CODING INTERVIEWER BEHAVIOR AS A METHOD OF EVALUATING PERFORMANCE[1]

Nancy A. Mathiowetz and Charles F. Cannell, Survey Research Center
The University of Michigan

I. Introduction

The use of computer assisted telephone interviewing (CATI) systems gives researchers greater ability to maintain quality control over the interviewing process. Monitoring interviewer performance is particularly well suited to a CATI system which permits unobtrusive observation of the interviewer's behavior through the use of a slave screen which replicates the interviewer's screen for the monitor. This technique also allows the supervisor to observe the interviewer's accuracy in recording responses. Monitoring interviewers is not a novel idea; most survey organizations do some monitoring of face-to-face and telephone interviewers. However, this usually consists simply of listening to interviews and noting problems as they become apparent. This approach to monitoring interviewers does not provide the researcher with information from which inferences concerning the quality of data or interviewer variation can be made. This paper describes a technique for systematically evaluating telephone interviewers based on coding their behavior. Results of the initial application of these techniques to a computer-assited telephone survey are also presented. The monitoring system described here is based on one developed earlier using tape recordings of face-to-face interviews to code.[2]

Programmed monitoring allows the researcher or supervisor to evaluate the most important aspect of an interviewer's performance; the techniques used in interacting with the respondent. The monitoring technique we used identifies the major categories of interviewer behavior and classifies each behavior as correct or incorrect or inappropriate, according to the concepts and specifications for that particular study. For example, there are several categories which assess the interviewer's delivery of a question. Was the question read exactly as printed? Did the interviewer make minor changes in reading without changing the intent of the question? A list of codes used in evaluating the interviewers' techniques is given in Figure 1.

The coding system has three major purposes.
1) In initial training, it helps to teach the novice interviewer which interviewing techniques are acceptable and which are not.
2) It serves as a basis for interviewers and supervisors to review work by coding interviews and discussing the problems which the coding reveals. After monitoring an interview, the supervisor and interviewer review both the good and bad aspects of the interviewer's performance. During these review sessions, both positive comments as well as corrective instruction can be given to the interviewers.
3) It provides an assessment of an interviewer's performance, which can be compared both with the performances of other interviewers and with the individual's own performances during other interviews. In order to make such comparisons, the distribution of good and poor behavior for each interviewer is compared with the distribution for all interviewers. The monitored data may also be useful in identifying questions which give the interviewer problems. For example, questions frequently asked incorrectly are awkardly worded; those frequently repeated were not readily understood by the respondent, etc.

Figure I

CODES FOR MONITORING INTERVIEWER BEHAVIOR

QUESTION-ASKING
- 11 Reads question exactly as printed
- 12 Reads question incorrectly-minor changes
- 16 Reads question incorrectly-major changes
- 17 Fails to read a question

REPEATING QUESTIONS
- 21 Repeats question correctly
- 25 Repeats question-unnecessarily
- 26 Repeats question-incorrectly
- 27 Fails to repeat question

DEFINING/CLARIFYING
- 31 Clarifies or defines correctly
- 35 Defines or clarifies-unnecessarily
- 36 Defines or clarifies-incorrectly
- 37 Fails to define or clarify

SHORT FEEDBACK
- 41 Delivers short feedback-correctly
- 45 Delivers short feedback-inappropriately
- 46 Delivers short feedback-incorrectly
- 47 Fails to deliver short feedback

LONG FEEDBACK
- 51 Delivers long feedback-correctly
- 55 Delivers long feedback-inappropriately
- 56 Delivers long feedback-incorrectly
- 57 Fails to deliver long feedback

PACE/TIMING
- 65 Reads item too fast or too slow
- 66 Timing between items-too fast
- 67 Timing between items-too slow

OVERALL CLARITY
- 75 "unnatural" manner of reading item (poor inflection, exaggerated or inadequate emphasis, "wooden" or monotone expression)
- 76 Mispronunciation leading to (possible) misinterpretation

II. Behavior Codes

The coding system is quite flexible and can be adapted to the purposes of a particular study. The codes shown here (see Figure I), for example, were developed for a methodological study using experimental interviewing techniques.

Some explanation of the codes is necessary before describing the operation of the technique. The system is organized around the major activities of an interviewer: (1) question-asking; (2) repeating questions; (3) defining and clarifying; and (4) giving feedback. The 10-50's codes are

used to identify concrete behavior (or lack of behavior) and determine whether its occurrence was correct and appropriate. The 60 and 70 codes require that the monitor evaluate the quality of the delivery.

III. Procedures

In our application of the system, the interviewing supervisors were trained as monitors. We felt that the same person who monitored the interviewer should provide the interviewer with an evaluation of the work. This became an important part of the supervisor's functions. Monitoring takes time from other supervisory functions and the researcher usually needs to compromise between the amount of monitoring considered ideal with the added time and cost involved. By allocating 40 hours per week, we were able to monitor and provide interviewers with evaluations from a major segment of the interview from approximately 15% of the interviews taken during eight weeks of interviewing.

Throughout the study period, reliability coding was done. The codes for evaluating concrete behavior (10-50) were fairly constant over the entire field period (mean = 88.4%, s.d. = 3.2) whereas similar to the pre-production reliability measures, the subjective codes (60 and 70) showed less consistency both between and within monitors (mean = 79.3%, s.d. = 5.2%).

Although it was not difficult to identify or to train supervisors to code types of behavior, it was far more challenging to identify specific models of pace and clarity of speech. Little research has been done on how the quality of speech affects responses. It may, however, be the key as to what distinguishes a highly successful interviewer from others. For this reason, we thought it important to code such behaviors, even though the reliability was lower than we would have liked.

IV. Analysis

There are several questions which can be addressed in analysis of the monitoring data. They include:

1. Do interviewers differ in their behavior among question types? Among individual questions?
2. Does behavior differ over the time of the study period? Do interviewing techniques improve with experience?
3. How markedly do interviewers differ in their correct and incorrect use of techniques among themselves?

An examination of monitoring codes by individual questions is not instructive since the cell sizes were small. Significant findings, however, did result from comparisons of behavior for types of questions.

The major type of question which we considered might show differences in interviewer behavior was whether it was open or closed. Table 1 compares the mean proportion of open and closed questions exhibiting various interviewer behaviors. The statistic was calculated in two steps. First, since there was an equal number of questions observed per monitored interview, the proportion of each specific type of behavior within a major category of behavior was calculated for each interview, e.g., the proportion of questions read correctly over all questions read for that interview. The mean of these proportions for open and closed questions, across all interviews is reported in Table 1. One reason for making this calculation was that any one interview could be monitored for one or more segments of eligible questions. Once again, proportion of behavior within each interview minimizes the effects of any one interview contributing more than the mean number of questions per interview to the statistic.

TABLE 1. MEAN PROPORTION OF VARIOUS INTERVIEWER BEHAVIORS BY QUESTION TYPE

Behavior	Mean proportion across interviews: Open Q's	Closed Q's	t[a]
I. Question Delivery:			
A. Correct reading	95.8%	95.4%	N.S.[b]
B. Minor changes	1.9	3.7	1.89*
C. Major changes	.5	.4	N.S.[b]
D. Fails to read	1.8	.5	2.98**
	100.0%	100.0%	
II. Probing:			
A. Correct repeated Q and/or definition	6.1%	8.3%	2.01*
B. Incorrect use or probe	1.6	1.4	N.S.[b]
C. No probe required	92.3	90.3	
	100.0%	100.0%	
III. Feedback:			
A. Correct feedback	17.1%	30.5%	7.08***
B. Incorrect or inappropriate feedback	1.8	1.0	N.S.[b]
C. Fails to give feedb'k	.7	1.2	N.S.[b]
D. No feedback req'd	80.5	67.3	
	100.0%	100.0%	

N = 208 interviews

[a] t-test based upon a paired comparison of behavior for open and closed questions. Due to the unequal number of questions monitored for any one interview, proportion of each behavior within an interview was calculated. The mean of that proportion is reported here.

[b] N.S. = $p > .06$

* $p \leq .06$

** $p \leq .01$

*** $p \leq .001$

Interviewers behaved differently in asking open and closed questions, as reflected in the higher proportion of questions in which the interviewer made minor changes in the question wording and the increased need to define terms. Many of the closed questions used in this study were scale items or lists of similar questions. The increased length of questions which introduce scales to a respondent, may merely by the number of phrases in the question, provide the interviewer with a greater opportunity to err. The repetitive nature of lists of similar questions may introduce an element of boredom for the interviewer, the result being a less precise delivery of the question. In examining separately the data for the scale questions and for lists of similar, single phrase questions, it appears that the

length of the question may be the best explanation for interviewer variance in delivery of the queries. The increased length of these questions is due primarily to the number of response categories the interviewer must read.

Open questions, however, present the interviewer with the greater challenge to elicit an adequate response. Probing is less often done correctly for open questions, usually as a result of an interviewer's failure to repeat a question when necessary. Further explanation of the mean proportions reported for feedback behavior is necessary. The experimental design of the study included three forms of the questionnaire, differing in the amount of instructions and feedback given to the respondent. The last three rows of Table 1 show a significant difference in feedback for open and closed questions. The difference in the amount of incorrect feedback may be due, not to the question type, but rather to the nature of the feedback for the two question types. Open questions usually had long feedback statements given to respondents following an adequate response; closed questions had more short feedback statements. Of more interest than the amount of "incorrect" behavior is the difference in the amount of inappropriate feedback given for open and closed questions.

Several changes in the interviewers' behavior took place over the course of the study period. Although only slightly variable, the percent of questions read correctly increased until the seventh week of the study, at which point the percentage dropped. There are two possible explanations for this phenomena. One probable explanation for this finding lies in the change of the nature of the respondents late in the study. Through the first seven weeks of interviewing, many of the respondents required little or no persuasion to consent to the interview, whereas the later respondents were the result of difficult refusal conversions. However, one could also explain these findings as a result of interviewers becoming bored or forgetful late in the study. Unfortunately, the effects of learning and the differences in respondents cannot be separated. The percentage of correct and appropriate probing increased steadily throughout the study.

Table 2 presents a summary comparison of interviewer variation for the major categories of behavior. The mean of the proportion of behavior within each interview for each interviewer was used for this comparison. Even with the special emphasis given to training interviewers for this study and the constant feedback given to them throughout the study, interviewers showed a significent variation for several of the behaviors. Forthcoming analyses will be to determine whether a small set of interviewers was responsible for a large proportion of the variance. If this proves to be true, one may want to compare the results of the interviewing data excluding those interviewers with the results calculated for all interviewers.

TABLE 2. DESCRIPTIVE MEASURES OF INTERVIEWER VARIATION

Behavior	Mean proportion across interviewers		
	Mean[a]	S.D.	F[b]
I. Question Delivery:			
A. Read correctly	95.7%	4.6%	2.34***
B. Minor changes	2.6	2.8	2.08**
C. Major changes	.4	1.7	11.48***
D. Fails to read	.8	2.4	1.69*
E. Question read too fast	6.6	7.7	3.60***
F. Unnatural delivery	.4	1.3	2.92***
II. Probing:			
A. Correct probing	7.9	4.6	N.S.[c]
B. Incorrect probing	1.5	1.8	N.S.[c]
III. Feedback:			
A. Correct feedback	24.9	12.2	N.S.[c]
B. Incorrect	2.0	2.9	3.59***
C. Inappropriate	.8	1.8	1.62*
D. Fails to give feedback	1.2	2.0	N.S.[c]

N = 26 interviewers

[a] Proportion of each behavior calculated within each interview; mean proportion for each interviewer was determined. Reported mean equals the mean of all interviewers' mean proportions.

[b] F statistic calculated from a one-way analysis of variance.

[c] N.S. = $p \geq .05$

* $p < .05$ ** $p \leq .01$ *** $p \leq .001$

Conclusions

Monitoring interviewers is neither a new idea nor one which is unique to either telephone or computer-assisted telephone interviews. Tape-recorded personal interviews have been coded in a manner similar to that described here. However, the centralized telephone interviewing unit provides the researcher with the ability to have more control over the sampling of interviews to monitor, and can do so without the interviewer's knowledge of when they are being monitored. The use of a CATI system greatly aided the means by which we randomized our selection of interviews.

The unique quality of this monitoring technique is that it uses a specific set of codes to identify and classify interviewer behavior. Thus, both the interviewer and supervisor are aware of the basis for evaluating performance. By recording the observation codes, the researcher has the means by which to understand how the interaction among the questionnaire, interviewer, and respondent affects the quality of the interviewing data.

FOOTNOTES

[1]This material is based upon work supported by the National Science Foundation under Grant No. SOC78-07287. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[2]Cannell, Charles F., Sally A. Lawson, and Doris L. Hausser. A Technique for Evaluating Interviewer Performance. Ann Arbor, Michigan: Survey Research Center, The University of Michigan, 1975.

REFERENCES

Cannell, Charles F. and Sally Robinson. "Analysis of Individual Questions," Chapter 11 in Working Papers on Survey Research in Poverty Areas, John B. Lansing, Stephen B. Withey, and Arthur C. Wolfe (eds.). Ann Arbor, Michigan: Survey Research Center, The University of Michigan, 1971.

Marquis, Kent H. "Purpose and Procedure of the Tape Recording Analysis," Chapter 10 in Working Papers on Survey Research in Poverty Areas, John B. Lansing, Stephen B. Withey and Arthur C. Wolfe (eds.). Ann Arbor, Michigan: Survey Research Center, The University of Michigan, 1971.

Morton-Williams, Jane. "Use of Verbal Interaction Coding for Evaluating a Questionnaire," Quality and Quantity (Britain), 13, 1979:59-75.