# On the use of the interaction coding technique

Prüfer, Peter; Rexroth, Margrit

# On the Use of the Interaction Coding Technique

by

Peter Prüfer and Margrit Rexroth

from: ZUMANACHRICHTEN Nr. 17, November 1985

(Title of the original: Zur Anwendung der Interaction-Coding Technik)

translated by Dorothy Duncan

# On the Use of the Interaction Coding Technique

## 1. Introductory Remarks

The discussion of "data quality" is gaining increasing importance in social science survey research. This discussion centers on the demand for valid survey data, which has grown markedly during the past few years. The methodological focus is on the data gathering process, since it is particularly in this area that certain influences may have a negative effect on data quality. Such effects may, for one, be caused by the interviewer when inappropriate behavior on his part leads to respondent reactions at odds with the question's purpose; unclearly worded questions can also have a negative effect on data quality.

In the past the attempt has been made to use interviewer training and pretests to counter certain influences exerted by interviewer and instrument, with the aim of discovering any existing flaws in the instrument and achieving the best possible starting conditions for data collection. In this connection it seemed of interest to test a technique that enables one through observation of the entire interview process to draw conclusions about both the interviewer and the instrument. This is the "interaction coding technique," based on the work of Charles Cannell et al.

The present article addresses the question of the extent to which this technique can help improve the quality of survey data. Two separate studies were carried out: The first was aimed at determining the suitability of the interaction coding technique for registering, analyzing and evaluating interviewer behavior. The second study centers on the instrument--the questionnaire or individual questions. The goal was to determine whether--along with traditional pretesting methods--the interaction coding technique provides additional information on the "functioning" of specific questions beyond that offered by conventional techniques.

As for the interviewer, he has been the subject of countless research projects in the past. This highlights the importance of his role in conducting empirical social research. Indeed, it is he--as the "agent" of the researcher--whose task is to gather data in the form of valid answers from respondents with the help of the instrument of the questionnaire. It is with good reason that the survey interviewer is referred to by Charles Cannell (1981) as the "gatekeeper" to the attitudes, experiences and perceptions of the respondent.

The research cited in the literature is in agreement that respondents' answers can be influenced by the interviewer in a number of different ways. Studies of these "interviewer effects" can be traced back as far as Rice (1929). The results show that these effects, i.e. undesirable changes in the data, can be caused by external characteristics of the interviewer such as sex, age or skin color, as well as by his behavior. It has become clear that the attitudes of the interviewer toward the topics covered in the questionnaire can indeed influence respondents' answers. Such influence can, for example, be caused by leading probes or questions used by the interviewer that steer the respondent in a particular direction; by the omission of non-directed questions; by erroneous recording of responses or by other nonverbal behaviors.

In such cases the quality of the data would be substantially reduced by "wrong" interviewer behavior. These interviewer-caused effects can be avoided or reduced by means of appropriate training. Two steps must be taken to achieve this end:

1. Observing and recording "wrong" interviewer behavior, i.e. behavior that deviates from the established rules;

2. Intervention measures to correct such "wrong" behaviors.

The interaction coding technique seems appropriate for implementing these steps. The question of the extent to which this technique is indeed capable of solving the above-mentioned tasks was the focus of the first study.

As noted above, the instrument or individual questions, along with interviewer behavior, can negatively affect data quality. However, determining the quality of a question is not unproblematical, since we have practically no objective, empirically-tested criteria to accomplish this. This unsatisfactory situation is described by Cannell et al. as follows:

> The least scientifically rigorous aspect of survey research is the development and testing of questions. It is ironic that the creation of the measuring instrument is based primarily on past experience with only a few "common sense" principles as guidance.

The usual practice in survey research has been to gain information on the quality of questions from the pretest that precedes the survey proper. This generally means that experienced interviewers conduct a number of interviews and report to the researcher on the problems caused by individual questions. These interviewer reports are both subjective and unsystematic, and usually limited to serious problems in the interviewing situation. Thus the final instrument (=questionnaire) is based primarily on the subjective assessment of the researcher and only in small part on empirical findings.

The literature contains attempts to formulate, on the basis of the author's own experience, generally valid rules for constructing questions. Frequently, however, such attempts are no more than rather trivial guidelines (e.g., "Avoid questions that might steer the respondent in a particular direction"). The very title of an early standard text containing guidelines on question wording (Payne, 1951: The Art of Asking Questions) provides some indication of what the 100 rules contained in the book are like. Significantly, the last paragraph reads as follows:

> Actually you won't need this check list type of stimulus for long because most of these things are only common sense anyway. Having once been pointed out, they should stay with you pretty well with perhaps only an occasional reading for a refresher.

The systematic studies of these topics, such as those by Schuman and Presser (1981) or Sudman and Bradburn (1982), published during the past few years--some of which, however, deal only with certain types of questions--make it clear that real efforts are being made to establish an empirical basis for question formulation. Our second study represents an attempt to use the interaction coding technique to help assess question quality through empirically-based conclusions; there were indications that the interaction between interviewer and respondent might shed light on question quality.

## 2. Study 1: Evaluation of Interviewer Behavior Using the Interaction Coding Technique

### 2.1. General Remarks on the Evaluating System

The first work done with the interaction coding system is based to a substantial degree on the system described in Cannell et al. (1975). The technique is set up to allow for the analysis and simultaneous evaluation of interviewer behavior in a face-to-face interview. This technique is simpler to apply than is the technique the authors modified for later work and used for determining question quality; thus the technique for evaluating interviewer behavior described in Cannell's work offers a good starting point for the testing of such evaluation in general. It is not the entire process of social interaction between respondent and interviewer that is analyzed using this technique, but only interviewer behavior during the survey. In Study 1, then, the technique focuses on only one of the participants.

## 2.2. Structure and Application of the Technique -- The Coding System

### 2.2.1. Structure of the Technique

The prerequisite for applying this technique is that the interview be tape-recorded. The interviewer also writes down the respondent's answers on the questionnaire, so that the written record can be checked against the tape-recording. When the tape is played back, all verbal activity of the interviewer during the interview is evaluated using a detailed coding system. The coding system consists of a listing of all behaviors (all verbal activities) that might be exhibited by an interviewer during the interview. These behaviors are divided into four categories:

Category I: includes all behaviors that have to do with asking the question;

Category II: those behaviors that concern clarification and non-directed probes;

Category III: encompasses all other behaviors;

Category IV: concerns maintaining the set order.

Within each of these categories a distinction is made between appropriate and inappropriate behaviors. This division is based on certain rules for interviewer behavior, drawn from the basic conception of the specific type of questionnaire. It should be noted that even in standardized interviews there is no absolutely binding set of rules for all possible behaviors. For example, there are rules regarding the introductory phase of an interview, the establishment of the interview atmosphere and encouragement to be provided to the respondent, which are to be applied at the discretion of the individual researcher. However, there are rules that can be regarded as generally binding and that provide the interviewer with an orientation for his behavior in a standardized interview. These include, for example: reading the text of the question as written, maintaining question order, strict adherence to the interviewer's instructions, absolute neutrality etc.

A less standardized questionnaire type would require the coding system to be based on rules adapted to meet the particular situation. The technique described here deals exclusively with the standardized interview. For pragmatic reasons, a code value is attached to each behavior. As shown in Table 1, this classification is based on a certain system.

Table 1: Overview of the code groups in the coding scheme

| | Code group | Code value |
|---|---|---|
| I. ASKING THE QUESTION | | |
| Appropriate behavior (correctly asking the question) | 10 | 11-13 |
| Inappropriate behavior (incorrectly asking the question) | 20 | 21-23 |
| II. CLARIFICATION/NON-DIRECTED PROBES | | |
| Appropriate behavior (non-directive clarification or non-directed probes) | 30 | 31-36 |
| Inappropriate behavior (directive clarification or non-directed probes) | 40 | 41-47 |
| III. OTHER BEHAVIOR | | |
| Other appropriate behavior | 50 | 51,58 |
| Other inappropriate behavior (verbal) | 60 | 62-68 |
| Other inappropriate behavior (nonverbal) | 70 | 71-75 |
| IV. SKIP INSTRUCTIONS | | |
| Correctly following skip instructions | 80 | 81 |
| Error in following skip instructions | 90 | 91,92 |

Thus, for example, all appropriate behaviors in Category I, "Asking the question," are included in Code Group 10. Within this Group 10 codes 11, 12 and 13 stand for appropriate, correct behavior in reading the question text. All inappropriate behaviors in this category are included in Code Group 20. Here codes 21, 22 and 23 stand for inappropriate behavior in reading the question text.

### 2.2.2. Assigning Code Values

Each activity of the interviewer observed on the tape-recording--whether appropriate or inappropriate-- is assigned the proper code value. The code values are set down for each specific question; the number of codes given for each question depends on the interaction between interviewer and respondent. In conducting a standardized interview the task of the interviewer consists mainly in reading the text of the question as written, then accurately setting down the response. Ideally, then, only one code value, indicating the correct, word-for-word reading of the question text, is assigned for each question (Code 11). The indication of success in following skip instructions, which is assigned an additional code from the 80s or 90s, is an exception to this. However, if the respondent requests clarification or if his response cannot be assigned precisely to one of the response alternatives, the interviewer must provide additional explanation, which is evaluated with the help of the appropriate code. The more extensive the communication between interviewer and respondent, the more code values become necessary. The behavioral rules defined in this coding system apply primarily to activities that can be heard. Code Group 70 is an exception; it encompasses inappropriate nonverbal behavior. It specifically involves behaviors such as the following:

– The interviewer records a response already received during the interview without asking the question again;

– The interviewer neglects to use a non-directed probe to clarify an inadequate response;

– The interviewer fails to clarify a question's meaning of a question when the respondent has misunderstood it.

We shall provide examples to explain the application and meaning of individual codes in the following section; subsequently we shall demonstrate, using a "live" survey situation (pretest interview for the 1984 welfare survey, Question 33), how inappropriate interviewer behavior and respondent reactions make it necessary for the interviewer to assign numerous code values.

2.2.2.1. Examples of the Meaning of Individual Codes

Code 12 (Appropriate behavior in the category "Asking the question," Code Group 10)

Meaning: Interviewer reads question text with minor alteration, without changing context; no key words are added, omitted or changed.

1. Question text:

Please tell me, using this list, who conducted this course.

Interviewer: Please tell me, using this list /here/, who conducted this course.

2. Text of question:

Into which of the categories listed here would you put this training course?

Interviewer: /The next question is:/ Into which of the categories listed here would you put this training course?

Comments: The interviewer makes minor changes in the text that have no substantive consequences and are thus acceptable.

Code 22 (Inappropriate behavior in the category "Asking the question," Code Group 20)

Meaning: Interviewer substantially alters wording of question; key words are added, omitted or changed.

1. Question text: How much is the net income of everyone in your household, taken together?

Interviewer: /Approximately/ how much is the total income of everyone in your household, taken together?

Comments: The addition of the word "approximately" alters the thrust of the question. Here the interviewer is attempting to relieve an embarrassing situation by allowing the respondent a certain amount of latitude in his response. This frequently occurs in connection with questions regarded as "touchy".

Code 34 (Appropriate behavior in the category "Clarification/non-directed probes," Code Group 30)

Meaning: Interviewer correctly repeats or clarifies respondent's answer in a nondirective manner.

Question text: . . . Do you consider this opportunity, for you personally, to be very good, good, not so good or not good at all?

Interviewer: . . . Do you consider this possibility, for you personally, to be very good, good, not so good or not good at all?

Respondent: Oh, I think it's quite good.

Interviewer: Do you consider it very good or good?

Comments: The interviewer is showing correct behavior by repeating the response alternatives given in the question text in order to render more precise the respondent's answer, calling on the respondent to settle on one precise alternative. Code 11 is given for the correct reading of the question text (= first interviewer activity) and Code 34 for clarifying the response (= second interviewer activity).

Code 44 (Inappropriate behavior in the category "Clarification/non-directed probe," Code Group 40

Meaning: Interviewer inaccurately summarizes the respondent's answer or decides for himself into which of the categories listed the response should be put.

Question text: Thinking of your own situation, how do you regard the possibility of going to school while you are working? Do you consider this possibility, for you personally, to be very good, good, not so good or not good at all?

Interviewer: Thinking of your own situation, how do you regard the possibility of going to school while you are working? Do you consider this possibility, for you personally, to be very good, good, not so good or not good at all?

Respondent: That's something for young people, I think I'm too old for that.

Interviewer: Then you consider it not so good.

Comments: The interviewer interprets the answer given by the respondent and assigns it to a response alternative; this is directive interviewer behavior, which produces manipulated data. Code 11 (first activity = correct reading of the question text) and Code 44 (second activity = interviewer decides into which category to put response).

2.2.2.2. Example of dialogue using a standardized question between the interviewer and the respondent and the assignment of a code value (pretest interview, "live" situation)

Wording of question in questionnaire:

INT.: SHOW LIST P. (List P shows a scale from 0 - 10 with verbalized end points.)

How much pressure do you feel from your job and housework, everything taken together? Please give the degree of pressure you feel on a scale of 0 to 10. "0" means that you feel "no pressure," "10" means that you feel pressure "to the limit of your endurance." The numbers in between are to show varying degrees of pressure.

Interviewer: (first activity)

How much pressure do you feel from your job and housework, everything taken together? Please give the degree of pressure you feel on a scale of 0 to 10.

Code 22: Interviewer substantially alters wording of question, omits key words, i.e. in this case the explanation of the scale.

Code 64: Incorrect technical procedure (disregarding list given).

Respondent:

. . . . (says nothing, interviewer allows too little time for thinking and goes on)

Interviewer: (second activity)

How much pressure do you feel from your job and housework, everything taken together?

Code 47: Interviewer allows respondent insufficient time to think.

Code 42: Interviewer does not repeat question correctly, i.e. as set down in the questionnaire.

Respondent:

Oh, gosh, that's hard to say. I mean I'm sure there are a lot of people who completely . . .

(Interviewer interrupts, both talk at the same time; respondent cannot be understood)

Interviewer: (third activity)

List P, look at list P . . . O, P--where's P?

Code 35: Interviewer correctly explains the technical procedure.

Code 62: Interviewer interrupts respondent.

Respondent:

Yes, but on the other hand it's supposed to be my subjective feeling . . .

(interviewer interrupts)

Interviewer: (fourth activity)

Yes, right.

Code 62: Interviewer interrupts.

Respondent:

. . . and not compared with other people.

Interviewer: (fifth activity)

Yes, exactly, just your subjective feeling.

Code 58: Interviewer makes permissible comment.

Respondent:

Seven. (Respondent answers with point 7 on the list)

Accordingly, the codes given the interviewer for this question are: 22, 64, 47, 42, 35, 62, 62, 58.

### 2.3. Purpose of Study 1

Along with testing a procedure like the interaction coding system for evaluating interviewer behavior, the aim was to analyze the performance of ZUMA's own interviewing staff.

As pointed out above, the general behavioral rules on which the coding system is based allow evaluation to take place according to quite objective criteria. This makes data comparability possible, in this case the comparability of data on interviewer performance. The following can be compared:

- Each interviewer's performance in individual behavioral categories;

- Each interviewer's overall performance in all categories;

- The overall performance of all interviewers in each category and in all categories taken together.

This insight into individual and overall performance makes it possible to do the following:

- Determine the performance level of an entire study staff;
- Provide systematic feedback to the interviewer on his behavior
  - in each interview,
  - in all interviews he conducts,
  - and on the overall behavior of all interviewers in a study, which allows the individual to compare his own performance with that of others (performance motivation);
- Carry out targeted training in all of the behavioral areas included in the coding system.

All members of the ZUMA interviewing staff at the time were included in the evaluation of interviewer behavior. They were all trained interviewers who had taken part in a multi-level program of basic training. Because of its structure and the opportunities it offers, the interaction coding system seemed appropriate for carrying out the necessary quality controls and offering further training for these interviewers.

## 2.4. Description of Field and Training Activities

### 2.4.1. Technical Implementation of the Field Work

The technique was tested in two pretests, carried out at a six-month interval. Twelve interviewers participated in each pretest, each of them conducting and tape recording three quota interviews [Translator's note: "Quoteninterviews", presumably interviews drawn from a quota sample); this resulted in a total of seventy-two interviews. All of the interviewers in the first pretest also participated in the second. In both studies--in line with the coding system--standardized questions were used that shared certain structural characteristics, i.e. they were primarily closed-end questions, with some open-ended follow-up questions. In each pretest the average interview lasted approximately 35-40 minutes.

### 2.4.2. Feedback and Training Phase

After the field work for the first pretest was completed, the tapes were played back and a code value was assigned to each interviewer's behavior for each question. For each interviewer the code values were entered in a table for each interview and each question, then summarized on a separate sheet. This summary provided the interviewer with standardized information on the appropriate and inappropriate behavior he exhibited in his interviews, and gave an overall score evaluating his performance in each behavioral category (cf. Table 1, behavior categories I-IV) and in all categories taken together. Thus each interviewer was able to assess his own performance in each interview and for all interviews together, as well as to compare his performance with that of other interviewers.

The reaction of the interviewers to the use of this technique was very positive. They were pleased to have the opportunity for feedback about their own performance level and to be able to compare it with that of the staff as a whole.

Following this phase of written feedback, targeted training activities were carried out with each individual; one of each interviewer's tapes, generally that with the most errors, was played back and specific instances of correct and incorrect behavior were discussed. The analysis of each individual's interview situations and his "right" and "wrong" behavior produced a more pronounced learning effect and greater interest than had earlier, more theoretical training activities. As our results will show, these targeted training activities had a positive effect on the subsequent pretest.

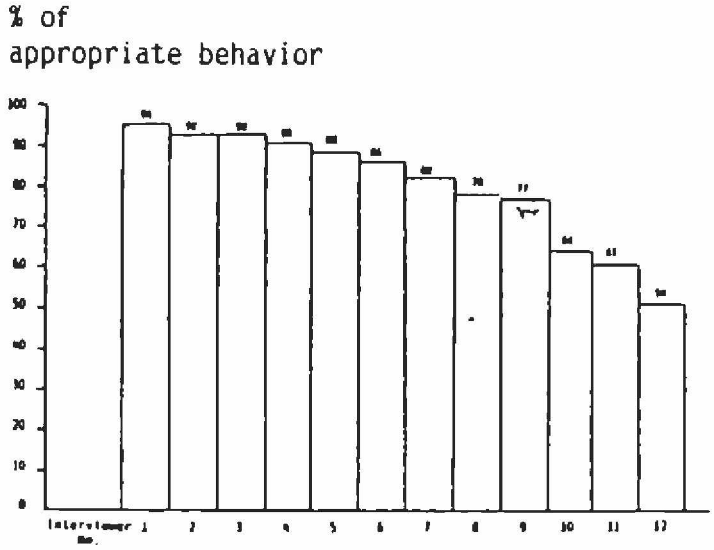
It should be noted that no individual sessions were held during the feedback phase of the second pretest; feedback was provided on each interviewer's performance only in the form of a summary sheet.

## 2.5. Results with Regard to Interviewer Behavior

The evaluation of interviewer behavior carried out in the first study using the technique described above produced an overall percentage of appropriate behavior for each of the twelve interviewers in all of the interviews he conducted. A score of 90 indicates that 90% of all behavior registered for the interviewer in question was considered to be appropriate.

In the first study there was a wide range of "interviewer performance," from 95% for the best to 50% for the worst interviewer (cf. Figure 1).

Figure 1. Individual interviewer performance (first study) (percentage of appropriate behavior)



A comparison of these scores among individual interviewers is useful only when one merely compares the percentages of appropriate behavior. It appears to be problematical to go further and attempt to interpret the performance scores as a measure of the quality of an interviewer's performance, since the assignment of a particular code value does not indicate degrees of the "quality" of a certain behavior. A type of behavior can only be coded as "right" or "wrong", it is impossible to differentiate between these two extremes A serious, consequential error by an interviewer (e.g an error in following skip instructions) is given the same weight--with the number 1-- as a careless error (e.g "interviewer makes superfluous comment") in calculating the performance score. Thus it is entirely conceivable that an interviewer who has a habit of making unnecessary comments might end up with a lower score than another who makes few, but serious, errors. This point should be taken into consideration in a coding system aimed at evaluating interviewer behavior, for example by making it possible to achieve a qualitative differentiation in the evaluation of specific interviewer activities.

In further analyzing interviewer behavior, we focused on clarifying two questions:

1 How are appropriate and inappropriate interviewer activities distributed among the individual categories?

2. Are there typical distributions of error among the various categories for "good" and "bad" interviewers?

Table 2 shows the distribution of appropriate and inappropriate interviewer activities in the individual categories on the basis of 35 interviews during the first pretest; a total of 1415 activities were coded.

Table 2: Distribution of appropriate and inappropriate interviewer activities among the individual categories (first study)

| | Appropriate activities | Inappropriate activities |
|---|---|---|
| Category I "Asking the question" | 52.8% | 3.6% |
| Category II "Clarification/non-directed probes" | 15.6% | 9.4% |
| Category III "Other behavior" | 2.0% | 11.2% |
| Category IV "Skip instructions" | 4.8% | 0.6% |
| Total | 75.2% | 24.8% |
| | 100.0% | |

The table shows, first of all, that some 3/4 (75.2%) of all interviewer activities are classified as appropriate, and 1/4 (24.8%) as inappropriate. It can also be concluded from the table that inappropriate behavior is concentrated in Categories II "Clarification/non-directed probes" and III "Other behavior." This is particularly striking in Category III, where far more instances of error than of correct behavior were recorded. This is primarily because the coding scheme in this area contains more cases of inappropriate than appropriate behavior (11:2). Thus a kind of behavior that is assigned to Category III is quite likely from the outset to be considered inappropriate.

Table 3: Comparison of error structure between the groups of the three best and the three worst interviewers, based on the average percentage of error per category (first pretest)

| | BEST GROUP | | WORST GROUP | |
|---|---|---|---|---|
| | Appropriate behavior | Inappropriate behavior | Appropriate behavior | Inappropriate behavior |
| Category I "Asking the question" | 100.0% | 0.0% | 81.7% | 18.3% |
| Category II "Clarification/ non-directed probes" | 90.7% | 9.3% | 43.0% | 57.0% |
| Category III "Other behavior" | 13.0% | 87.0% | 2.3% | 97.7% |
| Category IV "Skip instructions" | 100.0% | 0.0% | 90.7% | 9.3% |

When one compares the distribution of inappropriate behavior between the group of the three best and the group of the three worst interviewers (cf. Table 3), group-specific error distributions are very much in evidence. The "best" interviewer group is marked by an absolute lack of error in Categories I and IV, a relatively low error rate in Category II and a high error rate in Category III--which, however, as pointed out above, is to a large degree inherent in the system itself. While the "good" interviewer shows no error in the categories where "scripted rules" (such as reading the question) are to be followed (Categories I and IV), the interviewers in the "worst" group exhibit inappropriate behavior even here. The clearest difference between the two groups, however, is in Category II "Clarification/non-directed probes," where only 9.3% of the behavior in the "best" group was inappropriate, in contrast to more than half, or 57%, in the "worst" group. These results make it clear that subsequent training to correct problems that have emerged should be aimed particularly toward improving behavior in the Category "Clarification/non-directed probes."

The results of the second pretest were intended to demonstrate whether and to what extent targeted training indeed led to an improvement in the performance of the interviewers involved.

Figure 2: Individual interviewer performance (second pretest) (percentage of appropriate behavior)

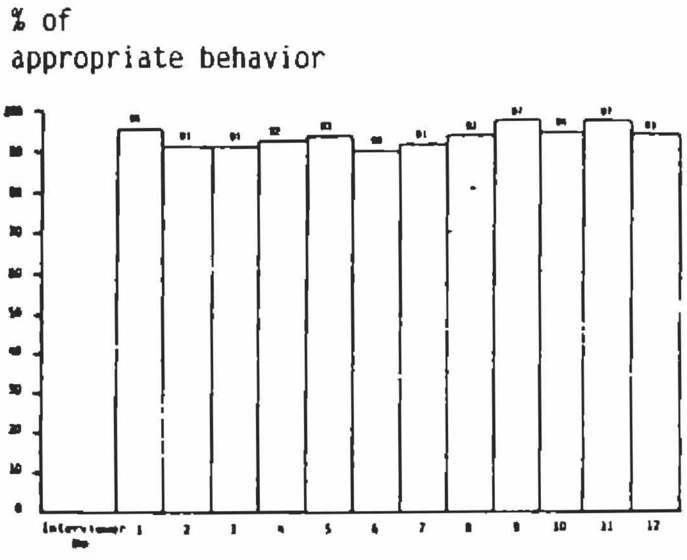


Figure 2 shows that the variation in individual interviewer performance was greatly reduced, which indicates a clear improvement in performance, particularly by bad interviewers. However, this can only to a limited degree be interpreted as a genuine improvement in quality, since an analysis of the tapes of the second pretest showed that the interviewers, aware that they were being evaluated, tended to avoid error, i.e. they became significantly more cautious. Apparently nearly all interviewers attempted to avoid situations calling for behavior that might be in danger of being classified as inappropriate. The knowledge that their behavior was being evaluated led them to be more "formal" than was the case in the first pretest.

Similarly, a direct comparison of the results of the first and second pretests is constrained by the fact that a questionnaire was used in the second pretest that differed in content from that of the first pretest, although their structures were similar, and that the target individuals recruited for the second study were not the same as those in the first, so that they might react differently, a factor that could affect interviewer behavior especially where clarification was required (Category II). This also applies to the distribution of appropriate and inappropriate interviewer activities in specific categories (cf. Table 4).

Table 4. Distribution of appropriate and inappropriate interviewer activities among the individual categories (second pretest)

| | Appropriate activities | Inappropriate activities |
|---|---|---|
| Category I<br>"Asking the question" | 71.4% | 2.8% |
| Category II<br>"Clarification/<br>non-directed probes" | 12.2% | 1.1% |
| Category III<br>"Other behavior" | 2.3% | 2.8% |
| Category IV<br>"Skip instructions" | 6.8% | 0.6% |
| Total | 92.7% | 7.3% |

A comparison with the results of the first pretest (cf. Table 2) shows that the error rate decreased noticeably particularly in Categories II and III. As pointed out above, one should be cautious in interpreting these results as quality improvement, since precisely in these categories the interviewers are able to limit certain activities to avoid potential error (e.g. a minimum of clarification, omission of additional comments, etc.), and this indeed occurred in the second pretest.

There was a genuine improvement in quality in Categories I and IV ("Asking the question" and "Skip instructions"), both of which categories are independent of both the questionnaire and the respondent, and both of which offer the interviewer little opportunity to exert influence because of their strict rules.

Changes in behavior as compared with the first pretest are especially striking in the results of the three "worst" interviewers (Numbers 10, 11 and 12 in Figures 1 and 2), as reflected in the individual categories (cf. Table 5).

Table 5: Comparison of error structure between the groups of the three best and the three worst interviewers, based on average percentage of error per category (second pretest)

| | BEST GROUP | | WORST GROUP | |
|---|---|---|---|---|
| | Appropriate behavior | Inappropriate behavior | Appropriate behavior | Inappropriate behavior |
| Category I "Asking the question" | 95.7% | 4.3% | 94.7% | 5.3% |
| Category II "Clarification/ non-directed probes" | 90.3% | 9.7% | 90.0% | 10.0% |
| Category III "Other behavior" | 18.0% | 82.0% | 75.3% | 24.7% |
| Category IV "Skip instructions" | 90.7% | 9.3% | 100.0% | 0.0% |

A comparison with the corresponding results of the first pretest shows drastic reductions in every category in the error percentages of the "worst" group. Here, too, for the above-mentioned reasons, the reduction in error in Categories II and III should not be interpreted simply as an improvement in the quality of interviewer performance; for Categories I and IV, however, one can assume that the reduction in error frequency does indeed stem from positive effects of training.

A comparison of the two groups illustrates the consequences of the overcautious error-avoidance behavior shown by the "bad" interviewers. In Category III "Other behavior" this group actually achieved error rates considerably lower than those of the "best" group. If we compare these results of the "best" group with their performance in the first pretest, we see that the error percentages in Categories II and III remained relatively constant and even increased in Categories I and IV. This might be interpreted as a "laurel effect": Recognition that they had performed well in the first pretest led to carelessness and lack of concentration.

In addition to permitting a limited interpretation of the purely quantitative interviewer performance results and shedding light on sources of error, a comparison of the first and second pretests shows certain changes in behavior resulting "purely" from training activities. As mentioned above, these changes are found in connection with those interviewer activities that are independent of both the respondent and the questionnaire. This applies to a great extent to the category "Asking the question," which involves reading the question text. A comparison of the quantitative results (cf. Tables 2 and 4) shows no marked change in the share of appropriate versus inappropriate behavior in this category; however, an examination of individual code values in this category indicated a qualitative improvement as compared with the first pretest in that there was a noticeable decline in the interviewers' tendency to make slight changes in reading the text (Code value 12, considered appropriate behavior). Instead there was an increased tendency to read the test of the question completely correctly (Code value 11).

Within the category "Other behavior" as well, and especially in Code Group 60, which basically includes activities that are independent of the respondent and the questionnaire, positive behavior changes were noted in the second pretest: the interviewers avoided inappropriate activities that were observed relatively frequently in the first pretest (e.g. Code 63 "Interviewer gives personal opinion or evaluation" or Code 66 "Interviewer unnecessarily rephrases respondent's answer"). This learning effect occurred particularly among the weaker interviewers.

In general terms, there are indications that the objective of reducing habitual, superfluous activities can be achieved in a relatively quick and problem-free manner by the interviewer--more quickly, for example, than he can learn appropriate, flexible activities in response to particular behavior on the part of the respondent.

Furthermore, there are indications that it is easier to learn and apply rules or instructions that the interviewer can take from the questionnaire ("scripted rules"), such as the text of the question, response alternatives, interviewer instructions and skip instructions, than it is to perform tasks that the interviewer is supposed to have learned in training and must apply flexibly to suit a particular interview situation. Increased training efforts are required in this area.

## 2.6. Assessment of the Interaction Coding Technique for the Evaluation of Interviewer Behavior

The interaction coding technique is indeed appropriate for the use described here, which is the analysis of interviewer behavior. It is clear that a study can be used to evaluate the overall performance of each interviewer with considerable objectivity, and to assess interviewer performance in individual areas of behavior. Thus this procedure offers the advantage of enabling one to target training to specific areas and individuals on the basis of quite objective quality controls. In addition, individual evaluations help to determine the overall performance of the entire interviewing staff.

Even bearing in mind the special circumstances mentioned above under which the interviews of the second pretest were conducted, we can conclude that the weaknesses in interviewer performance revealed by the interaction coding technique can be remedied; performance can be improved. This leads to a reduction in unintended interviewer influence in the survey, an aspect that is important in rendering the data collection process as nearly optimal as possible.

## 3. Study 2: Evaluation of Question Quality

## 3.3. General Remarks on the Evaluation System

The attempt to use the interaction coding technique to draw conclusions about question quality is based on the system's procedural methods themselves. As we have noted, each verbal activity, i.e. each kind of behavior described in the coding system, is assigned a code value. Thus the more activities that occur for each question, the more code values are assigned. It can be assumed that this applies particularly to unclearly operationalized question texts, which induce the respondent to request clarification and the interviewer to offer additional explanation, more than it does to questions that provide a clear stimulus, which can be answered without additional activity, i.e. without requiring questions or clarification. Thus the number of code values, depending as it does on the length of interaction between respondent and interviewer, can be an indication of how well a question functions (cf. also the works of Morton-Williams, 1983, and Cannell, 1971).

The usual method of testing a question's ability to function is to conduct pretests. Here the interviewer plays an important role in that it is he who submits a report on his observations concerning individual questions and how well they function. These interviewer reports are influenced by the subjective perceptions of the individual interviewer and are usually limited to what the interviewer experiences as severe problems occurring in his interviews. This method cannot produce an objective and realistic impression of the interview situation. Since the interaction coding technique provides a means of systematically recording the entire course of the interview, it seemed logical to test the procedure and its utility for determining question quality.

## 3.2. Brief Description and Illustrations of the Technique

A new system of code values was developed by the authors to apply the interaction coding technique to the evaluation of question quality, a system intended to be used for the standardized oral interview. The application of this system is based on the same basic principles as those described in detail under Study 1, but it differs from that system in two major respects:

1. In addition to the catalogue of all interviewer behaviors, it contains a list of all possible kinds of behavior that the respondent might exhibit in the interview.

2. Its formal structure makes no distinction between appropriate and inappropriate behavior.

Regarding No. 1: In the Study 1 system of code values for evaluating interviewer behavior it became apparent that one weakness was that respondents' reactions were not recorded, since interviewer behavior is largely dependent on respondent behavior and in many cases can only be explained by taking it into account. It is essential to record respondent activity if one is to assess question quality by observing the amount of interaction between interviewer and respondent.

Regarding No. 2: The aim in using the interaction coding technique to evaluate question quality is to determine such quality by observing the formal characteristics of behavior or interaction, not by evaluating these characteristics. For this reason the coding scheme constructed for the present study does not distinguish between appropriate and inappropriate behavior. If such a system were to be based on an evaluation of behavior, this would add a subjective dimension to the evaluation of question quality.

The following remarks on the coding system do not deal with its basic structure and application, which have already been described thoroughly in our remarks on Study 1. Here we shall present only the formal organization of this system. The coding system consists of as complete as possible a list of behaviors that the interviewer and the respondent might show during an interview, such as: reading the question text, questions by the respondent about the meaning of the question, clarification by the interviewer, answer by the respondent.

- Complex 1 includes all descriptions of behavior that constitute a stimulus by the interviewer, i.e. reading the question or response alternatives, scales or items, that take place before the first response from the respondent. The code values for these behaviors are in Code Group 100.

- Complex 2 describes all behavior by the respondent. The code values are contained in Code Group 200.

- Complex 3 describes all interviewer behavior beyond the reading of the question, such as explanations or non-directed probes, behavior that takes place after the first response from the respondent. These kinds of behavior are represented by Code Group 300.

- Complex 4 concerns interviewer recording of responses on the questionnaire, and is represented by Code Group 400.

A detailed description of the coding scheme is to be found in Appendix 2.

### 3.2.1. Assignment of Codes

Every kind of behavior recorded on the tape--by the interviewer as well as the respondent--is assigned the appropriate code value. These code values are set down question for question, and the number of code values assigned for each question depends on the length of the interaction process between the interviewer and the respondent. Here are two examples of possible interaction between the interviewer and the respondent to illustrate the application of the coding system:

Example 1 shows an ideal question and answer sequence, i.e. the interviewer reads the question and the possible responses as written and the respondent gives an answer in accordance with the response alternatives without asking for clarification or making other comments:

Question text in questionnaire

Let's begin with a few questions on the economic situation: How would you generally describe the current economic situation in the Federal Republic:

INTERVIEWER: READ POSSIBLE ANSWERS.

- very good

- good

- partly good/partly bad

- bad, or

- very bad?

Interviewer

Code 111: reads question as written (as described above).

Code 121: reads possible answers as written (as described above).

Respondent

Partly good, partly bad.

Code 201: Respondent gives appropriate answer to a closed-ended question.

The codes are recorded in the order of their occurrence: 111, 121, 201.

Example 2 shows another possible case using the same question:

Question text in questionnaire

Let's begin with a few questions on the economic situation: How would you generally describe the current economic situation in the Federal Republic:

INTERVIEWER: READ POSSIBLE ANSWERS.

- very good

- good

- partly good/partly bad

- bad, or

- very bad?

Interviewer

Let's begin with a few questions on the economic situation: How would you generally describe the current economic situation in the FRG:

- very good

- good

- partly good/partly bad

- bad, or

- very bad?

Code 112: Interviewer reads question with minor alterations. Code 121: Interviewer gives response alternatives as written.

Respondent

Pretty lousy, I'd say.

Code 222: Respondent gives inadequate answer (not in accordance response alternatives given for the closed-ended question).

Interviewer

Please give one of the alternatives listed. I'll read them again: very good, good, partly good/partly bad, bad or very bad?

Code 303: Interviewer correctly repeats question or parts of question.

Respondent

Bad.

Code 201: Respondent gives appropriate answer to closed-ended question.

The following code values are assigned: 112, 121, 222, 303, 201.

## 3.3. Field Work and Questionnaire

### 3.3.1. Field Work

Before presenting the results of the study, we shall turn briefly to the field work and the instrument on which the study was based.

Sixty "address interviews" [Translator's note: "Adresseninterviews", presumably interviews conducted at a particular address] were conducted by the twelve ZUMA interviewers. The addresses were determined by the interviewers using a random method; the target person within the household was determined by means of an additional procedure. The distribution of demographic characteristics deviates from that of traditional population surveys. Since the present study is something of a pilot project, it does not claim to be representative in the composition of its sample. Care was taken only to ensure that differences in age, sex and education were adequately represented.

### 3.3.2. The Questionnaire

The questionnaire consisted of questions taken from the 1984 General Population Survey of the Social Sciences (ALLBUS). Since previous experience had shown that coding work is very time-consuming, it was necessary to limit the number of questions to be analyzed with the help of the coding system. Care was taken to include types of questions with varying degrees of complexity (questions with alternatives, questions with verbal scales, with numerical scales, requiring the respondent to rank items, in which cards or lists are presented, etc.). It should also be noted that in questions with batteries of items the individual items were treated as independent questions for coding purposes.

## 3.3. Results

### 3.4.1. Definition of Quality

Before examining the extent to which this technique enables one to draw conclusions about question quality, an attempt should be made to define the concept of "quality". An important indication of question quality is that the meaning intended by the researcher corresponds to that perceived by the respondent. This is why many researchers strive to word questions without any semantic unambiguity. It also explains efforts by researchers such as Belson (1982) to find out from the respondent after the interview how he perceived a question's meaning.

The problem, however, is that time limitations render it impossible to test question understanding in surveys that use large samples. Similarly, financial considerations usually prohibit a subsequent survey of the meaning of a stimulus. Since one can assume, first of all, that there are large variations among individuals in their understanding of question stimuli, while secondly there are only very inadequate means of testing this factor, one should attempt to avoid or maintain control of all factors in the interviewing process that might add to the variability of question comprehension.

As we have pointed out, it is first of all the interviewer who may intervene in the response process, which can distort the question content as intended by the researcher. Since intervention by the interviewer usually occurs when the respondent requests clarification or gives an inadequate response, the very number of cases in which the interviewer intervenes is one indicator of the quality of question wording, and thus of an important aspect of quality in general.

Of course, there are other ways of operationalizing in order to render more precise the concept of question quality. One logical possibility would be to observe to what extent a question leads to appropriate responses, i.e. to responses that correspond to the intention of the question or to the alternatives provided, without giving any indication of invalidity Question A, for example, might be regarded as qualitatively superior to question B if A leads to a larger number of "adequate" responses than B. Unfortunately one cannot completely rely this sort of operationalization, since intervention by the interviewer does not tell us whether the respondent would have given an appropriate response had there been no interviewer action. This uncertainty points back to the need to take into consideration the interaction between interviewer and respondent, which can be accomplished by the coding system on the basis of its structural composition.

### 3.4.2. Results for Quality Determination

Seventeen questions were selected for evaluation using the coding system, including some containing long batteries of items. The individual items were treated in the coding process as independent questions, so that the computation is based not on N=17, but on N=57 (questions). Since the project described here is something of a pilot study, i.e. it was intended to test the use of the coding system in this area, the results shown here are based on simple methods of analysis. The first analytical step was carried out to show whether observing how many codes were assigned to a particular question would provide a basis for conclusions about the extent to which that question functioned properly. Thus the basic analyses that were performed deal with elementary forms of interaction. More differentiated forms of interaction, which might be studied by observing certain patterns of code values, will have to be looked at in future analyses.

Table 6 shows the distribution of kinds of interaction between interviewer and respondent for all questions and all interviews conducted, given as percentages.

Table 6: Frequency of kinds of interaction for all questions and all interviews, given as percentages (N=3346)

I. Ideal cases: 29.9%

II. Cases with appropriate responses, interview did not proceed ideally: 60.5%

II. Cases not resulting in an appropriate response: 9.6%

The table distinguishes between cases

- that proceed ideally, i.e.: The interviewer reads the question as written and the respondent answers appropriately, with no additional activity taking place between the presentation of the stimulus and the response;

- that produce an appropriate response, while not proceeding ideally, for example The interviewer reads the question text as written, the respondent asks for clarification of some point, the interviewer correctly provides clarification, the respondent answers appropriately, i.e. in accordance with the response alternatives or the intention underlying the question;

- not resulting in an appropriate response, for example: The interviewer reads the question as written, the respondent cannot give an answer in accordance with the alternatives (difficulties in understanding that cannot be solved by the interviewer, "I don't know," "no response").

The distribution of types of interaction as presented in Table 6 is based on the "narrowest" definition of what constitutes "ideal" interaction. Ideal interaction in this sense means that the interviewer gives the question stimulus correctly, word for word (cf coding scheme in Appendix 2, code value 111). Then there is ideal interaction in a broader sense, in which the interviewer is allowed to deviate slightly from the text when reading the question, response categories or items. This would include the addition of words such as "and" or "or", or minor changes in words that do not alter the content of the question or its context (cf. coding scheme in Appendix 2, code value 112).

If one includes such cases in the definition of "ideal", the distribution presented in Table 6 shows the following shift (cf. Table 7):

Table 7: Frequency of kinds of interaction (with the ideal defined more broadly) for all questions and all interviews, given in percentages (N=3346)

I. Ideal cases: 49.9%

II Cases with appropriate responses, interview did not proceed ideally: 40.5%

III. Cases not resulting in an appropriate response: 9.6%

It appeared proper to include these cases as "ideal", since they differ only unsubstantially from the strictly-defined ideal cases, and--as can be seen in Table 7, Column II--they reduce by 20% the fraction of cases that did not proceed ideally. Thus one can assume that the cases remaining in Column II differ more radically from the "ideal" cases. The cases in Group III will be dealt with in detail below.

These two tables demonstrate how interaction may proceed for all questions. This can, of course, be carried out for each individual question as well; it is possible to determine for each question the share of cases that proceeded ideally, that produced appropriate responses but did not proceed ideally, and that failed to produce an appropriate response. These results offer a somewhat rough indication of the "functioning" or quality of a question. They show, in addition to the basic ability of the coding system to differentiate between case groups, the fact that additional activity beyond ideal behavior does indeed occur, a finding which, although it does not come as a particular surprise, can only be quantified with the help of a suitable coding system.

Table 8 demonstrates how additional activity, i.e. going beyond ideal interaction, is distributed among the individual questions. For this purpose the average number of additional activities was computed for each question and the results were divided into four groups. If interaction on one question proceeded ideally, 0 activities were recorded.

Table 8: Average number of additional activities per interview for all questions

| Frequency intervals | 0.1-0.9 | 1.0-1.9 | 2.0-2.9 | 3.0 and more |
|---|---|---|---|---|
| Number of questions (N=57) | 35 | 15 | 4 | 3 |
| Percentage | 61.4% | 26.3% | 7.0% | 5.3% |

The table shows that for 61.4% of the questions (N=57) the average number of additional activities beyond the ideal lies between 0.1 and 0.9, that is, in these cases, on average, the interview deviates only slightly from the ideal. It is striking that 38.6% of all questions show an average number of additional activities of more than 1.0, and 12.3% show more than 2.0. The averages seem to be relatively small. However, in each group there are cases that deviate sharply from the norm; even in the group with the lowest number of additional activities, the group from 0.1 to 0.9, there is one case that shows sixteen activities, and in the last group there are cases of up to thirty additional activities.

This section has so far been primarily concerned with activities that take place between the presentation of a stimulus and an appropriate reaction on the part of the respondent. Below we shall present different ways in which interaction can proceed where some individual questions are concerned. Here of particular interest are the cases defined in Tables 6 and 7 which show inappropriate response behavior. In the present study the percentage made up by this group (9.6%, cf. Table 7) turns out to be substantially higher than the missing value percentage found in most surveys. This is primarily because this group proved far from homogeneous when evaluated with the help of the coding system, and it was marked by a number of interesting details. These will be presented below in connection with interaction on specific questions (cf. Figure 3).

Figure 3: Excerpt from Question 9, items A, E and G

INTERVIEWER: SHUFFLE GRAY SHOW CARDS AND GIVE THEM TO RESPONDENT

And now on to another area: Here I have some opinions on the state and the economy in the Federal Republic. Please tell me whether you agree completely, agree to some extent, don't really agree or don't agree at all with each of these opinions.

INTERVIEWER: CIRCLE THE NUMBER OF ONE ANSWER FOR EACH OPINION

Agree completely 1

Agree to some extent 2

Don't really agree 3

Don't agree at all 4

Don't know 5

A In our society everyone must see to it that he gets somewhere on his own It doesn't help very much to join together with other people in political groups or unions to fight for your interests.

E When the benefits provided by the social security system, such as continued wages in the case of illness, unemployment and early retirement benefits, are as high as they are today, people don't want to work anymore.

G. By and large economic profits are distributed fairly in the Federal Republic today

Explanation of the technique of this question: The individual cards contain the text of each item (visual presentation); the scale of one to four is given to the respondent in the question text (oral presentation).

Table 9 Percentages of courses of interaction for three items in question 9

| Question | 9A | 9E | 9G |
|---|---|---|---|
| I Proceeded ideally | 15.0% | 20.0% | 16.7% |
| II Resulted in appropriate response, did not proceed ideally | 46.7% | 46.7% | 56.6% |
| III Did not result in appropriate response | | | |
| - Missings (KA = no mark in questionnaire, wn = don't know, vw = refused to answer) | 3.4% | 3.3% | -- |
| - Inadequate answer, entered adequately on questionnaire | 34.9% | 30.0% | 26.7% |
| - Total | 38.3% | 33.3% | 26.7% |

Looking at Group III, Table 9, which is made up of those cases that fail to result in an adequate response, we see that it consists of two subgroups:

1 The group of "missing values" with the variations "no information" (i.e. no mark on questionnaire), "don't know" and "refused to answer";

2. Cases in which an adequate response has been recorded on the questionnaire, which turns out in coding to be inadequate.

Regarding No. 1: This group includes cases in which no answer was given (recorded as code value 225 "Respondent unable to answer question, lack of information") or the respondent refused to answer (code value 226). In addition, those cases are of interest in which nothing was marked on the questionnaire or no information (=KA) is available. Using the coding system it was possible to deduce how the interview proceeded and to investigate the reasons why the interviewer failed to record any response.

It turns out that the interaction between interviewer and respondent on these questions consists of fairly lengthy interaction "chains" which result because the interviewer attempts to encourage the respondent to give an adequate response. If, however, his efforts prove fruitless, for example because the respondent insists that he can only answer with certain reservations or with additional conditions not contained in the fixed question stimulus, or because he refuses to make a clear choice, the interviewer is confronted with the dilemma of whether to mark nothing at all or to mark incorrectly the alternative--if available-- of "don't know" or "refused to answer." To have the interviewer provide handwritten explanations on the questionnaire would, to mention just two problems, take too much of the interviewer's time and cost too much money. Moreover, it would also interfere with the interview, and would not be feasible for questionnaires whose results are computed by machine. Thus experience in using the coding system has shown that in most cases the "no information" category does not indicate an actual mistake on the part of the interviewer. Instead, it is usually the "end result" of a long process of interaction that has not led to an adequate response despite the efforts of the interviewer.

The causes of such undesirable behavior, however, lie not only in flaws in question wording and incorrect behavior by the interviewer, but also in standardized questions that do not allow for "individual" reactions by the respondent, instead forcing him to conform to a fixed response.

Regarding No. 2: The second subgroup is made up of cases in which a response was entered on the questionnaire, but proved inadequate for the coding procedure. Such cases may involve inadequate responses in a narrow sense (code value 222 "Respondent gives inadequate answer to a closed-ended question") or may be invalid responses resulting from inconsistency (for example when someone modifies his response by giving additional comments = code value 205).

As a rule, these cases are not recognized when the results of the interview are examined. Such differentiation can only be accomplished with the help of a coding system like the one under discussion. The fact that such cases occur with great frequency, however, is shown by Table 9, with 34.9% for Item A, 30.0% for Item E and 26.7% for Item G. It should also be noted that the other items in Question 9 as well show a large proportion of similar inadequate responses. The reason for this relates to the questions themselves and can be explained by two factors: As a result of this question's lengthy battery of items (from A – H) including items with difficult content, as well as its nonvisual verbal scale with four choices, the respondent frequently failed to use the differentiated wording of the scale given, since he was unable to remember the various points, but answered in his own words or simply "agreed". Because of the long, drawn-out nature of this question, resulting from the large number of items, the interviewer, after initially proceeding according to instructions, began to neglect the required question to determine the exact response as given in the scale and instead decided for himself into which category the inadequate response should be put, thus resulting a marking in the questionnaire that appeared to be correct.

If we examine the share of cases that proceeded ideally for Question 9, items A, E and G (Column 1) as well as the proportion of cases resulting in an adequate response, while not proceeding ideally, there clearly seem to be very similar processes of interaction at work here: a small proportion of cases proceeding ideally and relatively large proportions both of cases with an adequate answer, but not proceeding ideally, and of cases not resulting in an adequate answer.

These results indicate that problems in responding to this question are reflected in the large percentage of adequate responses where the interaction did not proceed in an ideal manner. Furthermore, it can be assumed that similar problem structures exist for all three items of this question.

The demography question concerning vocational training, Figure 4, is a similar case.

Figure 4: Demography Question 54

INTERVIEWER: SHOW LIST 52. MULTIPLE RESPONSES PERMISSIBLE

Let's talk about your vocational training: Which of the following apply to your situation?

A. Still in training.

B. No completed training

C. Vocational/in-company training with certificate of completion, but no apprenticeship

D. Apprenticeship with completion of final examination

E. Vocational practicum, period as trainee

F. Completed vocational school

G. Completed advanced vocational school

H. Completion of requirements for certificate as master [Translator's note: e.g. master craftsman], technician's license or similar qualification

I. Completed technical college (including engineering school)

K. University degree

L. Other vocational qualifications, specifically:____________

As Table 10 shows, there were no cases in which the interaction for this question proceeded ideally. There was, however, a large percentage (91.7%) in which an adequate response resulted, although the interview did not proceed ideally, i.e. additional activity was registered or, in 8.3% of the cases, adequate responses were recorded in the questionnaire that in coding turned out to be invalid. These results indicate problems in responding to this question. Superficially, one might argue that since adequate responses were given in 91.7% of all cases, the goal has been achieved to an acceptable degree. If, however, one assumes--as mentioned above--that intervention in the questioning process by the interviewer may constitute a source of wrong behavior and indicate that the respondent has requested clarification, the fact that 91.7% of all cases result in an adequate response, but do not proceed ideally, points to flaws in this question.

Table 10: Percentages of kinds of interaction for Demography Question 54

Question: 54

I. Proceeded ideally: 0.0%

II. Resulted in adequate response, but did not proceed ideally: 91.7%

III. Did not result in adequate response

- Missings (KA = no mark on questionnaire, wn = don't know, vw = refused to answer): 0.0%

- Inadequate response, marked as adequate in questionnaire: 8.3%

- Total: 8.3%

A "normal" analysis of results would not indicate this type of problem, since in all cases adequate responses were recorded on the questionnaire. It is only when we differentiate, using the coding system, among different ways in which interaction proceeds with additional activity, and distinguish between subgroups of cases resulting in inadequate responses, that we can recognize the difficulties involved in an operational definition of this question. The causes lie in a lack of sufficient distinction between response alternatives. The respondent cannot immediately find the right category for his situation, which causes dialogue to occur between the interviewer and the respondent; the frequency with which this happens is reflected in the large percentage in Group II. In 8.3% of all cases the interviewer himself attempted to assign what were frequently open-ended responses to a particular category, resulting in inaccurate markings.

These examples show that the coding system, by differentiating among kinds of interaction and inadequate response behavior, is capable of providing information about the extent to which a question functions properly, information that cannot be gained by means of usual pretest methods.

We shall briefly address two additional results which, while not directly related, are nonetheless of interest

In order to explain the additional activities that take place over and above what would be the ideal interview, it makes sense to examine the extent to which these activities stem from certain respondent and interviewer characteristics. It is particularly interesting to determine the extent to which various degrees of additional activity are due to different interviewers, since an answer to this question could be helpful for selecting and training interviewers.

Table 11: Respondent characteristics and average number of questions proceeding ideally per interview

|  | Sex |  | Age |  |  |
|---|---|---|---|---|---|
| Respondent characteristics | male (N=38) | female (N=22) | up to 44 (N=30) | 45-59 (N=20) | 60 and up (N=10) |
| Number of ideal cases | 27.7 | 28.1 | 29.6 | 30.4 | 17.4 |

Only sex and age were taken into consideration as respondent characteristics. Table 11 contains the results, showing for each subgroup and all questions the number of cases in which the interaction proceeded ideally. In view of the small number of cases, non-parametric tests were chosen to carry out the statistical testing of the differences that emerged. Sex-related differences were tested using the Mann-Whitney U-test and the Merian test; the differences proved to be insignificant. At least in the case of this sample, then, there was no evidence of sex-related differences in the number of questions for which the interview proceeded ideally.

The situation is a different one as far as different age groups are concerned. Here the Kruskal-Wallis test showed a 1% level of significance. This significance was a result primarily of the differences between individuals over sixty years of age and the other groups. It appears likely, then, that more additional activities occur in interviews with older respondents.

Table 12: Average number of questions proceeding ideally per interviewer and interview

| Interviewer No. | Average No. of ideal cases per interviewer | No. of interviews | Standard deviation |
|---|---|---|---|
| 1 | 36.3 | 3 | 11.6 |
| 2 | 34.3 | 4 | 4.2 |
| 3 | 31.7 | 3 | 8.3 |
| 4 | 33.0 | 5 | 4.5 |
| 5 | 27.4 | 5 | 10.3 |
| 6 | 22.0 | 5 | 15.5 |
| 7 | 22.0 | 5 | 12.1 |
| 8 | 25.2 | 13 | 10.1 |
| 9 | 27.7 | 6 | 4.3 |
| 10 | 27.0 | 2 | 11.3 |
| 11 | 29.6 | 5 | 7.2 |
| 12 | 35.3 | 4 | 18.8 |

Table 12 shows the results for individual interviewers; the interviewer numbers are not the same as those in Study 1. Each line gives the average number of cases proceeding ideally for each interviewer, as well as the standard deviation in these ideal cases. The table shows, first of all, that there are indeed differences between interviewers. Some interviewers have apparently produced, on the average, more ideal cases than others. The standard deviations show that these differences involve not only mean values, but also dispersions. It is apparent that some interviewers more consistently bring about ideal cases of interaction than others. If one looks only at those interviewers who have conducted approximately the same number of interviews, between four and six, the differences in dispersion range from 18.8 to 4.2. These results suggest that it would be a good idea to use a coding system like the one we are discussing here to carry out combined evaluations of interviewers according to average number of ideal cases and dispersion. The amount of dispersion appears to be at least as important as the average, since the aim is to have interviewers who work "correctly" throughout.

### 3.5. Evaluation of the System as a Means of Determining Question Quality

In answer to the question posed at the outset, the results presented here point to the following conclusion: If we define the concept of quality in an operational sense, through interaction and adequate answers, this system indeed proves able to determine characteristics of quality that can tell us something about the "functioning" of a question.

First of all, the system can differentiate between different ways in which interaction proceeds, it can show to what extent additional activities occur, and it can point out interesting aspects of inadequate response behavior. These diagnoses of quality are determined systematically, since in order to evaluate a particular question the codes or types of behavior for interviewer and respondent in all interviews are taken into consideration, which is an advantage not offered by the traditional pretesting method. In the traditional method the interviewer reports in a rather unsystematic, scattered way on the problem areas in his interviews that he feels to be relevant.

However, a comparison of the two methods shows that concrete flaws in a question cannot be pinpointed by the coding system, since the actual causes for the "non-functioning" of a question cannot be deduced from the numerical results, i.e. the code values. For such purposes this technique would be too imprecise, unless one could develop a coding scheme that would cover not only general behavioral descriptions but also the characteristics of specific questions, thus making it possible to take into account such detailed additional information. The structure of such a coding scheme, and the code values it would contain, would have to be focused largely on substantive and technical problems specifically related to the individual questionnaire.

In sum, experience with this coding scheme shows that the interaction coding technique, with its objective and systematic procedure and detailed examination of sequences of interaction, can substantially aid in the evaluation of question quality, a contribution which, together with the more qualitative results of conventional approaches, can lead to well-founded pretest conclusions as to the functionality of an instrument.

This report on the interaction coding technique was written by Peter Pruefer and Margrit Rexroth.

## Bibliography

BELSON, W.A. The design and understanding of survey questions. London: Gower, 1981.

BRENNER, M. Patterns of social structure in research interview. In: M. BRENNER (Hrsg), Social method and social life. London, 1981, 115-158.

BRENNER, M. Response-effects of 'role-restricted' characteristics of the interviewer. In: W. DIJKSTRA & J. von der ZOUVEN (Hrsg.), Response behaviour in the survey-interview. London: Academic Press, 1982, 131-165.

CANNELL, Ch. & KAHN, R. Interviewing. In: G. LINDSEY & E. ARONSON (Hrsg.), Handbook of social psychology. Cambridge: Addison-Wesley, rev. ed. 1968, 526-595.

CANNELL, Ch., LAWSON, S. & HAUSSER, D. A technique for evaluating interviewer performance. Survey Research Center, Institute for Social Research, University of Michigan, 1977.

CANNELL, Ch., OKSENBERG, L., CONVERSE, M. Experiments in interviewing techniques. Field experiments in health reporting, 1971-1977. Survey Research Center, Institute for Social Research, University of Michigan, 1979.

CANNELL, Ch., MILLER, P., OKSENBERG, L. Research of interviewing techniques. In: S. LEINHARDT (Hrsg.), Sociological Methodology. San Francisco: Jossey-Bass, 1981, 389-437.

CANNELL, Ch., KALTON, G.W., FOWLER, F.W. Techniques for diagnosing cognitive and affective problems in survey questions. Survey Research Center, Institute for Social Research, University of Michigan, 1985.

GORDON, R. Interviewing: strategy, techniques and tactics. Homewood. The Dorsey Press, rev. ed. 1975.

HYMAN, H. et al. Interviewing in social research. Chicago: University of Chicago Press, 1975.

MATHIOWETZ, N., CANNELL, Ch. Coding interviewer behavior as a method of evaluating performance. Proceedings of the American Statistical Association, Survey Methods Section, 1980, 525-528.

MORTON-WILLIAMS, J. A study of question failure through the use of interaction coding. Invited paper, 44th session of the International Statistical Institute. Madrid, 1983.

PAYNE, S.L. The art of asking questions. Princeton, N.J.: Princeton University Press, 1951.

SCHUMAN, H., PRESSER, S. Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, 1981.

SUDMAN, S., BRADBURN, N.M. Asking questions. A practical guide to questionnaire design. San Francisco: Jossey-Bass, 1982.

Appendix 1: Assessment of the Coding System as a Means of Evaluating Interviewer Behavior

Code group 10: Correctly asking the question (including presentation of response alternatives)

Code values:

11: Interviewer reads question text as written or with only minor changes, such as "and" or "or", which do not distort the context.

12: Interviewer reads question text with minor changes but without changing the context; no key words are added, omitted or changed.

13: Interviewer reads question text as written and completely, although respondent answers before he is finished.

Code group 20: Incorrectly asking the question

Code values:

21: Interviewer reads question text as written, but alters response alternatives.

22: Interviewer substantially alters question text; key words are added, omitted, changed.

23: Interviewer gives expected response himself; respondent has not yet replied.

Code group 30: Non-directive clarification or nondirected probes

Code values:

31: Interviewer clarifies in a non-directive manner the answer given by the respondent, "non-directive" meaning that a probe by the interviewer does not restrict or alter the context of the question, nor is the context of the (possible) answer by the respondent restricted nor altered.

32: Interviewer repeats the entire question text as written.

34: Interviewer correctly repeats or clarifies the respondent's answer.

35: In response to a question from the respondent, the interviewer correctly clarifies or interprets the question text in his own words or with the help of the question text.

36: Interviewer pauses for an appropriate length of time to give the respondent time to think.

Code group 40: Directive clarification and probes

Code values:

41: Interviewer makes directive comments not contained in the question text, "directive" meaning that the interviewer gives the respondent possible answers or additional contents not contained in the question text. Comments by the interviewer such as "Was that the right one?" are also regarded as directive.

42: Interviewer incorrectly repeats question or parts of question, that is, in repeating the question he does not read the text as contained in the questionnaire.

43: Interviewer uses directive introductory sentence not contained in the questionnaire.

44: Interviewer incorrectly sums up the respondent's answer or independently assigns the answer to one of the response alternatives.

45: Interviewer interprets/clarifies question in a way that is at variance with the original text of the question.

46: Interviewer gives additional stimulus (e.g. explanations) in his own words during or after reading the question text, not necessarily in a directive way, but unnecessarily.

47: Interviewer gives the respondent insufficient time or no time at all to think or put his answer into words, going on with the questionnaire.

Code group 50: Other appropriate behavior during the interview

Code values:

51: Interviewer gives permissible assistance with the interview situation in general (not with individual questions), for example by making comments like "There are no right or wrong answers."

58: Interviewer makes other permissible remarks in connection with individual questions, such as "Should I read the question again?"

Code group 60: Other inappropriate behavior

Code values:

62: Interviewer interrupts respondent

63: Interviewer gives his own personal opinion or assessment, for example praising or criticizing the respondent, showing surprise or displeasure.

64: Incorrect technical procedure by the interviewer, such as: allowing the respondent to read the questionnaire, not following instructions about providing assistance, neglecting to correct the respondent for incorrect technical procedure.

65: Interviewer reads his instructions aloud, thinks aloud.

66: Interviewer unnecessarily rephrases respondent's answer

67: Interviewer lets respondent go off on tangent.

68: Interviewer uses technical terms in talking with the respondent, such as [Translator's note: using the English for a German respondent] "rating", "items", etc.

Code group 70: Behavior that cannot be heard on the tape

Code values

71: Interviewer writes down a response already received without asking the question.

72: Interviewer neglects to ask again or clarify a question when an inadequate response is given.

73: Tape is interrupted, contents are missing.

74: Interviewer incorrectly or incompletely records respondent's answer on the questionnaire.

75: Interviewer enters respondent's answer in the questionnaire illegibly.

Code group 80: Correctly following skip instructions

Code value

81: Interviewer skips one or more questions as directed by instructions.

Code group 90: Incorrectly following skip instructions

Code values

91: Interviewer fails to read a question he should have read.

92: Interviewer asks a question he should have skipped.

## Appendix 2: Overview of the Coding System as a Means of Evaluating Question Quality

### Complex 1: First behavioral step by the interviewer prior to the first reaction of the respondent

This complex includes all behavior by the interviewer as prescribed by his instructions that takes place prior to the first reaction of the respondent, i.e. before the respondent answers, asks for clarification or comments in any way on the question put to him, such as :

- reading the introduction

- reading the question text/response alternatives/scales/items

- explanation of the task of the respondent (e.g. the use of lists, instructions on multiple answers etc.)

### I. Introduction

- 101 Interviewer reads as written or with minor alterations that do not change the context an introduction given in the questionnaire for a group of questions or a single question (minor alterations are, for example, words such as "and" or "well")

- 102 Interviewer gives an introduction that is not contained in the questionnaire in a non-directive manner

- 103 Interviewer alters an introduction to a group of questions or an individual questions contained in the questionnaire in a directive manner, i.e. he makes substantial changes in the introduction

- 104 Interviewer gives an introduction not contained in the questionnaire that gives a false impression of the thrust of the question

- 105 Interviewer neglects to read an introduction given in the questionnaire and reads the question text without any introduction or transition

### II. Question text

- 111 Interviewer reads question text as written

- 112 Interviewer reads question text with minor alterations, without changing the context

- 113 Interviewer reads question text with major alterations, i.e. key words or parts of the question text are added, omitted or changed, or several minor changes are made which make the meaning of the question unclear

III. Answer alternatives/scales

- 121 Interviewer reads answer alternatives/scales as written

- 122 Interviewer reads answer alternatives/scales with minor alterations

- 123 Interviewer reads answer alternatives/scales with substantial alterations (analogous to Code 113)

IV. Items

- 131 Interviewer reads items as written

- 132 Interviewer reads items with minor alterations

- 133 Interviewer reads items with substantial alterations (analogous to Code 113)

V. Other improper behavior in reading the question text

- 141 Interviewer does not completely read question text/answer alternatives/scales/items, since target person answers before he has finished

- 142 Interviewer himself gives expected answer before the initial response by the target person, deducing the proper answer from remarks made by the respondent

VI. Skip instructions

- 151 Interviewer asks a question he should have skipped over

- 152 Interviewer does not ask a question he should have asked

VII. Additional explanations prior to the first response by the respondent

- 161 Interviewer gives appropriate explanations on the respondent's task in line with the question design, which are not given verbatim in the questionnaire, but as instructions for the interviewer (e.g. multiple answers possible)

- 162 Interviewer gives inappropriate explanations on the respondent's task in line with the question design, which are not given verbatim in the questionnaire, but as instructions for the interviewer (e.g. one response is permitted, but interviewer implies that multiple answers can be given)

- 163 Interviewer neglects to give an explanation of the respondent's task in line with the question design, which is not given verbatim in the questionnaire but is necessary for adequate response to the question (e.g. interviewer neglects to mention that multiple answers are possible)

- 164 Interviewer gives additional appropriate explanations not included in the questionnaire (e.g. "Take your time and look at all the cards")

- 165 Interviewer gives additional inappropriate explanations (e.g. interviewer expands on the question text by giving addition explanation before the respondent's initial reaction)

Complex 2: Respondent behavior

Complex 2 includes all kinds of behavior shown by the respondent

I. Appropriate responses

- 201 Closed-ended question, designated respondent [Translator's note: "Zielperson", literally "target individual"] gives appropriate response

- 202 Open-ended question, designated respondent gives appropriate response

- 203 Designated respondent thinks aloud, comments on the question

- 204 Designated respondent corrects response and gives another

- 205 Designated respondent gives appropriate response, but tone of voice, hesitation, explanations, inconsistency indicate an invalid response

- 206 Designated respondent comments on/explains response (e.g. a scale)

II. <u>End of response</u>

- 207 Designated respondent signalizes end of response, says that he has said all he wants to

III. <u>Request for clarification of question content and role behavior, both as related to specific questions and in general</u>

- 208 Designated respondent requests clarification of content or meaning of question text, item, scale or response alternatives

- 209 Designated respondent requests clarification of his task, <u>specifically related to a particular question</u> (e.g. is he to answer freely or by choosing an alternative)

- 210 Designated respondent requests repetition of entire question or parts of question

- 211 Designated respondent repeats question or parts of question to clarify context in his own mind

- 212 Designated respondent asks for information about agency commissioning the survey, goal/purpose of the project, selection procedure, role of respondent, <u>in general</u>

IV. <u>Confirmation of interviewer's clarification</u>

- 213 Designated respondent confirms clarification of respondent's and interviewer's role, related to an individual question or in general

- 214 Designated respondent accepts interviewer's offer to explain something again

V. <u>Request for reinforcement of response given</u>

- 215 Designated respondent asks whether his response is correct, seeks feedback from interviewer on his response

VI. <u>Confirmation of response</u>

- 216 Designated respondent confirms response already given in answer to interviewer's question

- 217 Designated respondent confirms answer incorrectly repeated by interviewer

VII. Polite behavior, not going off on a tangent

- 218 Designated respondent makes polite remarks, not going off on a tangent but not directly relating to the question (e.g. "Do you have enough room?")

VIII. Irrelevant comments, going off on a tangent

- 219 Designated respondent gives information that is irrelevant to the question, attempts to carry on discussion, conversation

- 220 Designated respondent responds to irrelevant comments by interviewer

- 221 Designated respondent contradicts interviewer's attempt at clarification

IX. Inappropriate responses

- 222 Closed-ended question, designated respondent gives inappropriate response

- 223 Open-ended question, designated respondent gives inappropriate response

- 224 Designated respondent criticizes question text, scale, items or response alternatives

X. Nonresponse to question

- 225 Designated respondent cannot answer question (lack of information, difficulty remembering, response alternatives do not apply)

- 226 Designated respondent refuses to answer

XI. Interruption of reading of question text

- 227 Designated respondent interrupts

XII. Intervention by third parties

- 228 Other persons present at the interview intervene in the conversation, answer in place of the designated respondent, inform or correct him

## Complex 3: Interviewer behavior following the respondent's initial response

Complex 3 includes all kinds of behavior by the interviewer that takes place after the initial response by the respondent and clarifies:

- role, goal and task of the respondent in general and in connection with individual questions

- interviewer's role in general and in connection with individual questions

- content/meaning of individual questions

I. Clarification of task and content of individual questions

- 301 Interviewer accurately clarifies task/response of the respondent (when it is unclear into which category it should be put) with regard to a question

- 302 Interviewer accurately clarifies content/meaning of. question text, item, scale, response alternatives

- 303 Interviewer accurately repeats question or parts of question

II. General clarification of role, goal, task

- 304 Interviewer . clarifies task, role, goal in general, not related to a specific question, gives information about agency commissioning survey, selection procedures etc.

III. Probing/encouragement to expand on answer to open-ended question

- 305 Interviewer uses nondirected probe(s), using his own words in a non-directive manner

- 306 Interviewer uses probes given for open-ended questions

- 307 Interviewer encourages respondent to give further information, using short words (e.g. "hmm"), signalizes attention and encouragement

- 308 Interviewer allows sufficient time for respondent to think

IV. Establishment of motivation/contact to respondent

- 309 Interviewer makes neutral remark, not going off on a tangent, e.g. in order to create a positive atmosphere, is polite

V. Control of behavior that goes off on a tangent

- 310 Interviewer halts irrelevant behavior, e.g. by reading the next question or making neutral comments that encourage respondent to continue with the interview

VI. Reinforcement of appropriate response

- 311 Interviewer reinforces appropriate response and the recording of that response, often only as a matter of politeness (e.g. "good")

VII. Making sure of response given

- 312 Interviewer makes sure of response given by asking additional question

VIII. Checking whether clarification or repetition of question is desired

- 313 Interviewer asks whether designated respondent needs to have question clarified or repeated

IX. Provocation to go off on a tangent

- 314 Interviewer provokes respondent into going off on tangents that have nothing to do with question/answer process

- 315 Interviewer neglects to keep irrelevant comments initiated by respondent under control

X. Questioning survey credibility by making negative comments

- 316 Interviewer makes negative comments on interviewer's role, aspects of the project or the researcher, individual questions

XI. Evaluation of respondent's answer

- 317 Interviewer gives personal opinion regarding the respondent's answer

XII. Directive behavior that influences the respondent's answer and does not correspond to the intention of the question

- 318 Interviewer interprets/makes comments in a directive manner/inadequately on the content/meaning of: question text, item, scale, response alternatives

- 319 Interviewer inaccurately repeats question or parts of question

- 320 Interviewer gives incorrect information about respondent's task (with regard to individual questions and in general)

- 321 Interviewer gives answer himself

- 322 Interviewer inaccurately rephrases respondent's answer

- 323 Interviewer asks additional question indicating that he considers the response to be inappropriate or incomplete (although it is correct); e.g. too-frequent use of additional questions such as "Can you think of anything else?"

- 324 Interviewer uses improper technical procedure (for questionnaire and material), e.g., forgetting to mix up cards

XIII. Omission of nondirected probes

- 325 Interviewer neglects to use nondirected probes or to clarify answer to open-/closed-ended questions

- 326 Interviewer neglects to use required probes

- 327 Interviewer allows respondent insufficient time to think

XIV. Interrupting respondent

- 328 Interviewer interrupts respondent

Summary of Complex 4: Recording responses on the questionnaire

I. Closed-ended questions

Instructions in questionnaire - Recording

401 One response permitted – Interviewer marks wrong code value

402 Multiple responses permitted – Interviewer marks one or more answer alternatives incorrectly, although the number of alternatives marked corresponds to the number of responses given

403 Multiple responses permitted – Interviewer marks fewer responses than given by respondent

404 Multiple responses permitted – Interviewer marks more responses than given by respondent

II. Open-ended questions

405 Parts of or entire record illegible

406 Interviewer alters information substantially in writing it down, recording more, less or changed information

407 Interruption of tape

408 Tape impossible to understand, garbled