Computerunterstütztes Pretesting von CATI-Fragebögen: das CAPTIQ-Verfahren

Faulbaum, Frank; Deutschmann, Marc; Kleudgen, Martin

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

COMPUTERUNTERSTÜTZTES PRETESTING VON CATI-FRAGEBÖGEN: DAS CAPTIQ-VERFAHREN

FRANK FAULBAUM, MARC DEUTSCHMANN & MARTIN KLEUDGEN

Der vorliegende Beitrag stellt ein am Sozialwissenschaftlichen Umfragezentrum der Universität Duisburg-Essen entwickeltes computerunterstütztes, auf dem Prinzip des Behavior-Coding basierendes Pretestverfahren für CATI-Instrumente unter Feldbedingungen vor. Feldpretests (auch: Beobachtungs- oder Standardpretests) für CATI-Instrumente erfordern die Evaluation von CATI-Fragebögen im Rahmen einer computerunterstützten telefonischen Pretesterhebung. Die Registrierung der während der Pretestinterviews auftretenden Probleme mit einzelnen Fragen kann traditionell von den Pretestinterviewern entweder retrospektiv nach der Beendigung des telefonischen Pretests durch Ausfüllen eines speziellen Beobachtungsinstruments oder durch Paper- und Pencil-Notierung der Probleme während einer Frage-Antwort-Episode vollzogen werden. Weder von der retrospektiven Methode noch von der Methode des gleichzeitigen schriftlichen Eintrags sind, insbesondere bei schwierigen Erhebungsinstrumenten reliable Ergebnisse zu erwarten. Im zuletzt genannten Fall muss davon ausgegangen werden, dass der Interviewer in der Regel so sehr auf das eigentliche Interview und die computerunterstützte Eingabe der Antworten konzentriert ist, dass er mit der zusätzlichen Aufgabe, Probleme schriftlich festzuhalten, überfordert wäre. Das hier vorgestellte Verfahren versucht, diese Defizite zu umgehen und zugleich die Vorteile computerunterstützter Telefonumfragen für Pretests zu nutzen. Ein Vorteil besteht etwa in der Möglichkeit, größere, nach dem Zufallsverfahren von Gabler und Häder (1997) gezogene Pretteststichproben zu verwenden, um anspruchsvollere statistische Verfahren schon in der Pretestphase einzusetzen. Ein weiterer Vorteil liegt in der raschen Verfügbarkeit eines Pretest-Datensatzes. Das hier beschriebene CAPTIQ-Verfahren (CAPTIQ: Computer Assisted Pretesing of Telephone Interview Questionnaires) erlaubt die sofortige computerunterstützte Codierung der Problemarten und liefert einen Datensatz, auf dessen Grundlage die Pretestergebnisse in ihrer longitudinalen Qualität sichtbar gemacht werden können. Dies geschieht mit Hilfe einer grafischen Darstellung, die als IPG (Interview Process Graph) bezeichnet wird.

Ähnlich wie in einem Elektrokardiogramm lassen sich Probleme mit Anwortskalen, Antworttendenzen, Lernprozesse, etc, insbesondere bei längeren Itembatterien identifizieren.

Observational pretesting or standard pretesting of CATI-questionnaires is not unproblematic because the recording of observed respondent behavior has either to be carried out during the interview itself or after completion of the interview, prepared by filling out observational forms. Recording during the interview often places a heavy additional burden on the interviewer above and beyond conducting the interview properly. Recording after the interview presents a challenge to reliability. In this paper we present a method for Computer Assisted Pretesting of Telephone Interview Questionnaires (CAPTIQ) which allows respondents' behavior to be coded during the interview without burdening the interviewer too much. The interviewers are able to code without interrupting the flow of the interview itself. The pretest data for each question and each respondent collected by CAPTIQ may be seen as longitudinal data which can be represented by a graph called IPG (Interview Process Graph). The IPG, rather like an electrocardiogram, reveals any problem zones occurring during the interview. As a result, problems concerning response scales, information collected on the learning processes initialized by respondents while processing item batteries. Comprehension difficulties related to question wording or other factors also manifest themselves in oscillations of the IPG.. The paper describes the CAPTIQ-Method and presents an illustration of the IPG by evaluating a CATI-Questionnaire used for a nationwide survey about health and media use.

1. Zielsetzungen

Intensive Forschungsanstrengungen im Bereich der Optimierung von Pretestverfahren (vgl. z.B. Exposito/Rothgeb 1997; Presser/Blair 1994; Prüfer/Rexroth 1996) haben bisher noch nicht zu einem praktisch einsetzbaren Verfahren zur Optimierung von CATI-Fragebögen unter Feldbedingungen geführt. Der vorliegende Beitrag berichtet über ein computerunterstütztes Pretestverfahren, das diese Lücke schließen soll. Im Unterschied zu den kognitiven Laborverfahren wie z.B. „*think aloud*“, „*paraphrasing*“, „*probing*“, etc. verlassen sich reine Beobachtungspretests nur auf die passive Beobachtung des Befragtenverhaltens. Idealerweise sollten den Beobachtungsverfahren andere Verfahren wie z.B. die Anwendung von Appraisal-Systemen (vgl. Lessler/Willis 1999) zur Bewertung der Qualität von Fragen anhand bestimmter Kriterien oder kognitive Pretests zum Fragenverständnis vorausgegangen sein.

Das hier vorgestellte, als *CAPTIQ – Verfahren* (Computer Assisted Pretesing of Telephone Interview Questionnaires) bezeichnete Pretestverfahren für CATI-Instrumente

stellt einen ersten Vorschlag dar, eine Codierung des Befragtenverhaltens *computerunterstützt* an einer *großen Zufallsstichprobe der Zielpopulation* durchzuführen und dabei gleichzeitig in den natürlichen Ablauf eines CATI-Interviews zu integrieren, ohne dass dieser Ablauf durch die Intervieweraktivität der Codierung gestört würde (vgl. auch Deutschmann/Faulbaum/Kleudgen 2003). Insofern handelt es sich über den klassischen Beobachtungspretest hinaus, bei dem der Interviewer die Reaktionen des Befragten nur notiert, um eine Variante des Behavior Coding. In seiner klassischen Form erfolgt die Codierung *durch den Forscher nach dem Interview* auf der Grundlage von Tonbandaufnahmen. Rein technisch betrachtet, wäre eine akustische Aufzeichnung auch bei CATI-Interviews möglich. Da hierfür jedoch das vorherige Einverständnis des Befragten eingeholt werden muss, sind systematische Auswahleffekte zu erwarten, die die Qualität der zufällig gezogenen Preteststichprobe gefährden könnten. Darüber hinaus stellt das Wissen von der Aufzeichnung des Interviews beim Befragten eine entscheidende Abweichung von den realen Feldbedingungen dar, die sich möglicherweise erheblich auf die Reaktionen des Befragten auswirken kann.

Da die computerunterstützte Verhaltenscodierung nicht durch den Forscher selbst, sondern durch Interviewer während eines normalen CATI-Interviews erfolgen muss, das sich im Vergleich zu anderen Erhebungsmethoden durch einen besonders starken Zeitdruck auszeichnet, muss das Codiersystem sehr einfach gehalten werden und für den Interviewer leicht handhabbar sein. Dennoch bleibt die Aufgabe des simultanen Interviewens und Codierens eine gewisse Belastung für den Interviewer, die nur durch eine ensprechende Schulung und die Auswahl kompetenter und erfahrener Interviewer reduziert werden kann.

Die Begleitforschung im Zusammenhang mit dem hier vorgestellten CAPTIQ-Verfahren ist noch nicht abgeschlossen. Insbesondere fehlen Studien zur Intercoder-Reliabilität, die zwangsläufig eine akustische Aufzeichnung erfordern würden.

2. Der Prozess der Codierung

Die im Folgenden dargestellten Codierregeln sind aus anderen Systemen des Behavior-Coding abgeleitet (vgl. Fowler/Cannell 1996; Morton-Williams 1979; Oksenberg/Cannell/Kalton 1991; Prüfer/Rexroth 1985) und an die Besonderheiten der computerunterstützten Telefonumfrage angepasst worden. Um eine simultane Codierung während des computerunterstützten Interviews zu ermöglichen, musste ein Weg gefunden werden, um das Codiersystem in die CATI-Software so zu integrieren, dass der Interviewer die Codierung während des Interviews bewerkstelligen kann. Dies geschieht dadurch, dass bestimmte Funktionstasten für spezifische Typen des Befragtenverhaltens reserviert werden.

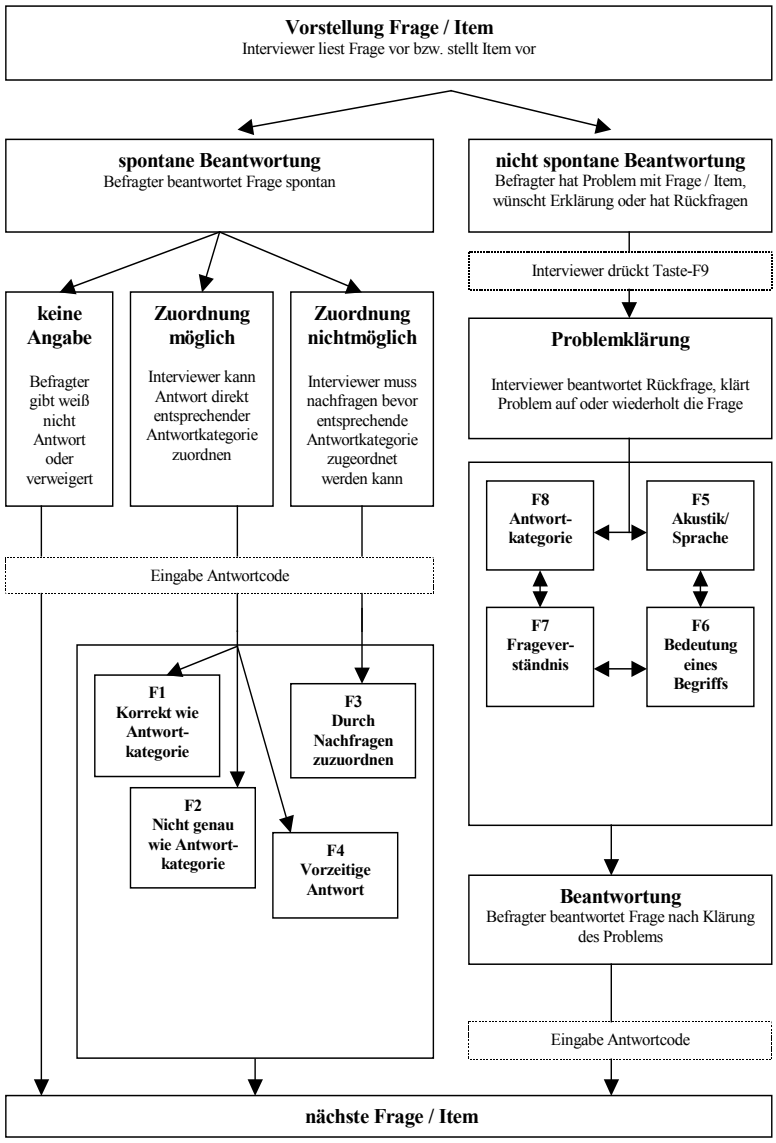
Die grundlegende Idee der Codierung des Befragtenverhaltens lässt sich durch das illustrieren, was Zouwen, Dijkstra and Ongena (2000) eine "paradigmatische Frage-Antwort-Folge" nennen. In einer paradigmatischen, idealen und unproblematischen Frage-Antwort-Folge liest der Interviewer korrekt die Fragen vor, stellt die Antwortkategorien vor und der Befragte gibt Antworten, die der Interviewer einer von mehreren möglichen Antwortkategorien zuordnen kann; d.h. der Befragte gibt nur *adäquate* Anworten. Es ist wohlbekannt, dass die Einstufung einer Antwort als adäquat nicht ausreicht, um die Bedeutung der Befragtenreaktion zu verstehen. So antworten etwa Befragte u.U. auch dann adäquat, wenn sie die Frage nicht verstanden haben. Leider gibt es im Rahmen eines Beobachtungspretests keine Möglichkeit zu entscheiden, wann dies der Fall ist. Ziel der hier beabsichtigen Codierung ist die Klassifikation der Antworten als *adäquat* oder *inadäquat* und die Identifikation *ausgewählter Typen nicht adäquater* Antworten. Da nur eine Codierung des Befragtenverhaltens und nicht des Interviewerverhaltens möglich ist, kann nicht entschieden werden, ob inadäquates Befragtenverhalten durch inadäquates Interviewerverhalten verursacht wurde. Die Wahrscheinlichkeit der zuletzt genannten Möglichkeit kann durch extensives Interviewertraining weiter reduziert werden. Überdies stellt dies bei großer Pretestichprobe und vielen Pretestinterviewern kein ernsthaftes Problem dar, da die Bedeutung systematischer Einflüsse durch Interviewer in der statistischen Analyse kontrolliert werden kann. Insbesondere können durch eine interviewerspezifische Auswertung (siehe unten) relative Schwächen der Interviewer aufgedeckt werden.

Abbildung 1 beschreibt Struktur und Prozess der Codierung sowie die während des Codierprozesses verwendeten Funktionstasten. Selbstverständlich kann die Zuordnung der Reaktionen des Befragten zu Funktionstasten auch anders programmiert werden. Das hier zugrundegelegte Codiersystem unterscheidet grundsätzlich zwischen zwei Arten von Befragtenverhalten:

- Spontane Beantwortung der Frage
- Nicht-spontane Beantwortung der Frage

Mit dem Begriff der *spontanen Beantwortung* soll ausgedrückt werden, dass der Befragte den Versuch unternimmt, in seiner ersten Reaktion auf die Frage eine adäquate Antwort zu geben. Diese Antwort kann entweder vollständig adäquat in dem Sinne sein, dass die Antwort durch den Interviewer einer der zulässigen Antwortkategorien zugeordnet werden kann oder sie ist insofern inadäquat, als der Interviewer eine Zuordnung nicht ohne Nachfragen zuordnen kann. Im Einzelnen können folgende *Varianten einer spontanen Beantwortung* unterschieden werden:

Abbildung 1: Ablauf der Codierung



- **Antwort entspricht genau einer Antwortvorgabe:** Befragter antwortet genau wie in der Antwortvorgabe vorgesehen, verwendet die gleichen Wörter wie in der Antwortvorgabe, Antwort ist problemlos zuzuordnen (Interviewer gibt den Antwortcode ein und codiert anschließend durch Drücken der Funktionstaste **F1**, dass die Frage problemlos zuzuordnen war).
- **Antwort entspricht nicht ganz genau einer Antwortvorgabe:** Befragter antwortet nicht genau wie in der Antwortvorgabe vorgesehen, verwendet andere oder ähnliche Wörter wie in der Antwortvorgabe, Antwort ist jedoch ohne weitere Nachfragen noch gut zuzuordnen (Interviewer gibt den Antwortcode ein und codiert anschließend durch Drücken der Funktionstaste **F2**, dass die Antwort nicht ganz genau der Antwortvorgabe entsprach).
- **Durch Nachfragen zuzuordnen:** Befragter antwortet zwar direkt, muss aber noch einmal gefragt werden, wo seine Antwort zugeordnet werden kann, Richtung der Antwort ist klar – die genaue Ausprägung aber noch nicht (Interviewer fragt nach, gibt die Antwort ein und codiert anschließend durch Drücken der Funktionstaste **F3**, dass eine Zuordnung nur durch Nachfragen möglich war).
- **Vorzeitige Antwort:** Befragter antwortet zu früh, kennt eigentlich noch nicht die gesamte Frage / Aussage und antwortet einfach während des Vorlesens der Frage/des Items (Interviewer gibt Antwort ein und codiert anschließend durch Drücken der Funktionstaste **F4**, dass die Antwort vorzeitig erfolgte).**Verweigerung oder „Weiß-nicht“** (direkte Eingabe der dafür vorgesehenen Antwortcodes Standardmode ohne anschließendes Drücken einer Funktionstaste).

Eine *nicht-spontane Beantwortung* liegt dann vor, wenn der Befragte in einer ersten Reaktion vom Interviewer eine Klärung verlangt, eher er eine Antwort abgeben kann oder sich zu einer „weiß nicht“-Antwort oder einer Verweigerung entschließt. In diesem Fall drückt der Interviewer in jedem Fall erst einmal die Funktionstaste F9, um damit zu codieren, dass der Befragte mit der Frage/dem Item Probleme hatte. Der Interviewer klärt dann in Übereinstimmung mit den Regeln des standardisierten Interviews, gibt nach Klärung die Antwort oder „weiß nicht“ oder „verweigert“ ein und codiert anschließend durch Drücken der Funktionstasten F5, F6, F7 oder F8, um was für eine Problemart es sich handelte, wobei folgende Problemarten unterschieden werden:

- **Akustik / Sprache:** Befragter hat den vorgelesenen Frage- bzw. Itemtext nicht richtig gehört, versteht die Sprache nicht gut, die Telefonverbindung ist schlecht, Geräusche in der Leitung (Interviewer drückt **F5**).
- **Bedeutung eines Begriffs:** Komplizierter Begriff wird nicht verstanden, Bedeutung von Fremdwörtern werden nicht genau verstanden, Befragter kennt einen Begriff oder ein bestimmtes Wort nicht (Interviewer drückt Funktionstaste **F6**).

- **Frageverständnis:** Befragter versteht die Frage/Item Aussage in ihrer Bedeutung nicht richtig, versteht den Sinn dieser Frage/Item nicht, versteht nicht, warum diese Frage gestellt wurde (Interviewer drückt Funktionstaste **F7**).
- **Antwortkategorien:** Befragter weiß nicht mehr, wie er zu antworten hat, hat die Antwortalternativen vergessen, zu komplizierte Antwortskala (Interviewer drückt Funktionstaste **F8**).

Diese Problemarten sind im Prinzip noch weiter differenzierbar, wobei bedacht werden muss, dass jede weitere Verfeinerung zusätzliche Anforderungen an das Training der Pretestinterviewer bedeutet. Dennoch wird gegenwärtig an einer Lösung des Problems gearbeitet, wie das Codiersystem weiter verfeinert werden kann, ohne dass das Interviewertraining zu anspruchsvoll wird. Insbesondere ist zu überlegen, ob die oben aufgeführten Problemarten in sich nicht noch zu heterogen sind und weiter zerlegt werden sollten.

Wie in jedem anderen System des Behavior-Coding, kann man auch in diesem Fall allein auf Basis der Pretestdaten nicht immer eindeutig entscheiden, ob die beobachteten Abweichungen vom idealen Frage-Antwortverlauf auf den Befragten, die Frage oder die Antwortkategorien zurückzuführen ist. Hierzu wäre eine detailliertere Betrachtung der entsprechenden Interviewsequenzen erforderlich. Durch das CAPTIQ-Verfahren wird aber eine Aufdeckung der problematischen Stellen im Interview ermöglicht. Da das Interviewerverhalten nicht beobachtet wird, kennen wir ebenfalls nicht das Ausmaß, in dem das beobachtete Befragtenverhalten vom Interviewer selbst beeinflusst wird. Wenn etwa Äußerungen beobachtet werden, die in die Problemart „Akustik/Sprache“ fallen, so kann das mangelnde akustische Verstehen etwa durch eine undeutliche oder zu leise Sprache des Interviewers bei gutem Hörvermögen des Befragten oder an einem schlechten Hörvermögen bzw. der akustischen Inkompetenz des Befragen bei deutlichem Sprechen des Interviewers liegen.

3. Visualisierung und Analyse der Pretestergebnisse: Der Interview Process Graph (IPG)

Wie bereits oben erwähnt, bietet das CAPTIQ-Verfahren die Möglichkeit, größere, zufällig gezogene Pretteststichproben einzusetzen, was nicht nur den Einsatz komplexerer statistischer Verfahren wie verschiedener Varianten der Faktorenanalyse, der multidimensionalen Skalierung oder der Analyse von Strukturgleichungsmodellen erlaubt und damit den Einsatz fortgeschrittener Verfahren der Itemanalyse, der Überprüfung theoretischer Validität, Konstruktvaliditat und Reliabilität schon in der Pretestphase ermöglicht. In seiner Effizienz nicht zu unterschätzen ist vielmehr die Anwendung von Verfahren der

Visualisierung, die einen Überblick über die Probleme mit Fragen/Items im vollständigen Ablauf des Interviews, also im Längsschnitt erlauben.

So lassen sich etwa die Häufigkeiten und Prozentwerte der vergebenen Codes für jede eineinzelne Frage bzw. jedes Item in der originalen Reihenfolge der Fragen und Items im Fragebogen abtragen. Es entsteht ein Verlaufsdiagramm, das für jede Frage und jedes Item die Häufigkeit des Auftretens der entsprechenden Probleme sichtbar macht. Diese Form der Visualisierung der Codeverteilungen im Zeitverlauf des Interviews nennen wir *Interview Process Graph (IPG)*, wobei der Begriff unabhängig von der Art des oder der dargestellten Codes definiert ist. Wir können so einen IPG für Verständniscodes, für Verweigerungen, etc. unterscheiden. In einem IPG können auch Eigenschaften verschiedener Codeverteilungen, z.B. Verständnisprobleme *und* Verweigerungen, gemeinsam abgetragen werden. Ferner lassen sich verschiedene Codes zu allgemeineren Problemarten zusammenfassen. So kann man etwa auf der Basis der vergebenen Codes die Verteilungen der in- adäquaten Reaktionen im Zeitverlauf darstellen.

IPGs erlauben die Identifikation möglicher Problemzonen im Interviewverlauf und die Analyse von Problemen mit Fragen/Items im Kontext von Nachbarfragen/Items, was z.B. bei längeren Itembatterien von Interesse ist. Sie erlauben auch die Visualisierung von Lern- und Anpassungsprozessen, die im Interviewverlauf auftreten. So kann man etwa erkennen, wie schnell Befragte lernen, eine bestimmte Form von Antwortskala zu handhaben.

Das CAPTIQ-Verfahren wurde bisher in vier Umfragen des Sozialwissenschaftlichen Umfragezentrums der Universität Duisburg-Essen erprobt. Als Beispiel sei eine Umfrage zum Thema „Medien und Gesundheit“ herausgegriffen, die im Rahmen einer Studie des Deutschen Diabetes-Forschungsinstituts (DDFI) an der Heinrich-Heine-Universität Düsseldorf durchgeführt wurde.[1] Es ging in dieser bundesweiten Studie um das Mediennutzungsverhalten der Bevölkerung bei der Gewinnung gesundheitsbezogener Informationen. Das Instrument beinhaltete sowohl einfache Ja/Nein-Fragen zu gesundheitlichen Beschwerden und Krankheiten als auch umfangreichere Itembatterien mit Zustimmungsskalen zum Thema Gesundheit, Fragen zur Informiertheit über bestimmte Erkrankungen sowie Fragen zur Mediennutzung allgemein. Insgesamt bezogen sich 124 Fragen/Items auf diese Inhalte, hinzu kamen Screening-Fragen sowie demographische Fragen, die allerdings nicht in den Pretest mit einbezogen wurden.

Für den CATI-Pretest wurden sechs besonders erfahrene Interviewer ausgewählt. Mit ihnen und einigen Supervisoren wurden dann die angesprochenen eingehenden

1 Wir danken an dieser Stelle besonders Prof. Dr. Werner Scherbaum für die freundliche Kooperation in diesem Projekt.

Pretestschulungen vorgenommen. Insgesamt wurden 100 CATI-Pretestinterviews mit dem CAPTIQ-Verfahren durchgeführt, die nach dem Verfahren von Gabler/Häder (1997, 1998, 1999) rekrutiert wurden.

Nach Anwendung des Behavior-Codings erschien es sinnvoll, die Fülle der erhobenen Informationen zum Befragtenverhalten zu den folgenden Kategorien zusammenzufassen:

- Spontane, adäquate oder nahezu adäquate Beantwortung (Codes: F1, F2, F4)
- spontane, inadäquate Beantwortung (Code: F3)
- nicht-spontane Beantwortung aufgrund eines Problems (Codes: F5, F6, F7, F8).

Mit dieser Klassifikation kann der gesamte Interviewverlauf im Interview-Process-Graph (IPG) gut und übersichtlich visualisiert werden. Es zeigt - ähnlich wie ein Elektrokardiogramm (EKG) die Ströme des menschlichen Herzens - die Problemzonen im Verlauf eines gesamten Telefoninterviews. Einfache und schwierige Antwortskalen können so genau identifiziert werden wie auch einsetzende Lernprozesse seitens des Befragten bei umfassenden Itembatterien. Verständnisprobleme und Schwächen einzelner Fragen machen sich durch extreme Ausschlägen im IPG bemerkbar.

Der in Abbildung 2 dargestellte IPG enthält die Prozentwerte für alle 124 Fragen mit inhaltlichem Bezug, mit denen die oben dargestellten Verhaltenskategorien bei den einzelnen Fragen/Items auftraten. Trotz ihrer eingeschränkten Lesbarkeit wurde die Grafik dargestellt, um einen visuellen Gesamteindruck des IPG zu ermöglichen.

Interessant scheint auch ein weiteres Phänomen, das bei dieser Itembatterie besonders gut zu sehen ist, aber auch bei anderen Itemfolgen auftaucht. Die erstgenannten Items zeigen gegenüber den letztgenannten schlechtere Werte im Antwort- und Befragtenverhalten. Dies könnte damit begründet werden, dass der Befragte seine zu erfüllende Aufgabe entweder nach jedem Item besser beherrscht oder der Befragte in bestimmte Antworttendenzen verfällt. Die Vorlage des ersten Items FR18_1 verursacht noch bei 17% der Befragten Probleme, so dass eine spontane Zustimmung oder Ablehnung ausbleibt. Diese Probleme sind vor allem auf das Item- bzw. Frageverständnis (7%) und Probleme mit der Antwortkategorie (6%) zurückzuführen. Bei 4% der Befragten musste das Item lediglich wiederholt vorgelesen werden. Die darauffolgenden Items verursachten deutlich weniger Probleme. Die entsprechenden Prozentwerte aus dem IPG sind in Tabelle 1 noch einmal zusammengefasst. Innerhalb dieser und anderer Itembatterien könnte also die steigende Anzahl spontan gegebener adäquater Antworten und die abnehmende Anzahl inadäquater und nicht-spontan gegebener Antworten mit einem einsetzendem Lernprozess des Befragten erklärt werden.

Abbildung 2: Beispiel eines Interview-Process-Graph (IPG)

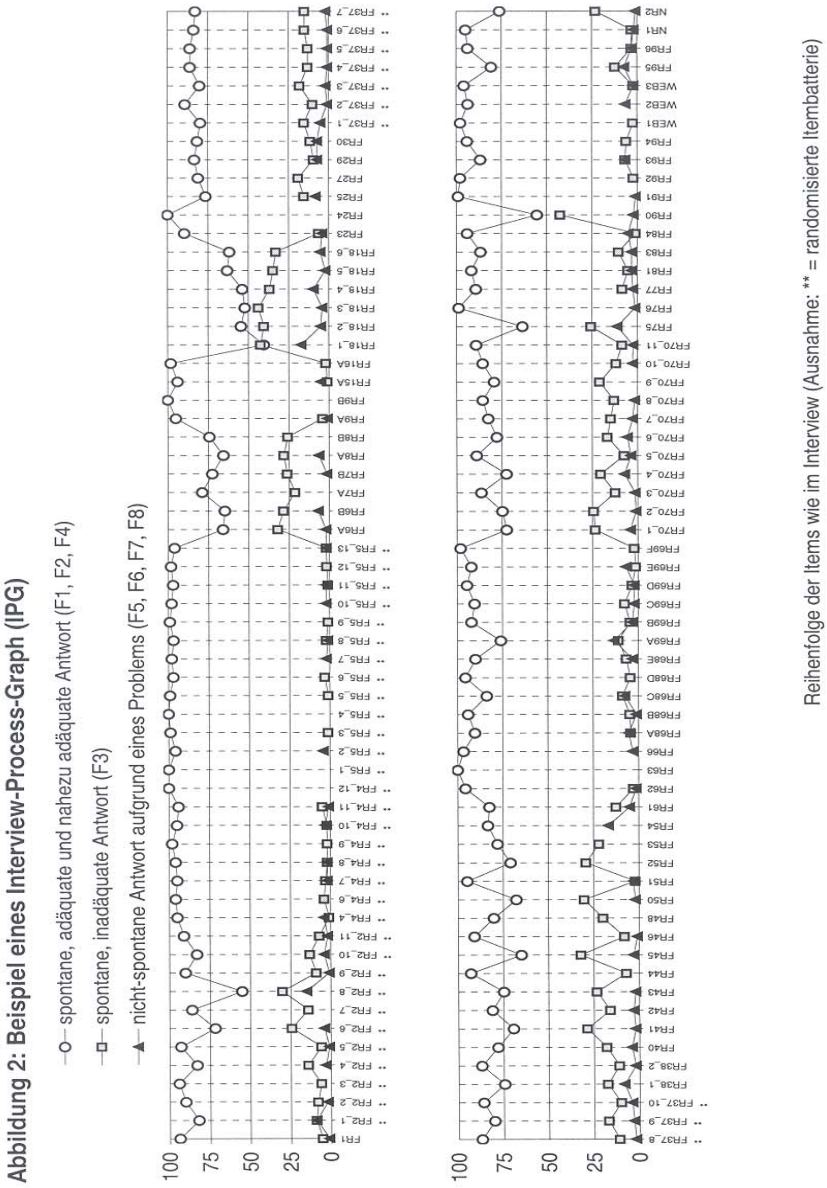


Tabelle 1: Anteilswerte adäquater und inadäquater Antworten (N=100)

	spontane, adäquate und nahezu adäquate Antwort (F1, F2, F4)	spontane, inadäquate Antwort (F3)	nicht-spontane Antwort aufgrund eines Problems (F5, F6, F7, F8)
FR18_1	40,4	42,4	17,2
FR18_2	54,5	40,4	5,1
FR18_3	52,1	43,8	4,2
FR18_4	53,7	36,8	9,5
FR18_5	63,0	34,8	2,2
FR18_6	61,7	33,0	5,3

4. Befragten- und interviewerspezifische Analysen

4.1 Befragtenspezifische Analyse

Bisher wurde die itemspezifische Analyse des Antwort- und Befragtenverhaltens unter dem Aspekt der qualitativen Bewertung des Instruments geschildert. Aufgrund der Möglichkeit, bei dem CAPTIQ-Verfahren relativ große Stichprobenumfänge einsetzen zu können, scheint die Methode aber auch für befragten- und interviewerspezifische Analysen des Befragtenverhaltens geeignet zu sein. So lassen sich etwa Fragen beantworten wie z.B.: Gibt es Befragtengruppen, die besonders viele Probleme bei bestimmten Fragen oder Formulierungen haben? Oder: Welche Merkmale der Befragten haben einen Einfluss auf ihr Antwortverhalten im Interview?

Als Beispiel greifen wir einige demographische Merkmale der Befragten (Geschlecht, Alter, Bildungsgrad) heraus und setzen sie in Beziehung zum Antwortverhalten. Tabelle 2 zeigt die Summenwerte der zusammengefassten Codes über alle Fragen/Items sowie über alle Interviewer. Eine Analyse der inadäquaten Beantwortungen ergibt durchaus sichtbare Unterschiede. Personen im Alter ab 45 Jahren und Personen mit niedrigerem Bildungsabschluss geben deutlich mehr spontane inadäuate Antworten als der Durchschnitt, Personen im Alter von 60 Jahren und darüber sowie Personen mit niedrigerem Bildungsabschluss geben auch mehr nicht-spontane inadäquate Antworten.

Dies sind zwar keine besonders überraschenden Ergebnisse, sie unterstreichen aber an dieser Stelle die Plausibilität des vorgestellten Verfahrens. Ähnliches fanden auch Prüfer und Rexroth (1985) in ihren Arbeiten über die Interaction-Coding-Technik heraus sowie

Reuband (1998) in seiner Befragung der Interviewer über der Verständnisprobleme seitens der Befragten.

Tabelle 2: Anteile demografischer Merkmale (N=100)

		spontane, adäquate und nahezu adäquate Antwort (F1, F2, F4)	spontane, inadäquate Antwort (F3)	nicht-spontane Antwort aufgrund eines Problems (F5, F6, F7, F8)
Geschlecht	männlich	86,8	9,9	3,3
	weiblich	84,2	13,0	2,8
Alter	16 - 29 Jahre	89,8	7,3	2,9
	30 - 44 Jahre	86,9	9,7	3,3
	45 - 59 Jahre	83,3	14,1	2,5
	60 Jahre und mehr	78,6	17,9	3,6
Schulabschluss	niedriger Schulabschluss	80,8	15,9	3,3
	höherer Schulabschluss	88,2	9,1	2,7
Gesamt		**85,3**	**11,6**	**3,0**

4.2 Interviewerspezifische Analyse

Mit umfangreicheren Pretestichproben lässt sich auch der Einfluss der Pretestinterviewer auf die Ergebnisse des Behavior-Codings überprüfen. Bereits auf der Ebene einfacher deskriptiver Analysen wird sichtbar, ob die Interviewer sich in ihrer Verhaltensklassifikation von einander unterscheiden. Da die Befragten bzw. die Telefonnummern durch das CATI-Programm nach Zufall auf die Interviewer verteilt wurden, haben wir eine Randomisierung über den Faktor „Interviewer“ vorliegen. Tabelle 3 stellt für jeden Interviewer die prozentualen Häufigkeiten der von ihm vergebenen Problemkategorien dar, und es werden durchaus Unterschiede sichtbar.

Während z.B. Interviewer BM bei durchschnittlich 6% seiner Interviews Antwortprobleme vercodete, waren es bei den Interviewern GA und ZI nur 1,6%. Auf der anderen Seite weist Interviewer ZI den höchsten Anteil an Befragten mit der Klassifikation „spontane inadäquate Antwort“ auf. Die Ergebnisse deuten darauf hin, dass ein solcher Pretest trotz intensiver Interviewerschulung nicht frei von Interviewereffekten ist. Das durch

diese Methode aufgedeckte quantitative Ausmaß dieser Effekte kann aber bei der Bewertung der Pretestergebnisse entsprechend mit berücksichtigt werden.

Tabelle 3: Codierung der Interviewer im Vergleich

	Anzahl der durchgeführten Interviews	spontane, adäquate und nahezu adäquate Antwort (F1, F2, F4)	spontane, inadäquate Antwort (F3)	nicht-spontane Antwort aufgrund eines Problems (F5, F6, F7, F8)
Interviewer: AE	12	84,3	11,5	4,2
Interviewer: BM	18	81,7	12,3	6,0
Interviewer: GA	20	92,2	6,2	1,6
Interviewer: KA	11	86,2	11,1	2,7
Interviewer: SC	12	86,1	10,5	3,3
Interviewer: ZI	27	82,9	15,5	1,6

5. Schlussbemerkung

Das computerunterstützte Behavior-Coding unter Feldbedingungen mit großen Zufalls-Preteststichproben nutzt die spezifischen Vorteile von CATI-Umfragen und ermöglicht Auswertungen, die bei den üblichen persönlich/mündlichen Pretests nicht oder nur mit sehr großem Aufwand möglich sind. Das Verfahren arbeitet auf seinem gegenwärtigen Entwicklungsstand noch recht grob und ist daher nur beschränkt in der Lage, konkrete Vebesserungsvorschläge für bestimmte Fragen und Itembatterien anzubringen. An einer Optimierung der Methode in dieser Richtung wird gegenwärtig gearbeitet. Immerhin ist es in der Lage, nicht nur einen Eindruck des Umfangs bestimmter Problemarten zu vermitteln, sondern auch detailliert bestimmte Problemzonen des Erhebungsinstruments im Längsschnitt aufzuzeigen. Die so identifizierten Bereiche können dann in einem weiteren Schritt mihilfe kognitiver Laborverfahren untersucht werden, um gesicherte Hinweise auf die Art der erforderlichen Veränderungen zu bekommen.

Das CAPTIQ-Verfahren ist für den routinemäßen Einsatz zum Pretesten von CATI-Instrumenten entwickelt worden. Die Module zur Codierung des Befragtenverhaltens lassen sich bei Verwendung der selben Software und mutmaßlich auch in alternativer Software im Prinzip für jeden CATI-Fragebogen implementieren. Damit ist es möglich, mit begrenztem Aufwand und in relativ kurzer Zeit einen Fragebogen unmittelbar vor der Hauptfeldzeit zu testen. Falls danach keine Änderungen an den Fragen nötig sind, lassen sich die erhobenen Pretestdaten ohne weiteres für die Haupterhebung nutzen, da der Pretest unter realen Feldbedingungen stattfindet und für die Befragten nicht wahrnehmbar ist.

Korrespondenzadresse

Prof. Dr. Frank Faulbaum
Lehrstuhl für Sozialwissenschaftliche Methoden/Empirische Sozialforschung
Sozialwissenschaftliches Umfragezentrum
Universität Duisburg-Essen, Standort Duisburg
Lotharstraße 65
47048 Duisburg
email: faulbaum@uni-duisburg.de

Literatur

Deutschmann, M./Faulbaum, F./Kleudgen, M., 2003: Computer Assisted Pretesting of Telephone Interview Questionnaires (CAPTIQ). In: Proceedings of the American Statistical Association, Survey Research Section, New York: ASA.

Exposito, J. L./Rothgeb, J. M., 1997: Evaluating Survey Data: Making the Transition from Pretesting to Quality Assessment. In: Lyberg, L. et al. (eds.), Survey Measurement and Process Quality. New York: Wiley.

Fowler, F.J./Cannell, C.F., 1996: Using Behavioral Coding to Identify Cognitive Problems with Survey Questions. S. 15-36 in: Schwarz, N./Sudman, S. (Hrsg.), Answering Questions: Methodology for Determining Cognitive and Communicative Process in Survey Research. San Francisco, Jossey-Bass.

Gabler, S./Häder, S., 1997: Überlegungen zu einem Stichprobendesign für Deutschland. ZUMA-Nachrichten 41: 7-18.

Gabler, S./Häder, S., 1998: Probleme bei der Anwendung von RLD-Verfahren. S. 58-68 in: Gabler, S./Häder, S./Hoffmeyer-Zlotnik, J. (Hrsg.), Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag.

Gabler, S./Häder, S., 1999: Erfahrungen beim Aufbau eines Auswahlrahmens für Telefonstichproben in Deutschland. ZUMA-Nachrichten 44: 45-61.

Gallhofer, I. N./Saris, W.E., 2000: Formulierung und Klassifikation von Fragen, in: ZUMA-Nachrichten 46: 43-72.

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

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

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

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

Prüfer, P./Rexroth, M., 1996: Verfahren zur Evaluation von Survey-Fragen: Ein Überblick. ZUMA-Nachrichten 39: 95-115.

Prüfer, P./Rexroth, M., 2000: Zwei-Phasen-Pretesting. ZUMA–Arbeitsbericht 2000/8.

Reuband, K.H., 1998: Der Interviewer in der Interaktion mit dem Befragten – Reaktionen der Befragten und Anforderungen an den Interviewer. S. 138-155 in: Statistisches Bundesamt (Hrsg.), Interviewereinsatz und –qualifikation. Band 11 der Schriftenreihe Spektrum Bundesstatistik.

Schaeffer, N.C./Maynard, D.W., 1996: From Paradigm to Prototype and Back Again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews. S. 65-88 in: Schwarz, N./Sudman, S., (Hrsg.), Answering Questions: Methodology for Determining Cognitive and Communicative Process in Survey Research. San Francisco: Jossey-Bass.

Van der Zouwen, J./Dijkstra, W./Ongena, Y., 2000: What Characteristics of Questions in Survey-Interviews make the Interaction Between Interviewer and Respondent 'Problematic' or Even 'Inadequate'? Department of Social Research Methodology, Vrije Universiteit, Amsterdam. Paper to presented on the Fifth International Conference on Logic and Methodology, Köln, Oktober 2000.

Willis, G.B./Lessler, J.T., 1999: Question Appraisal System – 1999. Washington: Research Triangle Institute.