Computer Assisted Pretesting of CATI Questionnaires (CAPTIQ)

Faulbaum, Frank

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

COMPUTER ASSISTED PRETESTING OF CATI QUESTIONNAIRES (CAPTIQ)

FRANK FAULBAUM

1. Introduction

Observational or standard pretesting of CATI-Questionnaires is not easily performed since in the strict sense this would mean that the recording of observed respondent behavior has to be done *during* the interview process. In this case, the coding system has to be designed in such a way that its handling does not influence the interviewer-respondent interaction. Otherwise the pretest would no longer constitute a pure field pretest but rather a pretest under specific conditions. In contrary to laboratory pretest methods like cognitive procedures (*think aloud*, *paraphrasing*, *probing*, etc.), pure observational pretesting exclusively relies on passive observation of respondents' behavior (for an overview of pretest methods see Exposito/Rothgeb 1997; Presser/Blair 1994; Prüfer/Rexroth 1996). Below, we present a method for Computer Assisted Pretesting of Telephone Interview Questionnaires (CAPTIQ) which allows

- a *behavior coding* of the question-answer episodes in *real-time* under *field conditions (standard pretest)*, i.e. during the interview at that time where the episode really takes place without interrupting the natural flow of the interview;
- the *reliable identification of certain types of problems* occurring during the interview
- the assessment of *respondent and interviewer specific influences* on data quality on the basis of pretest data;
- the immediate transfer of codes to a data file while the interview process is going on;
- the using of *big random samples* in order to *reduce the sampling error* of pretest results and to do *more complex statistical analyses* already in the pretest stage of questionnaire development

(see Kleudgen/Faulbaum/Deutschmann 2001; Deutschmann/Faulbaum/Kleudgen 2003; Faulbaum/Deutschmann/Kleudgen 2003). The approach is considered to be a first attempt to integrate coding of response behavior into a normal CATI interview. Associated with the pretest procedure is a specific graphical presentation of pretest results which is called IPG (Interview Process Graph). The IPG like an electrocardiogram reveals the problem zones occurring in the complete interview. By this method of presentation, it is possible to identify problems with response scales as well as possible learning processes initialized by the respondents while going through item batteries. Problems of understanding and weaknesses in question wording manifest themselves in oscillations of the IPG.

Behavior coding of respondent behavior, which basically constitutes a variant of standard pretesting methodology, in its traditional form tries to classify response behavior along the dimension adequate vs. inadequate. The coding is done with respect to each question in the questionnaire. In principle, this could either be done by categorizing the responses *after the interview* or *during the interview*. The first variant has the disadvantage of requiring automatic recording of the whole interview which, in turn, at least in Germany, requires the consent of the respondents. Since this might disturb the pure field character of pretesting and might introduce a bias into response behavior, the decision was to use the second variant, i.e. coding the response behavior during the interview. While behavior coding of tape-recorded responses after the interview has the apparent advantage that it could be done *by the researcher* himself, coding during the interview requires that the coding is done *by trained interviewers*. This, however, is not easy to deal with because of the higher time pressure in case of telephone interviews. The interviewer has to do coding and interviewing at the same time without interrupting or delaying the interaction between interviewer and respondent which might constitute a heavy burden on the interviewer. This kind of multi-tasking demanded on the interviewer could be circumvented by letting the coding be done not by the interviewers but by specifically trained personnel equipped with separate computers and headsets who does the coding in parallel with the interview. This strategy, however, would also require the agreement of the respondents. Furthermore, for big sample sizes it requires a costly equipment.

Observation and categorization of response behavior during the interview process require a quite simple coding system which could easily be managed by the interviewers. Nonetheless, the simultaneous task of interviewing and coding puts some burden on the interviewers who have to be trained extensively. Only the most competent and experienced interviewers should be selected for the pretest phase.

2. Coding system and coding procedure

The coding principles used are derived from behavior coding systems described elsewhere (see Fowler/Cannell 1996; Morton-Williams 1979; Oksenberg/Cannell/Kalton 1991; Prüfer/ Rexroth 1985, 1996) and adapted to the properties of the telephone mode. In contrary to PAPI, to which most coding procedures originally refer computer assistance allows the integration of the coding system into the CATI software (and, in principle also the CAPI software) by reserving certain keys for particular types of respondent behavior.

The basic idea of coding respondent behavior can be illustrated by what Zouwen/Dijkstra/ Ongena (2000) called a "paradigmatic question-answer sequence". In a paradigmatic, ideal and unproblematic sequence, the interviewer poses each question correctly and the respondent gives an answer which the interviewer is able to assign to one of the response categories. This, in fact, means that the respondent only gives adequate responses. Thus, the central aim of behavior coding and its underlying coding system is to classify for each question occurring in the interview the adequacy or inadequacy of the respondents' answers and to identify certain types of inadequacy. Since no coding of the interviewer-behavior is done, i.e. no real interaction coding is involved, we cannot decide whether an inadequate behavior of the respondent has been caused by inadequate *interviewer* behavior. The latter possibility can only be ruled out by an extensive interviewer training. Moreover, if a sufficiently high number of respondents is pretested and many interviewers are involved, the problem is not so serious since systematic interviewer influences can be accounted for in the statistical analysis.

The coding system is described systematically in figure 1. The basic types of behavior categories upon which the coding system is based are:

- **Spontaneous answer to the question:** The respondent in his first reaction tries to give a direct answer to the question or refuses the question.
- **Non-spontaneous answer to the question:** The respondent in his first reaction wants a further clarification by the interviewer before she/he gives an answer, refuses or says „don't know". Thus, this class of responses collects all those which cannot be counted as direct attempts to select a response category.

To each of these classes there corresponds a number of behavior subcategories leading to a specific code. The codes are entered into the computer by the use of function keys in order to allow for a rapid input**.**

The behavior subcategories belonging to the above basic category types are:

Subcategories for "Spontaneous answer to the question":

- *Answer corresponds correctly to the response categories (response scale)* and can be assigned to the response categories including the categories „refuse" or „don't know" without any problem (Interviewer presses function key F1 in order to indicate that the answer was assignable without problems)
- *Answer does not exactly meet the response categories*, but the response can be assigned to the response categories without further probes by the interviewer (press function key F2)
- *Answer is assignable after further probes:* Respondent answers directly but must be asked, to which response category his answer should be assigned (press F3)
- *Anticipated answer:* Respondent answers already while the question is read by the interviewer (press F4)

Subcategories for "Non-spontaneous answers to the question":

- *Question understanding/acoustics/ language:* Respondent does not clearly understand the question because of acoustic reasons or he knows the language not well enough or the phone connection is bad and there is noise in the phone line (press F5).
- *Concept meaning:* The meaning of a concept is not understood, the respondent doesn't know the concept or the word (press F6)
- *Question comprehension:* Respondent doesn't understand the meaning (sense) of the question. He doesn't understand why the question was posed (press F7)
- *Response categories:* Respondent forgot the response categories, response scale too complicated (press F8**)**

Of course, various subcategories can be rearranged according to certain properties and collected in new specific response classes like "adequate" or "inadequate".

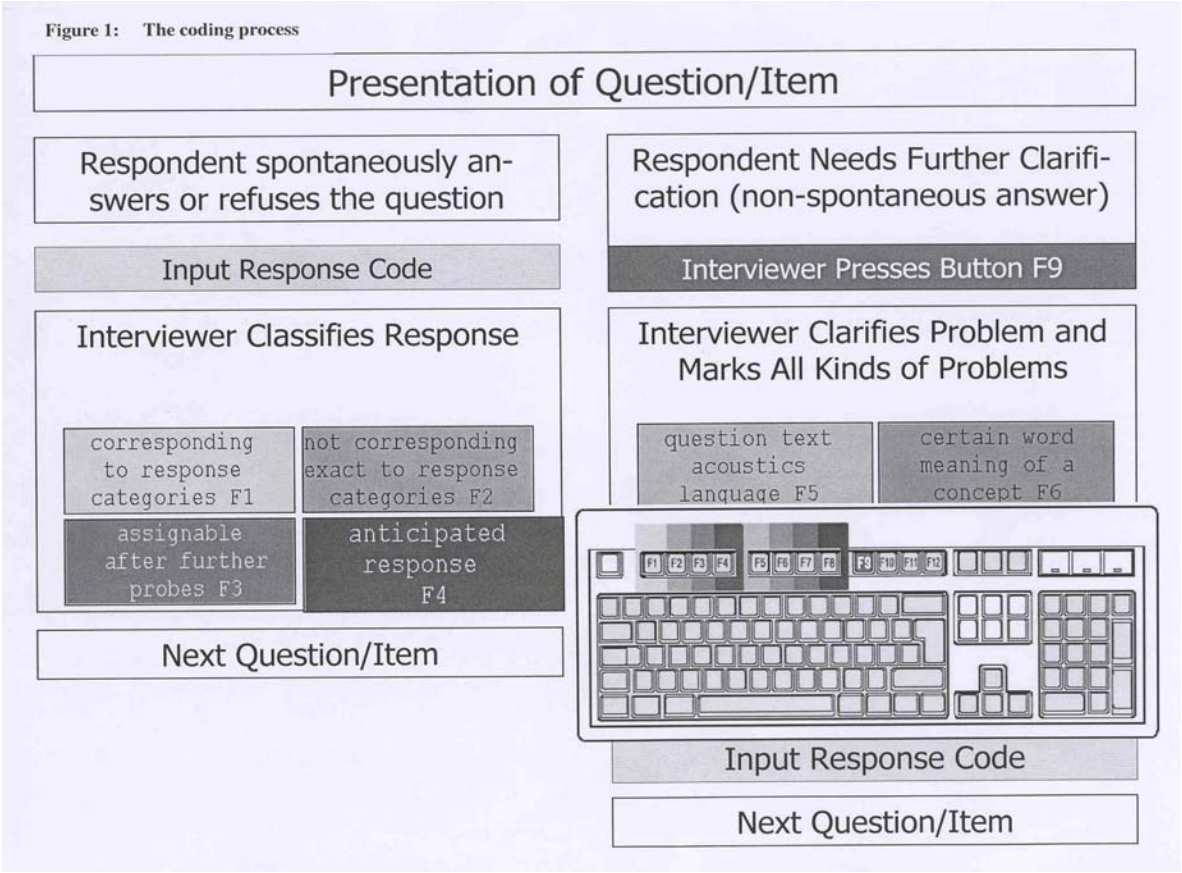
3. Analysis of pretest results

3.1 Structure of pretest data file and types of analyses

The pretest data file for each case contains the following information:

- Characteristics of respondent's interviewer (demographic variables, etc.
- For each question the response category including refusal information
- For each question and each coding category the classification code
- Further information about the interview (interview length, interviewer's impression about the respondent's behavior like cooperative attitude, etc.)

Figure 1: The coding process



These data admit different types of analyses: the analysis of interviewer differences in code statistics (frequencies, percentages) across questions and respondents, the analysis of differences in code parameters between types of respondents (male/female etc.) across interviewers and questions and the analysis of differences between questions or questions types in code parameters across interviewers and respondents. Examples of these types of analyses are given below. Of course, specific analyses for one interviewer, respondent or question can be done. A prerequisite for these analyses is a sufficient number of respondents and also questions. With a sufficient number of respondents also more complex statistical analyses like factor- and regression modeling or cluster analysis could be done.

3.2 Visualization of pretest results: The Interview Process Graph (IPG)

For each question of the questionnaire statistics of the different types of statistical coding results like frequencies, percentages, etc. of refusals and/or don't knows, of inadequate spontaneous responses, of comprehension problems, etc. can be plotted in various types of graphs we call *interview process graphs (IPGs)*. The horizontal axis of an IPG consists of the question numbers appearing in the same order as in the interview. The vertical axis refers to the statistics of certain types of coding. Thus, we can e.g. consider an IPG for the percentage of inadequate spontaneous responses, an IPG for total numbers of inadequate responses, an IPG for the percentages of meaning problems, etc.

IPGs allow for the identification of possible problem zones occurring during an interview and for the analysis of question/item problems in the context of neighbor questions/items which is especially important in case of big item batteries. They also permit the visualization of learning and adaptation processes occurring during the interview. One could e.g. visualize how fast the respondents learn to handle a certain type of response scale.

Figure 2 shows an example of an IPG. It is based on a CAPTIQ-pretest in a Health & Media Survey which dealt with media use and medical information seeking behavior. The sample size was 2.000. The questionnaire consisted of 124 questions of different types: simple yes/no questions about diseases and health problems, questions using various kinds of response scales for assessing the time dimension of health related behavior, item batteries for the identification of attitudes concerning different health topics using agreement scales as well as questions about knowledge of different diseases and the extent of media use in seeking medical information.

The size of the pretest sample was 100. The IPG in Figure 2 integrates different types of pretest information for all questions/items of the questionnaire: percentages of spontaneously given adequate and nearly adequate responses, percentages of spontaneously given inadequate responses and percentages of non-spontaneous answer

due to a problem. The codes defining theses response classes are indicated in the figure. The items indicated by a double star have been presented in a randomized fashion. We see that for some questions the percentages of adequate or nearly adequate responses were nearly 100 percent. An example are the thirteen questions named FR5_1 to FR5_13. The high percentages reflect the simplicity of the questions. The respondents were asked whether they already suffered from certain diseases. They had only to answer yes or no.

However, other items tell a completely different story. The item battery FR37_1 – FR37_10 introduced by the phrase „How do you feel personally informed about...“ followed by a list of different diseases like cancer/tumor, venereal diseases/Aids, heart condition, diabetes, etc. apparently seems to be more problematic. Respondents had to give a judgment on a verbal scale with respect to each disease. The scale values were (in English translation) „very well informed“, „well informed“, „somewhat informed“, „barely informed“, „not informed at all“ In 14% of all cases the interviewer could elicit an adequate answer only after further probes (spontaneous inadequate answer: Code F3).

A further example for weaknesses in an item battery is given by the six items named FR18_1 to FR18_6. The initial question was:

In the following I tell you some statements people sometimes make with respect to their health. Please tell me if you totally agree, almost agree, almost disagree or totally disagree.

Examples of items were:

- *My health is principally a matter of constitution and luck.*
- *My health is at first dependent of what I personally do.*
- *My health is determined by the physicians.*
- *Etc.*

On average, in 39% of the cases the respondents had to modify their spontaneous answers after probing by the interviewers in order to admit an assignment of the answer to an admissible response category. In addition, in 7% of the cases respondents apparently had problems and asked for clarification which may be seen as an indication of the larger complexity of task and a higher potential for response errors.

There is still another interesting finding which can well be illustrated by this item battery but which also occurs in other batteries. Items occurring earlier in the item list showed worse response behavior than items occurring later in the item list. This may either indicate the effect of further clarification in that respondents are becoming better in coping with the task in the sense of a learning process or that they return to constant response tendencies.

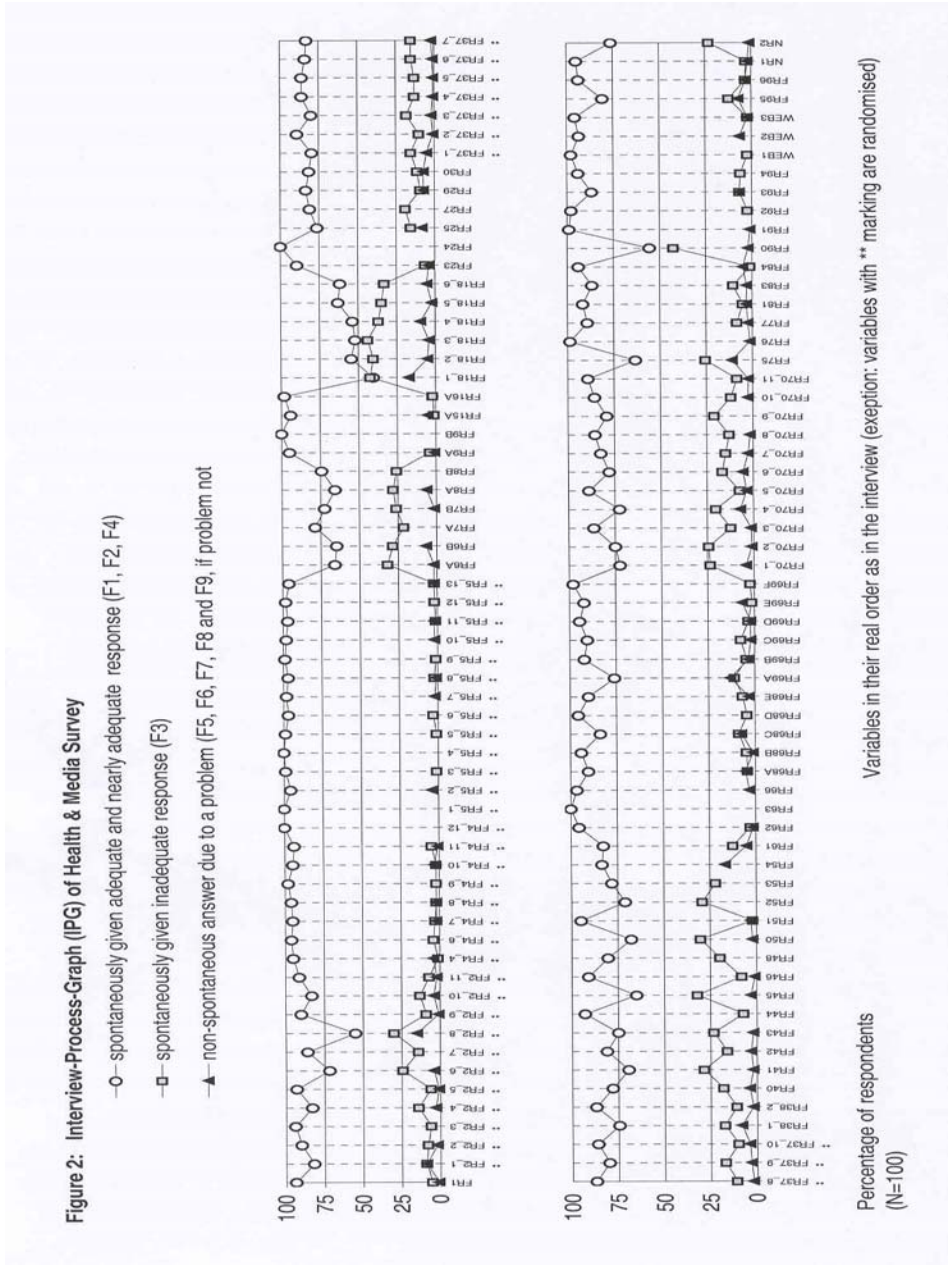


Figure 2: Interview-Process-Graph (IPG) of Health & Media Survey
–o– spontaneously given adequate and nearly adequate response (F1, F2, F4)
–□– spontaneously given inadequate response (F3)
–▲– non-spontaneous answer due to a problem (F5, F6, F7, F8 and F9, if problem not
Percentage of respondents (N=100)
Variables in their real order as in the interview (exeption: variables with ** marking are randomised)

The presentation of the first item FR18_1 causes problems for 17% of the respondents. The problems mainly concern the item or question understanding (7%) and problems with respect to the response categories (6%). In 4% of the cases the items only needed to be repeated by the interviewers. At the same time we observed an increase in the proportion of spontaneous adequate or nearly adequate answers (from 40.4% to 61,7%). The successive items were causing significantly less problems. The relevant percentages of the IPG are summarized once more in table 1.

Table 1: Proportions of adequate and inadequate answers

	spontaneous adequate or nearly adequate answer (F1, F2, F4)	spontaneous inadequate answer (F3)	non-spontaneous answer due to a problem (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

n=100

3.3 Respondent- and interviewer-specific analyses

3.3.1 Respondent-specific analysis

The preceding section concentrated on item-specific analyses of pretest data, i.e. on the quality of the instrument. The advantage of the CAPTIQ method is that it can handle larger sample sizes which also admit respondent- and interviewer-specific analyses. Thus, questions like “Are there specific subgroups of respondents having more problems with respect to certain types of questions than other subgroups” or “Which properties of respondents have the biggest influence on response behavior?” can, in principle be investigated.

As an example, let us consider the relationship between the demographic respondent variables “Gender”, “Age” and “Education” and the response behavior. Table 2 gives an overview of the proportions of various types of adequate and inadequate answers. The proportions are based on a summation of codes over items and interviewers. The table

shows significant differences between males and females. Females apparently give more spontaneous inadequate answers and more non-spontaneous answers due to a problem than males. The proportion of spontaneous inadequate answers also increases with age and decreases with education.

Table 2: Respondent-specific analyses: Demographic variables and response adequacy

		spontaneous adequate or nearly adequate answer (F1, F2, F4)	spontaneous inadequate answer (F3)	non-spontaneous answer due to a problem (F5, F6, F7, F8)
Gender	male	86.8	9.9	3.3
	female	84.2	13.0	2.8
Age	16 - 29	89.8	7.3	2.9
	30 - 44	86.9	9.7	3.3
	45 - 59	83.3	14.1	2.5
	60 years and more	78.6	17.9	3.6
Education	low	80.8	15.9	3.3
	high	88.2	9.1	2.7
Total		**85.3**	**11.6**	**3.0**

n=100

Though these results are far from surprising they underline the plausibility of the method. Similar results have been obtained by Prüfer/Rexroth (1985) in their work on interaction coding and by Reuband (1998).

3.3.2 Interviewer-specific analysis

Under the condition of big pretest sample sizes already simple statistical description may reveal interviewer differences with respect to the classification of behavior types. In the pretest example from the Health and Media Survey the respondents were randomly selected for the pretest sample and randomly assigned to the interviewers so that differences in proportions are not considered to be confounded with other background

variables. Table 3 shows for each interviewer the proportions of respondents who gave spontaneous adequate or inadequate answers non-spontaneous answers due to a problem.

It can easily be recognized that there are important differences between the interviewers. While, e.g., interviewer „BM“ coded non-spontaneous answers due to a problem in 6% of the cases, interviewers „GA“ and „ZI“ assigned these codes only in 1,6%. of the cases. Interviewer „ZI“ had the highest proportion of the behavior category „spontaneous inadequate answer. The results indicate that the intense interviewer training did not lead to a full standardization of coding behavior.

Table 3: Example of interviewer-specific analysis: Comparison of interviewers

	Number of complete interviews	Spontaneous, adequate or nearly adequate answer (F1, F2, F4)*	Spontaneous, inadequate answer (F3)*	non-spontaneous answer due to a problem (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

n=100
* percentages

4. Conclusions

The CAPTIQ-Method was specifically designed for evaluating CATI-Instruments with comparatively large pretest samples. The device is far from ideal. In fact, it has to rely on rather robust and rough coding principles. However, this does not mean that further refinements and modifications could not be done. In this respect the work presented here only represents a first step. What is needed in any case, are studies of intercoder reliability.

It is just the roughness of the method which guarantees its applicability to large pretest sample sizes which, in turn, allows for the application of more sophisticated statistical methods in the analysis of pretest data. Above, only the results of elementary inspections of the IPGs have been reported. More sophisticated analyses could involve factor

analyses and clustering of inadequate responses for the identification of problem types, methods of serial statistical analysis, subgroup analyses taking into account age, gender and other socioeconomic variables, etc.

The use of CAPTIQ is not limited to classical pretest applications which mainly concentrate on question quality. In addition, the method may also be used for the identification of interviewer-related as well as respondent-related causes of quality. Thus, response behavior is conceived to be decomposable into a respondent part, an interviewer part and a question wording part.

As a kind of observational pretest method CAPTIQ ideally should constitute the last member in a chain of pretesting stages all dealing with the improvement of the same instrument. It is clear that, at first, the standard rules for designing good questions should be followed (see Fowler 2001; Fowler/Mangione 1990) though in most research this is not the case. Also appraisal systems for questionnaires could be used (see e.g. Willis/Lessler 1999) at the first stage. The number of inadequate responses is expected to be substantially reduced if cognitive pretests are done before. In any case, the procedure serves diagnostic purposes. Though it is not able in every case to put into concrete terms what exactly has to be changed in the questions the procedure can give hints where to look for. It can also indicate problems not due to the question wording but rather to respondent- or interviewer-related properties.

CAPTIQ may also be useful if no extensive pretesting can be done. In most surveys which are not devoted to academic or governmental research but are done by commercial companies usually no extensive pretesting is taking place because of costs. Questionnaires are designed and then immediately submitted to the field. In these cases the method presented here could offer a quite cheap and routinely applicable method for the identification of severe questionnaire problems by inspecting the Interview Process Graph.

Contact

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

References

Deutschmann, M./Faulbaum, F./Kleudgen, M., 2003: Computer Assisted Pretesting of Telephone Interview Questionnaires (CAPTIQ). 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

Faulbaum, F./Deutschmann, M./Kleudgen, M., 2003: Computerunterstütztes Pretesting von CATI-Fragebögen. ZUMA-Nachrichten 52, S.20-34.

Fowler, F. J., 2001: Why it is easy to write bad questions. ZUMA-Nachrichten 48, S.49-66

Fowler, F. J./Mangione, Th. W., 1990: Standardized survey interviewing: Minimizing interviewer-related error. Newbury Park

Kleudgen, M./ Faulbaum, F./ Deutschmann, M., 2001: Computer assisted observational pretesting of CATI-questionnaires. Paper presented at the International Conference on Methodology and Statistics, Ljubljana.

Morton-Williams, J., 1979: The use of "Verbal Interaction Coding" Evaluating a questionaire. Quality and Quantity 13, 1979: S.59 – 75.

Oksenberg, L./Cannell, Ch./Kalton, G., 1991: New Strategies for pretesting survey questions. Journal of Official Statistics 7, S.349 – 365.

Porst, R., 1998: Im Vorfeld der Befragung: Planung, Fragebogenentwicklung, Pretesting. ZUMA-Arbeitsbericht, 98/02.

Presser, S./Blair, J., 1994: Survey pretesting: Do different methods produce different results? Sociological Methodology, S.73 – 104.

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

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

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

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 presented on the Fifth International Conference on Logic and Methodology, Köln, October 2000.

Willis, G. B./Lessler, J. T. (1999): Question Appraisal System-1999, Research Triangle Institute.