STANDARDIZED SURVEY INTERVIEWING

Floyd J. Fowler, Jr.
Thomas W. Mangione

INTRODUCTION

One goal of survey management is to have the interviewing be standardized. By this we mean the data collected should be unaffected by which particular interviewer conducts an interview. Failure to achieve this can produce either bias, estimates which are systematically different from some measure of the true population value, and/or simply unreliability, inflating the standard errors of estimates. In either case, minimizing interviewer effects is highly desirable.

There are four basic behavioral techniques that interviewers are taught to minimize their effects on survey data:

1. They are to read survey questions exactly as they are written.

2. When a respondent gives an inadequate answer to a question, one that does not meet question objectives, the interviewer's response, probing that inadequate answer, should be standardized and nondirective so that the likelihood of one answer over other answers is not affected.

3. There should be no interviewer discretion in the recording of survey answers: the interviewer is not to record an answer that has not actually been given by the respondent; when the respondent is supposed to answer in his or her own words, verbatim recording of answers is required.

4. The interviewer is to carry out the interaction with a respondent in a nonbiasing way, particularly refraining from presenting information about him or herself or commenting on the respondent's answers in ways that would indicate a preference for some kinds of answers over others.

In order to get interviewers to perform in these ways, interviewers receive a certain amount of training before they begin interviewing. Once they begin, researchers exercise some kind of supervision over the interviewer's work. Practices in both of these respects, however, differ widely. On the training side, interviewer training programs for some academic survey organizations and for government agencies often last five or more days. However, surveys are also done by interviewers who receive only a few hours of training; in some cases the researchers themselves have no direct role in or knowledge of the training that interviewers receive.

On the supervision side, all survey organizations review interviewer costs. If probability samples are used, response rates will be calculated. Conscientious survey organizations also review completed interviews to make sure that instructions are followed and question objectives are met. However, none of these activities relates to the quality of the interviewing process itself, the way in which the question and answer process is carried out. Yet, it is the way that the interviewer carries out the interaction with respondents which is the key to standardized interviewing.

When personal interview studies are being carried out in people's homes, the only ways to directly supervise an interviewer's process performance are observation and tape recording. Neither is a norm in current survey research practice. Of course, telephone surveys from centralized facilities provide great potential for monitoring interviewer behavior.

Although there is no question that some training of interviewers is essential, and probably most researchers would agree that some observation or tape recording is valuable, heretofore neither researchers nor those who would contract for survey research have had any empirical basis for saying how much training is enough and how much will be gained from investing in tape recording or observation. This paper presents data from a large-scale field experiment designed to provide that kind of information.

METHOD

The data were created by carrying out a special-purpose experiment to test the efficacy of four different training programs and three different approaches to the supervision of field interviewers. Sixty persons who met usual standards for being a survey interviewer, but without previous professional interviewing experience, were recruited, hired, and randomly assigned to one of four training programs. Three of these training programs were designed to replicate typical training experiences in survey research for basic interviewing skills: a program of less than one day, which consisted of a two hour lecture, a demonstration interview, reading a manual, and no supervised practice. A two-day program and five-day program were designed to approximate the two extremes of training programs most common among academic survey organizations. These differed from the one-day session in the extent to which interviewers were involved in discussions of procedures and, in particular, in the extent to which they had supervised practice and exercises. In addition, as a way of gaining a

benchmark on the full potential of training to influence interviewers, a fourth program was designed which consisted of ten days of training in basic interviewing skills.

Once training was completed, interviewers were randomly assigned, in a balanced design, to one of three programs of supervision. In each program, interviewers had a once-a-week conversation with a supervisor on the telephone. However, the content of the conversation varied. In Supervision Level I, interviewers received feedback only about their production efficiency, the number of hours they worked, and their response rates. In Level II, interviewers in addition received routine evaluation from a review of a sample of their completed interviews. In Level III, interviewers tape recorded all of their interviews; a sample was reviewed and evaluated each week. This was the only supervision level which allowed direct feedback on the interviewing process.

A critical feature of the design was that each interviewer received an assignment of 40 addresses which was a random subsample of the total sample. In this way, differences in interviewer behavior and results could be attributed to the interviewers and not to idiosyncracies of their samples.

Interviewers used a specially constructed half-hour health survey questionnaire, designed to include a sampling of various types of survey items: opinion and factual, open-ended and closed, difficult and easy, sensitive and not sensitive. Many of these items were identical to those used in the National Health Interview Survey.

Data that could be used to help evaluate interviewer performance, in addition to the health interviews, included a tape recorded practice interview which all interviewers conducted after training but before beginning production, a telephone reinterview with almost all of the health survey respondents about their reaction to the interviewer and the interview, and a self-administered questionnaire that interviewers completed about their interviewing experience after their work on the study was over. Also, of course, the interviewers who were assigned to Supervision Level III, which consisted of equal numbers of interviewers from each training program, tape recorded all of their interviews, except when respondents demurred (a reasonably rare event) or when their tape recording equipment malfunctioned.

Of the 60 interviewers who completed training and were given an assignment, 52 completed their entire assignment, which produced an average of 23 interviews. Five other interviewers completed a random half of their assignment. Hence, the analysis presented here is based on the results of 57 interviewers. Much of the analysis is focussed on the 20 interviewers whose work was tape recorded.

THE VALUE OF TRAINING

A key reason for tape recording interviews is to find out how interviewers actually carry out their job. Without direct observation, we have no real information about how well interviewers perform the tasks they are asked to do. Hence, our best information about the quality of standardized interviewing techniques comes from the twenty interviewers in our study who tape recorded all their interviews.

Some four hundred interviews that they tape recorded were carefully coded by specially trained coders. Among other things, coders tabulated the rate at which interviewers changed question wording, used directive probes, failed to record answers according to instructions, or gave respondents feedback between questions which was inappropriate, usually because it was commenting on the answer in an evaluative way. In addition, based on these and other counts, coders rated each interviewer's performance in each tape recorded interview on a four point scale: excellent, satisfactory, needs improvement, or unsatisfactory. Topics covered by these ratings included reading questions, probing open-ended and closed-ended questions, recording open-ended and closed-ended questions, and creating a nonbiasing interpersonal environment.

The results of this coding, which, of course, was done by coders who were unaware of the training background of the interviewers involved, are displayed in Table 1, tabulated by the kind of training program to which interviewers were assigned. The results are relatively uniform. On all measures except that pertaining to recording answers to closed questions, there was a significant effect of training on interviewer behavior. The clearest difference is that those who received less than one day's training were consistently much less good at interviewing than those who received two or more days training. There is, however, a general increase associated with training, with the more difficult interviewer tasks, such as probing and recording open-ended answers, distinctively benefiting from increased training.

THE EFFECT OF SUPERVISION

We were interested in measuring the extent to which interviewers changed during the course of the study. To give us a measure, we divided each interviewer's total assignment randomly in half, asking him or her to complete the first half before beginning the second. In this way, we were able to compare the results of early and later work on comparable samples.

Table 2 presents the same measures presented in Table 1 by whether an interview was taken in the first or second half of an interviewer's assignment. On average, interviewers took approximately 23 interviews; so, each half consisted of approximately eleven interviews.

It can be seen that with respect to all but one measure, using directive probes, there was no significant improvement associated with experience under the tape recorded method of supervision. For the most part, interviewers skills remained stable or improved slightly between the first and second half of their assignments.

We thought it possible that the least trained interviewers would improve most. This was not true. Table 3 shows two representative results. The evaluation of tape reviews produced virtually no improvement in the 1-day trainees, even though they had far to go to be satisfactory. Interestingly, it seemed to be those who had 5 days of training who showed the most improvement when they were tape recorded.

We were particularly interested in finding out what happened to interviewer skills when the interviewing process was not directly supervised. In the absence of tape recordings, we had only two sources of information about how interviewers performed: ratings made by respondents and the self ratings of interviewers of their own work. Of the various skills that we were interested in, the only one that we had some hope that respondents could rate was whether or not interviewers read questions exactly as worded. The quality of probing, recording, and the interpersonal environment created by the interviewer were beyond the respondents' abilities to evaluate, as indicated by low correlations between their answers and the coders ratings for corresponding interviews. It turns out that the respondents' ratings of how precisely interviewers read questions correlated modestly, .35, with our tape coders' ratings. Moreover, the interviewer's own self ratings of their performance in that regard correlated well with those of tape coders. Thus, it seems as if respondents and interviewers have some information to give about the way that interviewers actually read questions in the interview.

As one would glean from the data in Table 2, the performance level of interviewers on average was fairly stable after they completed training. All interviewers, regardless of their supervision program, tape recorded a practice interview after training but before beginning their assignment. The skills that they evinced in that practice interview, including how well they succeeded in reading questions exactly, correlated well with the way they performed in the actual survey according to our tape reviews. Hence, we reasoned that these preproduction ratings should also correlate with the respondents' reports and with the interviewers' self ratings.

Building on this information, Table 4 presents a very interesting pattern. In Table 4, when interviewers were tape recorded, respondent ratings and interviewer self-ratings of how well they read questions correlated significantly with coding of preproduction performance. However, for interviewers who were not tape recorded, there is no relationship between the way that respondents and interviewers said that these interviewers read questions and the level of skill they showed just after they completed training. This suggests that behavior deteriorated for a significant number of interviewers who were not taped.

The results of another approach to examining this issue are presented in Table 5. Here, the ratings of respondents of how well interviewers read questions are compared for the first half and the second half of an interviewer's work. It can be seen that when interviewers were tape recorded, there is a seven percentage point increase in the rate at which respondents thought that interviewers read questions exactly as worded. In contrast, when interviewers were not tape recorded, there was a five percentage point decrease over in the rate at which respondents said that interviewers read questions exactly as worded.

DISCUSSION

There are three main conclusions that we derive from these analyses.

1. There is a major difference between the interviewing skills demonstrated by people who receive two or more days of training and those who receive less than one day of training. There are many survey studies done that utilize interviewers who receive training that is equal to or less than the training received by our one-day training group. Many survey organizations contract with interviewing pools with which they have no direct contact and no direct knowledge of the kind or quality of training in basic interviewing skills that interviewers have received. One-time studies sometimes are done by recruiting volunteers or other new interviewers who are given only an hour or two of training. In this study, interviewers were paid to read a professional interviewer manual. The training was done by highly experienced field supervisors. This three-hour training program was far from the worst training program to which interviewers could be exposed. Yet, to repeat, the skill levels that even they demonstrated were clearly far inferior to those produced by those who received more training in basic interviewing skills.

2. Second, it is important to realize that even with relatively intensive supervision interviewers did not get much better in critial interviewing skills with experience. Improvement was modest. For the most part, the level of skill interviewers had when they completed training was close to the highest level of skill they would ever have as interviewers. Very intensive and detailed supervision of the interviewing process only produces stability or, possibly, a slight improvement.

3. Perhaps the most important implication of our data, however, is that the stability of performance associated with tape recording is not to be dismissed lightly. Our data support the conclusion that without direct supervision of the question and answer process, the performance of some interviewers is likely to deteriorate over time. Rather than improving skills, experience is likely to erode them.

Indeed this would seem to be a very likely outcome on logical grounds as well as on the basis of our data. Being a standardized, nondirective interviewer involves using a relatively complex set of skills. In the absence of direct observation of these skills in practice, interviewers are neither rewarded for performing well at these difficult tasks, nor can they have much sense that the way they perform the interviewing task is important to the organization for which they work. At the same time, it is clear that there are pressures felt by interviewers from respondents to relate to them in more personal and less formal ways. Hence, we think it is quite reasonable that if interviewer performance is not evaluated through observation or tape recording, some interviewers will tend not to be rigorous about using the interviewer techniques they are taught.

It should be added that our evidence for this hypothesis does not rest solely on the data presented here. When we looked at interviewer effects, the extent to which differences in answers could be associated with the interviewer, we found a significant interaction between the kind of training interviewers received and the kind of supervision they received. Those with the least training benefited distinctly from having their work tape recorded. Without it, those interviewers were very inconsistent. In addition, the best trained interviewers, despite the fact that they had the best skills at the the time training was over, proved to be inconsistent if they were not tape recorded. We think it likely that those highly trained but less supervised interviewers behaved like "old pros"; they were self confident; they knew how to do the job and "freelanced" in inconsistent, and possibly counterproductive, ways, unless they were held accountable for their interviewing by tape recorded supervision.

In conclusion then, these analyses suggest that interviewers who receive only a few hours of training in basic interviewer techniques will not have basic standardized interviewing skills. Perhaps the most distinctive contribution of the analyses, however, is to point to the importance of the combination of training and supervision in predicting and maintaining interviewer performance. In particular, our analyses suggest that the inclusion of direct information about the interviewing process, either through tape recording or observation, may be an essential technique for producing standardized survey interviewing.

Table 1
Selected Measures of Interviewer Behavior from Coding Taped Interviews
By Training Program
(Supervision Level III Only)

Percentage of Interviews Rated Excellent or Satisfactory	Length of Training Program				
	< 1 day	2 days	5 days	10 days	P**
Reading Questions as Worded	30	83	72	84	<.01
Probing Closed Questions	48	67	72	80	<.01
Probing Open Questions	16	44	52	69	<.01
Recording Answers to Closed Questions	88	88	89	93	N.S.
Recording Answers to Open Questions	55	80	67	83	<.01
Non-biasing Interpersonal Behavior	66	95	85	90	<.01

* Less than 0.5.
** Based on f test; df corrected for intraclass correlation within interviewer scores. Based on about 320 interviews coded for 20 interviewers.

Table 2

Selected Measures of Interviewer Behavior from Coding of Taped Interviews for 1st and 2nd Half of Interviewers' Assignment (All Training Groups, Supervision Level III Only)

	Portion of Assignment		
	1st Half	2nd Half	P **
Percentage of Interviews Rated Excellent or Satisfactory			
Reading Questions as Worded	63	70	N.S.
Probing Closed Questions	61	73	<.05
Probing Open Questions	42	48	N.S.
Recording Answers to Closed Questions	88	91	N.S.
Recording Answers to Open Questions	68	74	N.S.
Non-biasing Interpersonal Behavior	81	87	N.S.

** Based on F test; df corrected for intraclass correlation within interviewer scores. Based on about 320 interviews coded for 20 interviewers.

Table 3

Rating of Two Interviewer Behaviors from Coding 1st and 2nd Half of Interviewers' Assignment by Training Program (Supervision Level III only)

Percent Rated Excellent or Satisfactory Reading Questions as Worded

Length of Training Program	Portion of Assignment	
	1st Half	2nd Half
1 day	32	28
2 days	83	84
5 days	64	79
40 days	85	83

Percent Rated Excellent or Satisfactory Probing Open Questions

Length of Training Program	Portion of Assignment	
	1st Half	2nd Half
< 1 day	18	13
2 days	57	54
5 days	44	61
10 days	70	69

Table 4

Correlation of Production Measures of Question Reading to Pre-Production Rating[1] By Type of Supervision

Pre-production Rating of Question Reading	Measures of Question Reading: Production Tape Rating	Average Respondent Rating	Interviewer Self Rating
Supervision Levels I & II (Untaped) (N=37)	N.A.	-.02	.06
Supervision Level III (Taped) (N=20)	.65**	.35*	.59**

* p = .06
** p = <.01
1. Based on coding a practice interview taken after training but before beginning production.

Table 5

Difference in Percentage of Early and Late Respondents' Ratings of Question Reading by Level of Supervision*

Level of Supervision	Percentage Difference Between Early and Late Respondents who Said Questions Read Exactly
Level I (Not taped)	-5
Level II	-5
Level III (Taped)	+7

* Negative number means there were fewer respondents giving a response in the 2nd half of an interviewer's assignment.