“ ein. Zusätzlich beschrieb der Interviewer in diesen Fällen das Antwortverhalten des/der Befragten nach Ende des Interviews in einem Zusatzbericht.

Quantitatives Ergebnis des Problem Coding bei der Wahlfrage:
In 7% aller Fälle war die Zahl „1“ gecodet, was für einen formal nicht korrekten Befragungsablauf stand. Dieser Wert kann als Qualitätsindikator dieser Frage - auch im Vergleich zu dem anderer Fragen des Fragebogens - gesehen werden. Mit 7% unerwünschtem Befragtenverhalten signalisiert er durchaus ernstzunehmende Probleme bei der Beantwortung der Wahlfrage.
Qualitatives Ergebnis des Problem Coding bei der Wahlfrage:
Die aufgetretenen Probleme waren einerseits typische Skalenprobleme: In 7 Fällen mußte die Skala noch einmal erklärt werden, in 6 Fällen war unklar, ob man einen Wert nennen oder auf der Liste ankreuzen soll. Andererseits gab es Verständnisprobleme: Es war 22 Befragten unklar, an welche Wahlen sie denn bei der Frage konkret denken sollen: Generell an Wahlen, speziell an die Europa-Wahl, um die es ja generell im Fragebogen ging oder an Betriebsratswahlen.
Der entscheidende Vorteil des Problem Coding gegenüber dem Behaviour Coding liegt also darin, daß man neben dem quantitativen Häufigkeitswert als Qualitätsindikator durch die Beschreibung des konkreten Befragtenverhaltens bzw. des Problems auch Hinweise auf die konkreten Ursachen für den Qualitätsmangel einer Frage erhält.

Bei ZUMA hat sich die Problem Coding Technik bewährt. Für den Interviewer bedeutet der Einsatz der Technik allerdings eine hohe Anforderung, die nur durch entsprechende Schulungsmaßnahmen erfüllt werden kann.

2.1.3 Random Probe

Die "Random-Probe-Technik", wurde von Schuman (1966) mit dem Ziel entwickelt, das Fragenverständnis bei geschlossenen Fragen in Hauptstudien zu überprüfen. Dabei wählt jeder Interviewer vor dem Interview nach einem Zufallsverfahren eine bestimmte Anzahl von Fragen aus, bei denen Zusatzfragen (Probes) zum Fragenverständnis gestellt werden. Beispielsweise wurden in der von Schumann erwähnten Studie pro Fragebogen jeweils zehn der insgesamt 200 Items zufällig ausgewählt. Als Probes standen dabei folgende drei Formulierungen zur Verfügung (Originaltext S. 241):

1. „Would you give me an example of what you mean?“
2. „I see - why do you say that?“
3. „Could you tell me a little more about that?“

In seiner Studie demonstriert Schuman an zwei Fragen eindrucksvoll die Eignung seiner Random-Probe-Technik und kommt zu folgendem Ergebnis (S. 244):

„The answers to these questions show excellent variation, intercorrelate well, are significantly related to a number of background variables, and are relevant to an important hypothesis. But the random probes suggest that the questions were reasonably well understood by less than half the sample.“

Obwohl die Random-Probe-Technik von Schuman ursprünglich zum Einsatz in Hauptstudien vorgesehen war, scheint sie auch in Testerhebungen zur Evaluation von Fragen sinnvoll anwendbar zu sein.

2.1.4 Intensive Interview (Belson)

Belson (1981 und 1986) kann überzeugend nachweisen, daß auch formal korrekten Antworten ein falsches, d.h. vom Fragenkonstrukteur nicht intendiertes Verständnis des Frageinhalts zugrunde liegen kann. Mit der üblichen Pretteststrategie, aus den Reaktionen bzw. Antworten der Befragten Rückschlüsse auf das Fragenverständnis zu ziehen, sind solche Fälle, bei denen trotz formal korrekter Antwort ein falsches Fragenverständnis vorliegt, nicht zu erkennen. Belson empfiehlt dafür ein Verfahren, bei der Befragte nach der Durchführung eines Standard-Pretest-Interviews zum Verständnis von drei bis vier bereits vorher festgelegter Fragen intensiv befragt wird. Belson nennt dieses zweite Interview "Intensive Interview, das mittels eines zweistufigen Vorgehens erhoben wird:

1. Der/die Interviewer/in liest die zu testende Frage sowie die aus dem Standard-Pretest-Interview bereits vorliegende Antwort noch einmal vor. Der/die Befragte wird anschließend gebeten, eine Beschreibung darüber zu geben, wie die Antwort zustande kam, wobei der/die Interviewer/in extensiv nachfragen soll.
2. Der/die Interviewer/in stellt eine oder mehrere fest vorgebene Fragen, um festzustellen, wie bestimmte Aspekte bzw. Konzepte, die mit dieser Frage verknüpft sind, interpretiert wurden.

Variationen dieser Technik sind auch unter den Bezeichnungen „Reinterview“ (Bailar 1986), „Double Interview“ (Gordon 1963) „Intensive Reinterview“ (Johnson/Woltman 1986), oder „Follow-Up Interview“ (Morton-Williams/Sykes 1984) bekannt.

2.1.5 Unstrukturierte/ qualitative Interviews/ Tiefeninterviews

Unstrukturierte Interviews, Tiefeninterviews und ähnliche „qualitative“ Interviewformen können sinnvoll in einer frühen Entwicklungsphase des Fragebogens eingesetzt werden. Diese Interviews besitzen einen eher explorativen und experimentellen Charakter, d.h. sie dienen vorwiegend dazu, Ideen, Hinweise und Informationen zur Fragenkonstruktion zu generieren. Der Interviewer ist dabei von den Zwängen eines standardisierten Interviews befreit, d.h. er kann bei Bedarf z.B. nachfragen, hinterfragen oder alternative Fragenversionen anbieten.

2.1.6 Analyse der Antwortverteilungen

Über die Häufigkeitsverteilung von Antwortalternativen lassen sich - meist nur grobe - Rückschlüsse auf die Qualität einer Frage ziehen. Indikatoren für Fragenmängel sind dabei in der Regel

- nicht oder nur minimal besetzte Antwort-Kategorien,
- extreme Häufigkeitsverteilung über alle Antwort-Kategorien,
- hohe Häufigkeitswerte bei sog. „Ausweichkategorien“ , wie z.B. „weiß nicht“ (Befragte/r kann sich nicht entscheiden oder hat keine Informationen) oder „verweigert“ (Befragte/r möchte die Frage nicht beantworten).

Sinnvoll ist dieses Verfahren nur bei einer genügend großen Fallzahl.

2.1.7 Split-Ballot

Beim Split-Ballot-Verfahren werden zwei (oder mehr) Varianten einer Frage jeweils einer Teilgruppe der Befragtenstichprobe zur Beantwortung präsentiert. Unterschiede in den Antwortverteilungen werden dann auf die unterschiedlichen Fragevarianten zurückgeführt.
Unter dem Aspekt der Evaluation von Fragen hat das Split-Ballot-Verfahren zum Ziel, eine Entscheidung für diejenige Fragenvariante herbeizuführen, die letztendlich zum Einsatz kommen soll. Diese Entscheidung trifft der Forscher normalerweise auf der Grundlage der Häufigkeitsverteilungen bzw. auf Grund von statistischen Analysen. Unter dieser Voraussetzung sollte ein Feld-Pretest, bei dem ein Split-Ballot-Verfahren eingesetzt wird, einen Stichprobenumfang von mindestens 100 Interviews haben.
Neben Analyse- und Verteilungsaspekten können aber auch Pretestbeobachtungen als Entscheidungsgrundlage für eine bestimmte Formulierungsvariante einer Frage dienen. Dabei kann es sich um Pretestinformationen unter Einsatz eines traditionnellen Standard-Pretests handeln, aber auch um Beobachtungen aus anderen Verfahren wie z.B. Nachfaßfragen zum Verständnis bestimmter Frageninhalte (Probingverfahren).

2.2. Kognitive Laborverfahren

Aus der interdisziplinären Zusammenarbeit von Kognitionspsychologen und Umfrageforschern, deren Beginn auf das Ende der 70-iger Jahre datiert werden kann, ging eine Reihe von Methoden hervor, die zwar nicht unbedingt neu waren, mit denen jedoch Informationen über kognitive Prozesse während des Frage-Antwort-Prozesses gesammelt werden können. Da diese Informationen Hinweise darüber geben, wie Befragte eine Frage bzw. bestimmte Elemente davon verstehen und interpretieren, sind sie damit auch zur Evaluation von Survey-Fragen geeignet. Diese Methoden sind in den letzten Jahren unter der Bezeichnung „kognitive Laborverfahren“ bekannt geworden. Dabei muß das „Labor“ nicht unbedingt mit technischer Ausrüstung, wie z.B. Tonband, Videorecorder oder Einwegscheibe bestückt sein; in den meisten Fällen genügt ein schlichter Büroraum. In der Regel wird bei diesen Verfahren mit nur wenigen Befragten gearbeitet.

Im folgenden sollen die wichtigsten kognitiven Laborverfahren kurz vorgestellt werden.

- **Think-Aloud**

  Diese Technik kann als die zentrale kognitive Technik überhaupt bezeichnet werden. Der/die Befragte wird aufgefordert, „laut zu denken“ und dabei seine sämtlichen Gedankengänge, die zur Antwort führen bzw. führten zu formulieren. Ziel dabei ist, aus den Äußerungen der Befragten Hinweise darüber zu erhalten, wie die ganze Frage oder einzelne Begriffe verstanden wurden. Die - üblicherweise auf Tonträger - aufgezeichneten Formulierungen werden auch als "verbal protocols" (vgl. z.B. Ericsson/Simon 1980) bezeichnet.
  Über die Anwendung der Think-Aloud-Technik in der Umfrageforschung finden sich in der Literatur wenig klare Instruktionen. So weisen Blair/Presser (1993) anhand einer Befragung von 68 akademischen Institutionen in den USA nach, daß es beim Einsatz der Methode keine klaren Empfehlungen bezüglich Auswahl und Schulung der Interviewer, Anzahl der durchzuführenden Interviews, Bandaufzeichnungen und Analyseverfahren gibt.

Bei der Anwendung der Think-Aloud-Methode gibt es zwei Vorgehensweisen:

1. Die Befragten werden aufgefordert, laut zu denken, **während** sie ihre Antwort formulieren. Diese Vorgehensweise bezeichnet man als **Concurrent-Think-Aloud-Methode**.
2. Die Befragten werden aufgefordert, nach der Beantwortung der Frage darüber nachzudenken, wie die Antwort zustande kam. Diese Vorhehensweise ist bekannt unter dem Begriff **Retrospektive-Think-Aloud-Methode**.

Es gibt Befürworter für die eine und für die andere Vorgehensweise. So sprechen sich z.B. Sudman/Bradburn/Schwarz (1996), für die retrospektive Variante aus, da Befragte erfahrungsgemäß den Prozeß, der zur Antwort führte, nicht immer in Worte fassen können. Die befragte Person wird dann im nachhinein, d. h. nach der Formulierung ihrer Antwort gebeten, ihre Überlegungen zu beschreiben, die zur Antwort führten. Aus einer „Concurrent“ -Vorgehensweise wird so zwangsläufig eine „retrospektive".
Ähnliche Erfahrungen wurden auch in der Feldabteilung beim Einsatz der Concurrent-Think-Aloud-Methode gemacht. So waren bei zu skalierenden Meinungsfragen nur etwa die Hälfte der Befragten in der Lage, vor Nennung eines Skalenwertes ihre Gedanken laut zu formulieren, die letztendlich zu der Entscheidung für einen Skalenwert führten. Die Concurrent-Think-Aloud-Methode stellt an die Befragten hohe Anforderungen, die nur unter detaillierter Anleitung überhaupt erfüllt werden können.

Die Think-Aloud-Methode wurde zur Evaluation von Fragen in unterschiedlichen Bereichen erfolgreich eingesetzt:

1. <u>Bei retrospektiven Fragen:</u>
   Loftus (1984) setzte beispielsweise die Concurrent-Variante ein, um zu klären, wie Befragte bei der Frage, wie häufig sie in den letzten 12 Monaten bei einem Arzt gewesen sind, vorgehen: Überlegen die Befragten vom gegenwärtigen Zeitpunkt ausgehend rückwärts oder umgekehrt, vom Zeitpunkt von vor 12 Monaten bis in die Gegenwart? Die Concurrent-Think-Aloud-Methode war in der Lage zu

zeigen, daß bei autobiographischen Gedächtnisfragen Befragte eher in der „Vergangenheit-Gegenwart-Richtung“ denken.

Ergebnisse eines Einsatzes der Concurrent-Think-Aloud-Methode bei Zuma zeigten, daß bei retrospektiven Faktfragen zu Alltagsgeschehnissen dagegen, wie z.B. Fernsehkonsum der letzten 7 Tage, die Zeiten weder vorwärts noch rückwärts aufaddiert werden, sondern in den meisten Fällen eine Schätzung des durchschnittlichen Verhaltens pro Tag zugrunde gelegt wird, um dieses dann für den entsprechenden Zeitraum hochzurechnen.

Die Methode erweist sich als sinnvoll, um bei retrospektiven Faktfragen den Antwortprozeß transparent zu machen. Kenntnisse dieser Art ermöglichen dann bessere und präzisere Formulierungen dieses Fragentyps.

2. <u>Bei Meinungsfragen:</u>

In einer Studie der Zuma-Feldabteilung wurde der Einsatz der Concurrent-Think-Aloud-Methode bei Meinungsfragen an 31 Fällen überprüft, wobei die Methode nicht wie üblich im Labor, sondern im Feld mit speziell geschulten Pretestinterviewern eingesetzt wurde. Dabei wurde unter anderem das in einer Allbus-Studie erhobene Item „Ein Mann schlägt sein 10-jähriges Kind, weil es ungehorsam war“ in die Überprüfung mit einbezogen. Das Item ist mittels einer 4- Punkte-Skala (sehr schlimm/ziemlich schlimm/weniger schlimm/überhaupt nicht schlimm) zu bewerten. Durch die Methode des lauten Denkens wurden Probleme, die Befragte bei der Bewertung des Items hatten, deutlich. Es handelte sich dabei um die gleichen Probleme, die bereits bei der Durchführung eines Standard-Pretests bei diesem Item bekannt waren. Beim Einsatz der Concurrent-Think-Aloud-Methode traten die Probleme allerdings weit häufiger auf (in 29% aller Fälle, in denen laut gedacht wurde), als dies beim Standard-Pretest der Fall war (2%). Um zu verdeutlichen, welcher Art die genannten Probleme waren, sei hier ein Beispiel aus den Tonbandaufzeichnungen angeführt:

Befragte:

„Das kommt darauf an, in was für einer Situation das ist. Ich meine, es kann Situationen geben, in denen ein Kind es ganz extrem darauf anlegt und wo es auch mal gerechtfertigt ist, ein Kind zu schlagen. Es aber auch nicht richtig schlagen, sondern ihm mal eine Ohrfeige geben oder so...Aber es kommt halt darauf an, wenn

der Mann seinen Frust gerade hat, dann ist es „sehr schlimm". Eine Ohrfeige wäre nicht so schlimm. Also das kann man so nicht bewerten."

Das laute Denken demonstriert die Unzulänglichkeiten der Itemformulierung, in diesem Fall eine zu breite Generalisierung der zu bewertenden Situation.

3. Zur Überprüfung von Hypothesen:
   Bishop konnte 1992 nachweisen, daß sowohl die Concurrent- als auch die Retrospektive-Think-Aloud-Methode auch zur Hypothesenüberprüfung sinnvoll eingesetzt werden kann. Er wendet die Methode bei bereits bekannten Experimenten bezüglich Fragenabfolge und Kontexteffekten wie z.B. dem bekannten Experiment von Schuman und Presser (1981) zu „Communist and American Reporters" an, und weist nach, daß das, was Befragte bei ihrer Antwort laut dachten, genau dem entsprach, was Schuman und Presser als Erklärung des Kontexteffekts formulierten.

- **Probing**

Beim Probing handelt es sich um eine altbekannte Interview-Technik, die z.B. als zentrales Element Bestandteil der bereits beschriebenen Verfahren „Random Probe" von Schuman (1966) und „Intensive Interview" von Belson (1981) sind. Dabei wird eine gegebene Antwort vom Interviewer durch eine oder mehrere Zusatzfragen (Probes) „hinterfragt", um mehr Informationen zu erhalten.
Je nachdem, ob das Probing während des Interviews oder danach durchgeführt wird, werden folgende Bezeichnungen verwendet:

Follow-Up-Probing: Probing sofort nach der spontanen Antwort.

Post-Interview-Probing: Probing nach dem Interview.

Unabhängig vom Probing-Zeitpunkt kann auch nach der Aufgabenstellung, auf die sich das Probing bezieht, unterschieden werden. Hier werden z.B. von Oksenberg/Cannell/Kalton (1991) zwei weitere Probing-Varianten erwähnt:

Comprehension Probing: Probing zum Fragenverständnis.

Oksenberg/Cannell/Kalton (1991) nennen drei Varianten des Comprehension Probing:

1. Befragte sollen die Bedeutung eines bestimmten Begriffs in einer Frage erläutern.
2. Befragte sollen Aspekte ihrer Antwort erläutern.
3. Befragte sollen erläutern, wie klar verständlich ein Begriff für sie war oder welche Probleme sie beim Verständnis eines Begriffs hatten.

Information Retrieval Probing: Probing zu Aspekten der Informationsbeschaffung. Sinnvolles Anwendungsgebiet sind besonders retrospektive Faktfragen.
Beispiel:
Frage: *"Wann waren Sie zum letzten Mal beim Zahnarzt?"*
Information Retrieval Probing*: "Wie schwer fiel es Ihnen, die Frage zu beantworten?"*

- **Confidence Ratings**

Befragte sollen nach der eigentlichen Antwort den Grad der Verläßlichkeit ihrer Antwort bewerten (meist mit Hilfe einer Skala).
Beispiel:
Frage: *"Wie lange haben Sie in den letzten sieben Tagen insgesamt ferngesehen?"*
Confidence Rating: *"Was würden Sie sagen: Ist Ihre Angabe sehr genau, ziemlich genau, eher ungenau oder grob geschätzt?"*

- **Paraphrasing**

  Befragte sollen - nach der Beantwortung - die Frage mit eigenen Worten wiederholen bzw. formulieren.
  Erfahrungsgemäß gehen Befragte dabei unterschiedlich vor: Die einen versuchen, sich möglichst wörtlich an den Fragetext zu erinnern, die anderen versuchen, den Inhalt der Frage in eigenen Worten wiederzugeben. Drei Versuche von Befragten, den Text einer Frage zu wiederholen, sollen im folgenden als Beispiel angeführt werden:

  Fragetext:
  "*Im Vergleich dazu, wie andere hier in Deutschland leben: glauben Sie, daß Sie Ihren gerechten Anteil erhalten, mehr als Ihren gerechten Anteil, etwas weniger oder sehr viel weniger?"* (Dabei wird die Skala optisch nicht vorgegeben)

  Nach Beantwortung der Frage werden die Befragten aufgefordert:
  "*Bitte wiederholen Sie die Frage, die ich Ihnen eben vorgelesen habe noch einmal in Ihren eigenen Worten. Wie lautete die Frage?"*

  Drei Antwortbeispiele aus kognitiven Laborinterviews, die in der Zuma-Feldabteilung erhoben wurden und bei denen das Paraphrasing-Verfahren zum Einsatz kam, sollen die Wirkungsweise dieser Technik demonstrieren:

  Befragter 1:
  „Glauben Sie, daß Sie in Ihrer jetzigen Tätigkeit - verglichen mit anderen in Deutschland Lebenden - den gerechten Anteil bekommen, weniger gerecht, einigermaßen gerecht oder ganz ungerecht“.

  Befragter 2:
  „Daß ich sagen sollte, daß ich im Vergleich zu anderen Bevölkerungsteilen über Maßen vom Sozialstaat profitiere“.

Befragter 3:
„Ob ich eigentlich mit dem, was ich besitze, was ich habe, mit dem, was ich tun kann, zufrieden bin“.

Die Technik kann einem Forscher einerseits aufschlußreiche Hinweise geben, welche inhaltlichen Aspekte Befragte mit einer Frage verbinden, und andererseits kann die Paraphrasing-Technik zeigen, ob der Fragetext in allen Aspekten erinnert werden kann (z.B. konnte die 4-stufige Skala nicht korrekt wiedergegeben werden).

- **Sorting - Verfahren**

Sorting - Verfahren sollen vornehmlich Hinweise darüber geben, wie Befragte Begriffe kategorisieren bzw. als Konzept verstehen. Es gibt drei Varianten:

1. Free Sort

Befragte sollen vorgegebene Items nach eigenen Kriterien gruppieren. Die Items werden dabei auf Kärtchen vorgegeben und sollen in selbstdefinierte Gruppen bzw. „Häufchen“ sortiert werden.

2. Dimensional Sort

Beim Dimensional Sort wird vorgegangen wie beim Free Sort, nur daß hier vorgegebene Items nach vorher festgelegten Kriterien sortiert werden sollen.

3. Vignette Classifications

Bei Vignette Classifications handelt es sich um eine Variante des Dimensional Sort. Beispielsweise sollen Befragte kurze Situationsbeschreibungen („Vignettes“) lesen und jeweils entscheiden, ob diese ihrer Meinung nach in die Überlegungen bei der Beantwortung einer vorgelegten Frage mit einbezogen werden sollen oder nicht. Vignette Classifications werden eingesetzt zur Bestimmung des Verständnisses bestimmter Begriffe.

- **Response Latency**

  Bei dieser Technik handelt es sich um die Messung der Zeit zwischen Präsentation der Frage und der Antwort.
  Die Möglichkeiten reichen dabei von exakter Messung (z.B. mittels Stopuhr oder Computer) bis hin zu einer groben Schätzung durch den/die Testleiter/in mittels Kategorien wie z.B. „kurz“ , „mittel“ , „lang“ . Diese subjektiven Schätzungen werden auch als „Qualitative Timing“ bezeichnet.
  Lange „Reaktionszeiten“ werden dabei in der Regel als Indikator für Fragenmängel interpretiert.

2.2.1 Bewertung der kognitiven Laborverfahren

Als Gesamtbewertung aller hier dargestellten kognitiven Laborverfahren lassen sich folgende Vor- und Nachteile nennen.

Vorteile dieser Techniken:

- Schnelle Durchführung
- Niedrige Kosten
- Die Techniken können innerhalb verschiedener Stadien der Fragebogenkonstruktion angewandt werden (z.B. kann sofort im Anschluß an eine Änderung einer Frage diese neue Version im Labor getestet werden).

Nachteile dieser Techniken:

- Diese Techniken beschränken sich vorwiegend auf die Evaluation einzelner Fragen und nicht auf den Fragebogen als Ganzes. Dies bedeutet, daß diese Labor-Techniken keinesfalls einen Test des gesamten Instruments - in welcher Form auch immer - ersetzen können.
- Durch die geringe Fallzahl besteht ein hohes Unsicherheitsrisiko bezüglich der Generalisierbarkeit der Ergebnisse.

2.3 Andere Verfahren

Im Folgenden werden Evaluationsverfahren kurz vorgestellt, die weder der Kategorie „Testerhebung im Feld" noch der Kategorie „kognitive Laborverfahren" zugeordnet werden können.

2.3.1 Focus Groups

Im Bereich der Evaluation von Fragen können Focus Groups auf zwei Arten sinnvoll eingesetzt werden:

1. In einer frühen Entwicklungsphase des Fragebogens können Focus Groups wertvolle Hinweise zu Akzeptanz oder Verständnis des Befragungsthemas, einzelner Themenbereiche, einzelner Fragen oder einzelner Begriffe geben.
2. Focus-Groups eignen sich besonders dafür, schriftliche Fragebogen zu testen. Dabei bearbeitet zunächst jeder der Gruppenmitglieder für sich den Fragebogen, wobei keine Möglichkeit für Rückfragen gegeben werden sollte, da eine möglichst realistische Bearbeitungssituation simuliert werden soll, vor allem, um für jedes Gruppenmitglied die individuelle Bearbeitungsdauer festhalten zu können. Anschließend können die Gruppenmitglieder allgemeine Eindrücke zum Fragebogen, wie z.B. zur Thematik, zur Bearbeitungsdauer oder zum Schwierigkeitsgrad äußern. Danach wird der Fragebogen Frage für Frage „durchgearbeitet", wobei die Gruppenmitglieder aufgefordert werden, zu jeder einzelnen Frage - soweit vorhanden - Kommentare, Verständnisprobleme oder Rückfragen zu äußern. Daneben werden vom Moderator an die Gruppe bereits vorbereitete Fragen zu einzelnen Fragen gestellt, überwiegend zum Verständnis der ganzen Frage oder einzelner Begriffe.
   Diese Vorgehensweise wurde bei ZUMA bereits mit Erfolg praktiziert.

Grundsätzlich empfiehlt es sich, eine Focus-Group-Sitzung auf Tonträger aufzuzeichnen, als „Notlösung" ist jedoch auch eine schriftliche Protollführung durch einen Co-Moderator denkbar.

Focus Groups sind für die Evaluation einzelner Fragen nur bedingt geeignet. Der Vorteil von Focus Groups liegt vor allem darin, daß mehrere Personen gleichzeitig „befragt“ werden können, ein entscheidender Nachteil ist darin zu sehen, daß soziale Interaktionen bzw. gruppendynamische Prozesse, die zwangsläufig bei einer Focus-Group-Sitzung auftreten, das bei der eigentlichen Befragung relevante Individualverhalten verzerrt darstellen bzw. nicht adäquat repräsentieren. Diese Nachteile sowie die - meist - geringe Fallzahl lassen es ratsam erscheinen, den Fragebogen vor seinem endgültigen Einsatz einem Feld-Pretest unter möglichst realistischen Bedingungen zu unterziehen.

2.3.2 Experten

Zur Beurteilung von Fragebogen eines beliebigen Entwicklungsstadiums können Experten zu Rate gezogen werden. Dabei sollten idealerweise mehrere Experten eingesetzt werden, die ihre Bewertungen zur besseren Vergleichbarkeit anhand eines vorgegebenen Kriterienkatalogs vornehmen. Von Lessler und Forsyth (1995) wurde beispielsweise ein detailliertes Codesystem entwickelt, mit dessen Hilfe Experten eine Frage nach ihren Merkmalen und Eigenschaften - auch im Hinblick auf die Aufgabenstellung für die Befragten - beurteilen können („Expert Questionnaire Appraisal Coding System“).

3. Zur Leistungsfähigkeit einzelner Verfahren

In der Praxis stellt sich die Frage, welches der zahlreichen Evaluationsverfahren für das jeweilige konkrete Umfrageprojekt geeignet ist.

Neben den bereits genannten (und bekannten) Vor- und Nachteilen einzelner Verfahren geben insbesondere zwei Forschungsarbeiten, bei denen verschiedene Verfahren verglichen wurden, interessante Hinweise auf die Leistungsfähigkeit der eingesetzten Verfahren:

- Oksenberg/Cannell/Kalton (1991)
- Presser/Blair (1994).

Übereinstimmendes Fazit beider Arbeiten:

- Es gibt keine Methode, die in allen Problembereichen zufriedenstellend arbeitet.

Oksenberg/Cannell/Kalton (1991) stellen in Ihrer Vergleichsstudie fest, daß sog. „Special Probes" (z.B. Comprehension Probes) zwar erfolgreich zur Aufdeckung von Verständnisproblemen eingesetzt werden können, weniger jedoch zur Identifizierung aller anderen Probleme. Bewährt hat sich in dieser Studie auch das Behaviour Coding, wobei sich allerdings auch hier zeigte, daß die Ursachen der Probleme nicht direkt erkennbar sind.

Presser/Blair (1994) berichten z.B., daß der Standard-Pretest im Vergleich zu anderen Verfahren am wenigsten reliabel ist. Im Gegensatz dazu ist das Behaviour Coding sehr reliabel auf Grund der Anwendung objektiver Regeln, es liefert aber keine Hinweise auf die Ursachen dieser Probleme. Kognitive Verfahren wie Probes und Think-Aloud-Verfahren liefern die meisten Verständnisprobleme, aber z.B. keine Interviewerprobleme. Das Verfahren der Expertenrunde liefert vergleichsweise die meisten Erkenntnisse und ist am kostengünstigsten, besitzt allerdings starke Defizite bei Hinweisen auf Interviewer-Probleme.
Auf der Grundlage dieser Ergebnisse empfielt es sich also, mehrere Verfahren einzusetzen. Da der Erkenntniswert der einzelnen Verfahren für unterschiedliche Problembereiche differiert, sollte der Einsatz der Verfahren sinnvoll kombiniert werden. Auf einen abschließenden Feld-Pretest sollte auf keinen Fall verzichtet werden.

4. Zusammenfassung

Bis Mitte der 80er Jahre stand der Pretest nur äußerst selten im Blickpunkt des wissenschaftlichen Interesses. Er galt zwar in der älteren methodischen Literatur als wesentlicher Bestandteil im Gesamtkonzept einer Umfrage - Studie, gleichzeitig gab es wenig „übereinstimmende" Anhaltspunkte in der Literatur für die konkrete Durchführung. In der Regel kam der Standard-Pretest zur Anwendung, obwohl man sich dessen Schwächen bewußt war und obwohl bereits 1966 Schuman zwecks besserer Information über das Verständnis von Fragen den Einsatz einer „Random Probe"

empfahl und Belson 1981 und 1986 auf Grund seiner Studien die Notwendigkeit sah, formal adäquate Antworten der Befragten zu hinterfragen.

Heute stehen zur Evaluation von Fragen eine ganze Reihe von Verfahren zur Verfügung. Die sozialwissenschaftliche Methodenforschung, die im Bereich der Fragebogenkonstruktion durch die Zusammenarbeit mit Kognitionsforschern in den letzten Jahren zu äußerst praxisrelevanten Erkenntnissen kam, bezog ab Mitte der 80er Jahre auch den Pretestbereich mit ein. Vor allem die amerikanische Literatur beschrieb den erfolgreichen Einsatz von „neuen" Verfahren wie z.B. „Think-Aloud", „Probing", oder „Paraphrasing", die bislang entweder in anderen Forschungsbereichen zur Anwendung kamen oder einfach „in Vergessenheit" geraten waren. Sie waren also nicht unbedingt neu, wurden aber wieder populär und für den Pretestbereich übernommen. Es sind die am häufigsten eingesetzten sogenannten kognitiven Laborverfahren.
Diese neuen kognitionspsychologischen Verfahren bieten den Vorteil, Einblick in die Gedankenprozesse der Befragten zu gewinnen, um so Probleme bei Fragen zu identifizieren. Im Gegensatz dazu ist die Identifizierung von Problemen beim Standard-Pretest ja nur dann der Fall, wenn Befragte selbst um Klärung bitten oder sich offensichtlich falsch verhalten.
Insbesondere hat der Einsatz solcher Verfahren dazu beigetragen, Erkenntnisse bei der Beantwortung retrospektiver Fragen zu gewinnen (vgl. z.B. Tanur 1992; Schwarz/Sudman1994).

Die Vielzahl aktueller Evaluationsverfahren wirft die Frage auf, welches Verfahren nun eingesetzt werden kann bzw. soll.
Sowohl die Ergebnisse von Vergleichsstudien[5] als auch eigene praktische Erfahrungen der Autoren sprechen dafür, mehrere unterschiedliche Verfahren sinnvoll zu kombinieren. Während beispielsweise Fowler (1995) für die Evaluation von Fragen die Verfahren Focus Groups, kognitive Laborinterviews und einen abschließenden Feld-Pretest mit Auswertung der Antwortverteilungen empfiehlt, plädieren die Autoren dieses Berichts für eine flexible, den jeweiligen Gegebenheiten eines Umfrageprojekts angepaßte mehrstufige Vorgehensweise, deren Abschluß ebenfalls ein

[5] siehe Kapitel 3

Feld-Pretest bilden sollte. Für diese Vorgehensweise führen die Autoren den Begriff „Multi-Method-Pretesting“ ein[6].

Daß damit ein höherer Aufwand an zeitlichen und finanziellen Ressourcen verbunden ist, liegt auf der Hand. Berücksichtigt man jedoch den nicht unerheblichen Zuwachs an wichtigen Informationen über Qualitätsmerkmale jeder einzelnen Frage, dann ist dieser Mehraufwand unserer Ansicht nach mehr als berechtigt.

[6] ein ZUMA-Arbeitsbericht Prüfer, P./Rexroth, M.:„Multi-Method-Pretesting“ ist in Vorbereitung

Literatur

Alwin, D., 1977: "Making Errors in Surveys: an Overview." S. 131 - 150 in: Sociological Methods and Research, 6, 1977.

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

Atteslander, P., 1984: Methoden der empirischen Sozialforschung. Berlin: Walter de Gruyter.

Babbie, E. R., 1973: Survey Research Methods. Belmont, Calif.: Wadsworth Publishing Co.

Backstrom, C. H./Hursh, G., 1963: Survey Research. New York: Wiley.

Bailar, B. A., 1986: Recent Research in Reinterview Procedures. S. 41 - 63 in: Journal of the American Statistical Association, 63, 1986

Bailey, K. D., 1982: Methods of Social Research. New York: Free Press.

Belson, W. A., 1981: The Design and Understanding of Survey Questions. Aldershot, England: Gower.

Belson, W. A., 1986: Validity in Survey Research. Aldershot, England: Gower.

Biemer, P. P./Groves, R. M./Lyberg, L. E./Mathiowetz, N. A./Sudman, S. (Hrsg.), 1991: Measurement Errors in Surveys. New York: Wiley.

Bishop, G., 1992: Qualitative Analysis of Question-Order and Context Effects: The Use of Think-Aloud Responses. S. 149-162 in: N. Schwarz/S. Sudman (Hrsg.), Context Effects in Social and Psychological Research. New York: Springer.

Blair, E. 1986: Processes used in the Formulation of Behavioral Frequency Reports in Surveys. S. 481 - 487 in: Proceedings of the Section on Survey Research, American Statistical Association, 1986.

Blair, J./Presser, S., 1993: Survey Procedures for Conducting Cognitive Interviews to Pretest Questionnaires: A Review of Theory and Practice. Survey Research Center, Universtity of Maryland.

Bolton, R., 1993: Pretesting Questionnaires: Content Analyses of Respondents' Concurrent Verbal Protovols. S. 280-303 in: Marketing Science, 12, 1993.

Bortz, J., 1984: Lehrbuch der empirischen Forschung für Sozialwissenschaftler. Berlin: Springer.

Boyd, H. W./ Westfall, R., 1964: Marketing Research. Text and Cases. Homewood, Ill.: R. D. Irwin.

Bradburn, N./Sudmann, S., 1979: Improving Interview Method and Questionnaire Design. Chicago: Jossey-Bass.

Cannell, C. F./Axelrod, M., 1956: The Respondent Reports on the Interview, in: American Journal of Sociology, 62, 1956.

Cannell, C./Lawson, S./Hausser, D., 1975: A Technique for Evaluating Interviewer Performance. Ann Arbor: The University of Michigan; Survey Research Center; Institute for Social Research.

Cannell, C./Kalton, G./Fowler, F., 1985: Techniques for Diagnosing Cognitive and Affective Problems in Survey Questions. Ann Arbor: The University of Michigan; Survey Research Center; Institute for Social Research.

Cannell, C./Oksenberg, L./Kalton, G./Bischoping, K./Fowler, F. J., 1989: New Techniques for Pretesting Survey Questions. Final Report. Ann Arbor: The University of Michigan; Survey Research Center. Boston: University of Massachusetts; Center for Survey Research.

Converse, J. M./Presser, S., 1986: Survey Questions. Handcrafting the Standardized Questionnaire. Beverly Hills: Sage.

DeMaio, Th.(Hrsg.), 1983: Approaches to Developing Questionnaires. Bureau of the Census. Office of Management and Budget. Statistical Working Paper 10.

Diekmann, A., 1995: Empirische Sozialforschung. Reinbek: Rowohlt.

Elliott, K./Christopher, M., 1973: Research Methods in Marketing. London: Holt, Rinehart & Winston.

Erbslöh, E., 1972: Techniken der Datensammlung I: Interview. Stuttgart: Teubner.

Ericsson, K. A./Simon, H. A., 1980: Verbal Reports as Data. S. 215 - 251 in: Psychological Review, 8, 1980.

Fink, A., 1995: The Survey Kit. Thousand Oaks: Sage.

Fink, A./Kosecoff, J., 1985: How to Conduct Surveys. Beverly Hills: Sage.

Foddy, W., 1993: Constructing Questions for Interviews and Questionnaires. Cambridge: Cambridge University Press.

Forsyth, B. H./Lessler, J. T., 1991: Cognitive Laboratory Methods: A Taxonomy. S. 393 - 418 in: P. P. Biemer/R. M. Groves/L. E. Lyberg./N. Mathiowetz /S. Sudman (Hrsg.): Measurement Errors in Surveys. New York: Wiley.

Fowler, F. J., 1984: Survey Research Methods. Beverly Hills: Sage.

Fowler, F. J., 1995: Improving Survey Questions. Thousand Oaks: Sage.

Fowler, F. J./ Mangione, T. W., 1990: Standardized Survey Interviewing. Beverly Hills: Sage.

Fowler, F. J., 1992: How unclear Terms Affect Survey Data. S. 218 - 231 in: Public Opinion Quarterly, 56, 1992.

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

Galtung, J., 1967: Theory and Methods of Social Research. New York: Columbia University Press.

Gordon, W. D., 1963: Double Interview. In: New Developements in Research. London: Market Research Society with the Oakwood Press.

Groves, R., 1989: Survey Errors and Surveys Costs. New York: Wiley.

Hippler, H.-J./Schwarz, N./Sudman, S. (Hrsg.), 1987: Social Information Processing. New York, Berlin: Springer-Verlag.

Hoinville, G./Jowell, R., 1978: Survey Research Practice. London: Heinemann Educational Books.

Hunt, S. D./ Sparkman, R. D./Wilcox, J. B., 1982: The Pretest in Survey Research: Issues and Preliminary Findings. S. 269 - 273 in: Journal of Marketing Research, Vol. XIX, 1982.

Hyman, H. H., 1955: Survey Design and Analysis: Principles, Cases and Procedures. Free Press, Glencoe, Ill.

Jabine, T. B./ Straf, M. L./ Tanur, J. M./Tourangeau, R. (Hrsg.), 1984: Cognitive Aspects of Survey Methology: Building a Bridge Between Disciplines. Washington, D. C.: National Academy Press.

Jobe, J. B./Mingay, D. J., 1989: Cognitive Research Improves Questionnaires. S. 1053 - 1055, in: American Journal of Public Health, 79, 1989.

Jobe, J. B./ Mingay, D. J., 1990: Cognitive Laboratory Approach Designing Questionnaires for Surveys of the Elderly. S. 518 - 524 in: Public Health Reports, 105, 1990.

Johnson, R. A./Woltman, H. F., 1986: Evaluating Census Data Quality Using Intensive Reinterviews: A Comparison of U.S. Census Methods and Rash Methods. S. 293 -.298 in: Proceedings of the Section on Survey Research, American Statistical Association, 1986.

Kahn, R. L./Cannell, C. F., 1967: The Dynamics of Interviewing. New York: Wiley.

Karmasin, F./Karmasin, H., 1977: Einführung in die Methode und Probleme der Umfrageforschung. Wien: Hermann Böhlaus Nachf..

Kidder, L. H., 1981: Research Methods in Social Relations. New York: Holt, Rinehart and Winston.

Kreiselmaier, J./Prüfer, P./Rexroth, M., 1989: Der Interviewer im Pretest. Mannheim: ZUMA-Arbeitsbericht 89/14.

Krueger, R. A., 1988: Focus Groups. A Practical Guide for Applied Research. Newbury Park: Sage.

Lansing, J. B./Morgan, J. N., 1971: Economic Survey Methods. Ann Arbor, Mi.: University of Michigan, Institute for Social Research.

Lessler, J. T./Forsyth, B. H., 1995: A Coding System for Appraising Questionnaires. In: N. Schwarz./S. Sudman (Hrsg.), Answering Questions. San Francisco: Jossey-Bass.

Lessler, J. T./Kalsbeek, W. D., 1992: Nonsampling Error in Surveys. New York: Wiley.

Lin, N., 1976: Foundations of Social Research. New York: McGraw-Hill.

Loftus, E., 1984: Protocol Analysis of Responses to Survey Recall Questions. In: T. B. Jabine / M. L. Straf / J. M. Tanur ./ R. Tourangeau (Hrsg.), 1984: Cognitive Aspects of Survey Methology: Building a Bridge Between Disciplines. Washington, D. C.: National Academy Press.

Loftus, E., 1986: Survey Remembering. S. 193 - 207 in: Second Annual Research Conference, March 23 - 26, 1986, Bureau of the Census, Proceedings.

Loftus, E., 1990: Memory. Surprising new Insights into how we remember and why we forget. U.S.A.: Addison-Wesley Publishing Company.

Loftus, W. F./Klinger, M. R./ Smith, K. D./Fiedler, J. 1990: A Tale of two Questions: Benefits of Asking More Than One Question. S. 330 - 345 in: Public Opinion Quarterly, Vol. 54, 1990.

Morgan, D. (Hrsg.), 1993: Successful Focus Groups. Newbury Parks: Sage.

Morton-Williams, J., 1979: The Use of "Verbal Interaction Coding" for Evaluating a Questionnaire. S. 59 - 75 in: Quality and Quantity, 13, 1979.

Morton-Williams, J./Sykes, W., 1984: The Use of Interaction Coding and Follow-up Interviews to investigate Comprehension of Survey Questions. S. 109 - 127 in: Journal of the Market Research Society, 26, 1984.

Noelle, E., 1971: Umfragen in der Massengesellschaft. Reinbek: Rowohlt.

Noelle-Neumann, E., 1970: "Wanted: Rules for Wording Structured Questionnaires." S. 191 - 201 in: Public Opinion Quarterly, 34, 1970.

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

Nuckols, R.C., 1953: A Note on Pre-testing Public Opinion Questions. S. 119-120 in : The Journal of Applied Psychology, 37, 1953.

Oksenberg, L.,/Cannell, Ch./Kalton, G., 1991: New Strategies for Pretesting Survey Questions. S. 349 - 365 in: Journal of Official Statistics, 7, 1991.

Oppenheim, A. N., 1966: Questionnaire Design and Attitude Measurement. New York: Basic Books.

Payne, S. L., 1951: The Art of Asking Questions. Princeton, N.J.: Princeton University Press.

Porst, R., 1985: Praxis der Umfrageforschung. Stuttgart: Teubner.

Presser, S./Blair, J., 1994: Survey Pretesting: Do different Methods produce different Results? S. 73-104 in: Sociological Methodology, 1994.

Prüfer, P./Rexroth, M., 1985: Zur Anwendung der Interaction-Coding-Technik. S. 2-49 in: ZUMA-Nachrichten, 17, 1985.

Prüfer, P./Rexroth, M., 1996: Multi-Method-Pretesting. Manuskript zum Vortrag ZUMA-interne Fortbildung. Januar 1996.

Rea, L. M./Parker, R. A., 1992: Designing and Conducting Survey Research. San Francisco: Jossey-Bass Publishers.

Reynolds, N./Diamantopoulos, A./Schlegelmilch, B., 1993: Pretesting in Questionnaire Design: A Review of the Literature and Suggestions for Further Research. S. 171-182 in: Journal of the Market Research Society, 35, Nr. 2, 1993.

Rossi, P. H./Wright, J. D./Anderson, A. B. (Hrsg.), 1983: Handbook of Survey Research. New York: Academic Press.

Royston, P./Bercini, D./Sirken, M./Mingay, D., 1986: Questionnaire Design Research Laboratory. S. 703 - 706 in: Proceedings of the Section on Survey Research, American Statistical Association, 1986.

Sanchez, M. E.: Effects of Questionnaire Design on the Quality of Survey Data. S. 206 - 217 in: Public Opinion Quarterly, 56, 1992.

Scheuch, E. K., 1973: Das Interview in der Sozialforschung, in: R. König (Hrsg.), Handbuch der empirischen Sozialforschung, Bd. I.

Schnell, R./Hill, P. B./Esser, E., 1995: Methoden der empirischen Sozialforschung. München/Wien: Oldenbourg.

Schrader, A. , 1971: Einführung in die empirische Sozialforschung. Ein Leitfaden für die Planung, Durchführung und Bewertung von nicht-experimentellen Forschungsprojekten. Stuttgart: Kohlhammer.

Schuman, H., 1966: The Random Probe: A Technique for Evaluating the Validity of Closed Questions. S. 218-222 in: American Sociological Review, 31, 1966.

Schuman, H./Presser, S., 1981: Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context. New York: Academic Press.

Schwarz, N./Sudman, S. (Hrsg.), 1992: Context Effects in Social and Psychological Research. New York: Springer-Verlag.

Schwarz, N./Sudman, S. (Hrsg.), 1994: Autobiographical Memory and the Validity of Retrospective Reports. New York: Springer.

Sheatsley, P. B., 1948: Some Uses of Interviewer-Report Forms in: Public Opinion Quarterly, 11, 1947/1948.

Sheatsley, P. B., 1983: Questionnaire Construction and Item Writing. In: Rossi, P. H./Wright, J. D./Anderson, A. B. (Hrsg.): Handbook of Survey Research. New York: Academic Press.

Sletto, R. R., 1940: Pretesting of Questionnaires. S. 193-200 in: American Sociological Review, 5, 1940.

Someren, M. v./Barnard, Y./Sandberg, J., 1994: The Think Aloud Method. A Practical Guide to Modelling Cognitive Processes. London: Academic Press.

Sommer, R./Sommer, B. B., 1980: A Practical Guide to Behavioral Research. New York/Oxford: Oxford University Press

Sudman, S./Bradburn, N., 1974: Response Effects in Surveys.Chicago: Aldine.

Sudman, S./Bradburn, N., 1982: Asking Questions. A Practical Guide to Questionnaire Design. San Francisco: Jossey-Bass.

Sudman, S./Bradburn, N./Schwarz, N., 1996: Thinking About Answers. The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.

Tanur, J. M. (Hrsg), 1992: Questions about Questions. New York: Russell Sage Foundation.

Tränkle, U., 1983: Fragebogenkonstruktion. In: Feger, H./Bredenkamp, J. : Enzyklopädie der Psychologie, Bd. 2: Datenerhebung.

Turner, C. F./Lessler, J./Gfroerer, J. (Hrsg.), 1992: Survey Measurement of Drug Use: Methodological Studies. Rockville: National Institute on Drug Abuse. U.S. Department of Health and Human Sevices.

Turner, C. F./Martin, E. (Hrsg.), 1982: Surveying Subjective Phenomena. Cambridge, Mass.: Harvard University Press.

de Vaus, D.A., 1986: Surveys in Social Research. London: George Allen & Unwin.

Warwick, D. P./Lininger, C. A., 1975: The Sample Survey: Theory and Practice. New York: Mc Graw - Hill.

Wellenreuther, 1982: Grundkurs: Empirische Forschungsmethoden: Königstein: Athenäum.

Williamson, J./Karp, D./Dalphin, J.R., 1977: The Research Craft: An Introduction to Social Science Methods. Boston: Little, Brown and Co..

Willis, G. B./Royston, P./Bercini, D., 1991: The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. S. 251 - 267 in: Applied Cognitive Psychology, 5, 1991.

Wilson, T. D., 1985: Questionnaire Design in free Context of Informational Research. In: M. Brenner, J. Brown, J. and D. Canter: The Research Interview. London: Academic Press.