COMPUTER ASSISTED PRETESTING OF TELEPHONE INTERVIEW QUESTIONNAIRES (CAPTIQ)

Marc Deutschmann, Frank Faulbaum and Martin Kleudgen
Survey Research Center, University of Duisburg, Germany

KEY WORDS: Pretesting, Questionnaire Development and Evaluation, CATI

Basic aims

This paper is concerned with the development of a pretest method for evaluating questionnaires for computer assisted telephone interviews under field conditions, i.e. with *observational* or *standard pretesting* of CATI-Instruments. 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. Observational pretesting ideally should be preceded by cognitive methods applied specifically to get an impression of whether the respondents’ understanding of the questions corresponds to the understanding intended by the researcher.

In case of pretesting PAPI (paper & pencil) instruments the recording of respondents‘ behavior may be done in different ways, by tape recording, by interviewers‘ post interview recall of special problems occurring while responding to single questions, etc. (for an overview of pretest methods see Exposito & Rothgeb 1997; Presser & Blair 1994; Prüfer & Rexroth 1996).

The approach to be presented here is considered to be a first attempt to integrate coding of response behavior into the actual CATI interview. Behavior coding which, in fact, is constituting a variant of standard pretesting methodology, in its traditional form tries to classify response behavior with respect to whether it may be considered adequate or 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 agreement 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 of coding the response behavior during the interview.

While behavior coding of automatically 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 the interviewer*. This, however, is not easy to deal with because of the time pressure which is known to be highest in case of telephone interviews. The interviewer has to do the coding without interrupting or delaying the interaction between interviewer and respondent. Apparently, the coding by the interviewer rules out the possibility of coding the interviewer behavior since the interviewer should not be allowed to code his own behavior.

The interviewers‘ categorization of response behavior during the interview process has the consequence that the coding system has to be very simple and could be handled quite easily by the interviewer. Nonetheless, the simultaneous task of interviewing and coding puts a heavy burden on the interviewers who have extensively to be trained. In fact, only the most competent and experienced interviewers should be selected for the pretest phase.

The research presented here is still ongoing. Up to now, no studies of intercoder reliability have been performed. Such studies would, of course, require that interviews are recorded and independently coded by different interviewers after the interview.

Coding system

The coding principles used are derived from behavior coding systems developed elsewhere (see Morton-Williams 1979; Oksenberg, Cannell & Kalton 1991; Prüfer & Rexroth 1985) and adapted to the properties of the telephone mode. In contrary to PAPI, 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 and 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. It is well known that the registration of a response as adequate is not sufficient for gaining a full understanding of the meaning of the respondent’s reaction. This would, as already mentioned above, require the application of cognitive methods in order to get an impression of how the respondent understood the question and an assessment of the question’s validity. But this is not the topic of this paper which only concentrates on observational pretesting with all its limitations.

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 first distinction on which the coding system is based on is that between *spontaneous* and *non-spontaneous* responses. Usually, *non-spontaneous* responses are conceived as delayed responses. In the present case, non-spontaneous responses are conceived as responses by which respondents signalize that they need further assistance by the interviewer in order to give an adequate answer. Thus, this class of responses collects all those which cannot be counted as direct attempts to select a response category. A direct answer could not spontaneously be given because of problems to generate a response in the required format.

Spontaneous responses are further subdivided in refusals and don't knows, assignable answers and non-assignable answers. Refusals and don't know responses are understood in the usual sense. Assignable answers are those where the answers can be assigned without problems by the interviewer to one of the answer categories or to one of the scale points of a response scale. These may also include cases where respondents give answers which constitute small deviations from the given answer categories as well as answers given before the interviewer has completed the question reading. Non-assignable answers are those where the interviewer is not able to assign the answer and is therefore forced to evoke an adequate response by neutral probes compatible with the rules of standardized interviewing. The non-assignable spontaneous answers constitute the inadequate spontaneous answers in a narrower sense.

To each of the above mentioned possibilities of spontaneous assignable and non-assignable responses there correspond certain code inputs. For his coding the interviewer uses prescribed function buttons F1 to F9:

In summary, the codes in case of spontaneous response are:

F1: *Response corresponds to response categories:*
Respondent answers precisely in accordance with the prescribed response categories and uses the same response wording. The answer can be assigned accurately by the interviewer.
F2: *No perfect correspondence between response and response categories:*
Respondent's answer does not perfectly fit the response categories; he uses other or similar words, but responses can be assigned without problems.
F3: *Response assignable after further probes:*
Respondent gives an answer which cannot be assigned by the interviewer without further probes.
F4: *Anticipated response:*
Respondent answers before interviewer finished the question.

If the *answer* is classified as non-spontaneous, one of the following codes apply:

F5: *Question text, acoustics, language:*
Respondent is not able to perceive the text acoustically; he doesn't understand acoustically what the interviewer said; the telephone connection is bad.
F6: *Concept meaning:*
The meaning of a concept is not understood by the respondent; respondent doesn't know a certain word.
F7: *Question comprehension:*
Respondent doesn't understand the meaning of the whole question or item; he doesn't understand the reason why the question is posed.
F8: *Response categories:*
Respondent has forgotten the response categories, response scale too complicated.

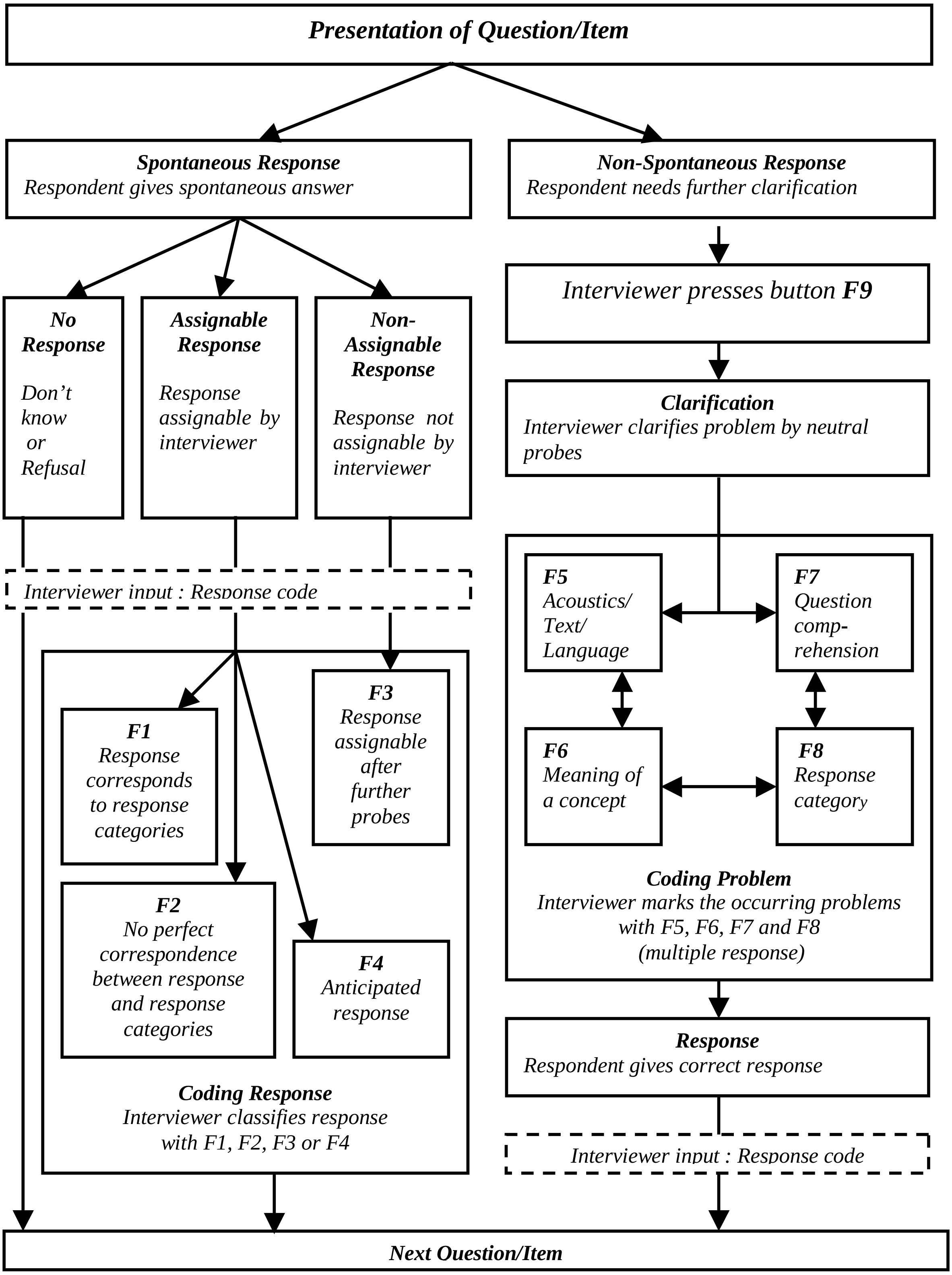
As in every behavior coding system we cannot in each case decide whether the observed deviations are due to the respondent or due to the question and its response categories. Since the interviewer behavior is not observed, we also don't know the extent to which the respondent behavior is influenced by interviewer behavior. With respect to some coding categories this influence may be more apparent as e.g. in the case of lack of acoustical comprehension. On the other hand, the respondent could have been rather old and suffering from acoustic incompetence.

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

For each question of the questionnaire statistics of the different types of 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 1: The coding process


Presentation of Question/Item
Spontaneous Response
Respondent gives spontaneous answer
Non-Spontaneous Response
Respondent needs further clarification
No Response
Don't know or Refusal
Assignable Response
Response assignable by interviewer
Non-Assignable Response
Response not assignable by interviewer
Interviewer presses button F9
Clarification
Interviewer clarifies problem by neutral probes
Interviewer input : Response code
F5
Acoustics/ Text/ Language
F7
Question comp- rehension
F6
Meaning of a concept
F8
Response category
F1
Response corresponds to response categories
F3
Response assignable after further probes
F2
No perfect correspondence between response and response categories
F4
Anticipated response
Coding Response
Interviewer classifies response with F1, F2, F3 or F4
Coding Problem
Interviewer marks the occurring problems with F5, F6, F7 and F8
(multiple response)
Response
Respondent gives correct response
Interviewer input : Response code
Next Question/Item

The CAPTIQ-Method has hitherto been used in two big surveys conducted by the Survey Research Center at the University of Duisburg. One survey, the Health & Media Survey, dealt with media use and medical information seeking behavior. The sample size was 2.000. In the second survey a random sample of 3.000 persons was asked about attitudes concerning gene examinations in case of psychic and mental illnesses. The examples selected for demonstrating the possibilities of the interview process graph were drawn from the Health & Media Survey. The questionnaire contained 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. Figure 2 shows an example of an IPG integrating 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. Let us turn, e.g. to the item battery containing 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 are:

- *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 interviewers could assign the spontaneously given responses only after additional probing. In 7% of the cases non-spontaneous responses due to a problem were given. Spontaneously given inadequate responses in absence of other types of inadequacy indicate that respondents used other response categories than they should use. Some of the respondents may have answered simply by “agree” or “disagree” and didn’t try to refine their responses according to the prescribed scale values. Table 1 allows a more detailed view on the results. In fact, it shows that in this item battery the percentages of adequate responses rise while those of inadequate ones decline. This fact indicates a certain learning effect. Nonetheless, the level of inadequate spontaneous response, remains considerable high (33.0%). On the basis of this result one should think about a revision of response scale values with respect to these items. The answering categories perhaps did not appear natural to the respondents. Perhaps the end points are not optimally chosen, etc.

Table 1: Results for item battery FR18_1 to FR18_6

	spontaneously given adequate and nearly adequate response	spontaneously given inadequate response	non-spontaneous answer due to a problem
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

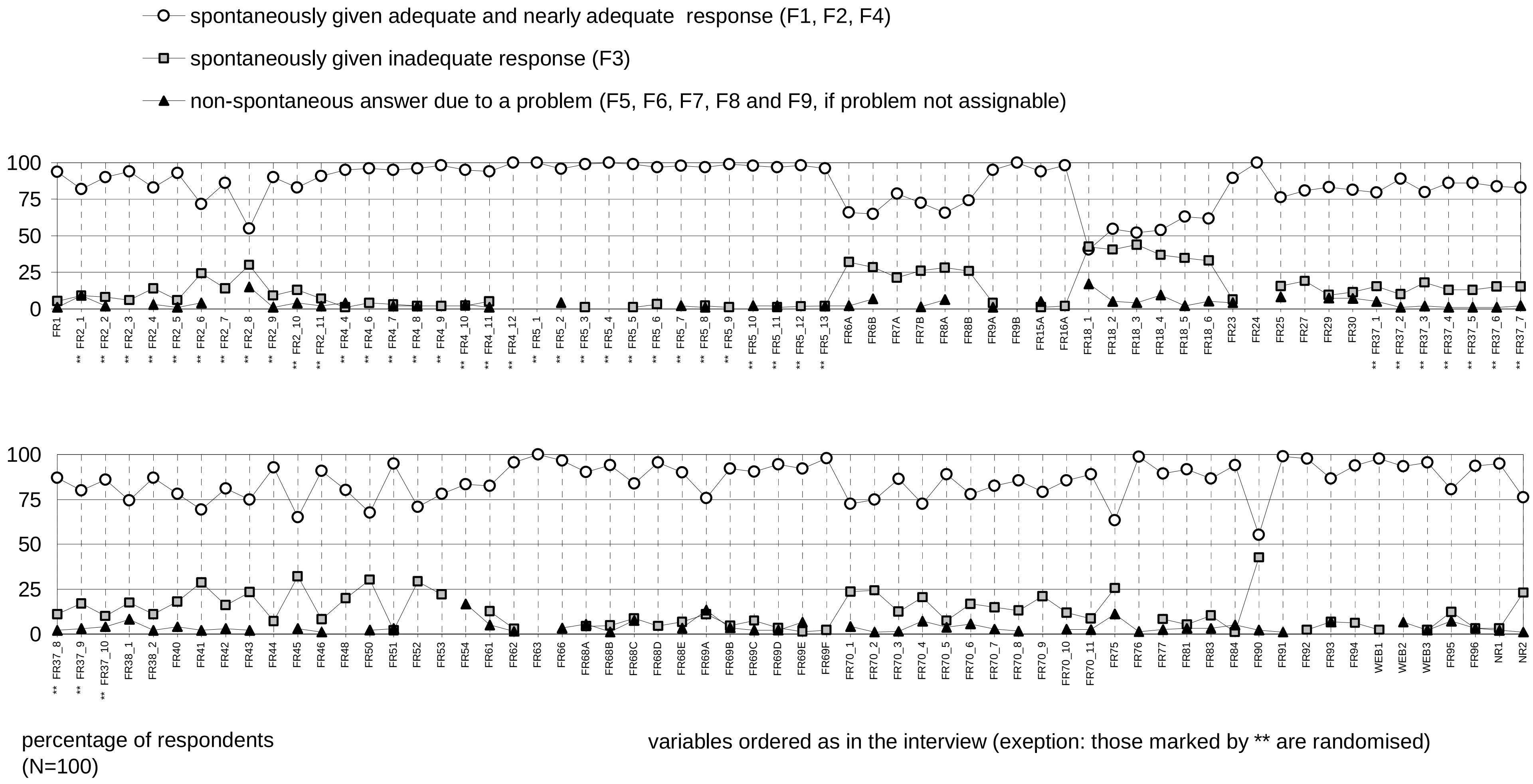
If we turn to the third column of table 1 we can see that the first item FR18_1 caused most of the non-spontaneous answers. A more thorough analysis revealed that most of these concerned comprehension problems. The subsequent items caused significantly less problems. This may indicate that the instruction for the use of the response scale which is intimately connected with the first item was perhaps not understood by the respondents.

Similar observations as those with respect to the item battery just referred to can also be made with respect to other types of question sequences. Of course, some of the systematic sequential effects like order or position effects disappear if randomization of item presentation is introduced.

An apparent advantage of an IPG is its use for the identification of item-specific effects. The battery of items F2_1** to F2_11** consisted in a set of randomly presented statements related to respondents’ feelings of well-being during the past seven days. All items with the exception of FR2_6** and FR2_8** were negative items; i.e. they described negative feelings while FR2_6** and FR2_8** described positive ones. The response categories were “very often”, “often”, “sometimes”, “seldomly”, “never”. As figure 2 shows the number of inadequate responses is highest for the two positive items. The items didn’t fit in to the negative context and apparently caused reactions of astonishment.

A final demonstration of the usefulness of IPGs concerns the detection of dependencies between scale values and frequencies of inadequate response. To this end, let us turn to questions FR40 to FR53 in figure 2.

Figure 2: Interview-Process-Graph (IPG) of Health & Media Survey



These questions referred to the frequency of use of various sources getting health-related information like television, internet, etc. All questions used the same response scale for the assessment of frequency of use. The category labels were: "daily", "at least one time per week", "at least one time per month", "more seldomly or never". In table 3 for selected questions FR45, FR46, FR51 and FR52 the percentages of the "never"-category and the frequency of different forms of inadequacy are tabulated. As can be seen, percentage of inadequacy covary negatively with the percentage of respondents having chosen the category "never".

Table 3: Dependencies between frequency of „never"-response and response inadequacy

	Proportions of „never"-response	Spontaneously given adequate and nearly adequate response	spontaneously given inadequate response	not spontaneous answer due to a problem
FR45	7,0	65,0	32,0	3,0
FR46	75,0	90,9	8,1	1,0
FR51	80,0	94,9	2,0	3,1
FR52	13,0	70,8	29,2	

This indicates that respondents had more problems if retrieval from their memory became more complex. It is simplest in case they never used the information source. All other response categories caused categorization problems. Improvements of the instrument after pretesting should have concentrated on the reformulation of the other response categories.

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 should not be done. In this respect it only represents a first step. What is needed in any case, are studies of intercoder reliability. But this makes sense only if the final stage of development is reached.

It is just the roughness of the method which guarantees its applicability to large pretest sample sizes. This, 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.

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). Perhaps the amount of inadequate responses would have been less if cognitive pretests had been done before.

Nonetheless, CAPTIQ is a useful and efficient method if no extensive pretesting could be done. In most surveys which are not devoted to academic or governmental research but are done by commercial firms usually no extensive pretesting is taking place. 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. Standard statistical analysis procedures could be applied to achieve this end.

References

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. (2001): Why it is easy to write bad questions, *ZUMA-Nachrichten*, 48: 49-66

Fowler, F. J. & Mangione, Th. W. (1990): *Standardized survey interviewing: minimizing interviewer-related error*. Newbury Park

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

Oksenberg, L., Cannell, Ch. & Kalton, G. (1991): New Strategies for Pretesting Survey Questions. *Journal of Official Statistics*, 7: 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*: 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.

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'? Paper presented on the Fifth International Conference on Logic and Methodology, Cologne, October 2000.

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