THE ROLE OF INTERVIEWER TRAINING AND SUPERVISION IN REDUCING INTERVIEWER EFFECTS ON SURVEY DATA

Floyd J. Fowler, Jr. and Thomas W. Mangione
Center For Survey Research - UMASS/Boston

INTRODUCTION

One goal of good survey methodology is to minimize the extent to which interviewers influence the answers they obtain. Although it is not possible routinely to dissociate the effect of interviewers from the idiosyncracies of the samples they interview, there are a number of reports in the literature in which interviewer effects have been estimated. (e.g. Bailar et al, 1977; Freeman and Butler, 1976; Groves and Kahn, 1979; Groves and Magilavy, 1980; Hanson and Marks, 1958; Kish, 1962). These studies show that in most survey instruments there is a range in the extent to which interviewers affect answers. Although the mix varies from study to study, perhaps a third of the items in most surveys have an intraclass correlation associated with the interviewer of .01 or higher (see Groves and Kahn, 1979). Although the exact impact of such an intraclass correlation, or rho, on the total error of a survey estimate depends on the average size of interviewer assignments, there is no question that these interviewer effects produce significant, unwanted error in many survey-based estimates.

Another goal, not necessarily related, is to maximize the accuracy of the answers reported in surveys. Although accuracy of survey reporting is not often assessed, there are studies which show that interviewers have a clear role to play in affecting the accuracy of survey answers (e.g. Cannell, 1977; Cannell and Fowler, 1964).

The specific aims of the analyses reported here were to examine the potential of interviewer training and supervision for increasing the standardization of interviewers and increasing the accuracy of data they obtain.

THE EXPERIMENT

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 the 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 interviewer training programs. Three of these were designed to replicate typical programs in survey research: a program that took less than one day, a two-day training program, and a five-day training program. In addition, as a way of gaining a benchmark on the full potential of training to affect interviewers, a fourth program consisted of ten days of training and practice.

Once training was completed, interviewers were randomly assigned, in a balanced design, to one of three programs of supervision. In level 1, interviewers received feedback only about production and response rates. In level 2, interviewers in addition received routine review and feedback about the quality of their completed interviews. In supervision level 3, all interviews were tape recorded, a sample was reviewed and interviewers received systematic feedback about the quality of their interviewing techniques.

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 the data collected by interviewers assigned to different training or supervision programs could be attributed to the interviewers and not to idiosyncracies of their samples.

Interviewers used a specially constructed 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.

Of the 60 interviewers who completed training and were given an assignment, 52 completed their assignment. Five other interviewers completed a random half of their assignment. Hence, the analysis presented here is based on the results of 57 interviewers.

Interviewer Effects

The intraclass correlation, or rho, has been proposed by many authors, including Kish (1962), as a measure of the effect of interviewers on data which itself is not affected by the number of interviews per interviewer. It is calculated as the percentage of the total variance that can be associated with interviewers. We carried out a random effects ANOVA using the interviewer as the random effect. The results permitted calculation of rho as follows:

$$\text{rho} = \frac{V_I^2}{\text{MSE} + V_I^2}$$

$$V_I^2 = \frac{\text{Mean Square Model} - \text{Mean Square Error}}{n}$$

n = average number of interviews per interviewer

Rhos were calculated for each of the 148 items in our survey instrument. Like others,

we found considerable range. We were most interested in examining the hypothesis that training and supervision are effective ways to minimize interviewer effects on those measures which are most affected by interviewers. Therefore, we restricted analysis of training and supervision effects to the 54 items which our ANOVA indicated ($p < .10$), were most likely to be influenced by interviewers.

Next, we calculated rhos for each of the 12 cells defined by the 4 training and 3 supervision treatments for each of those 54 items. Transforms of these intraclass correlations, Log (rho/(1-rho)), are the dependent variables used in a two-way analysis of variance. The results and actual rhos are presented in Table 1.

Table 1 is a very orderly table with two major exceptions: the interviewers who received ten days of training but were not tape recorded (Supervision Levels I and II) produce much higher rhos than one would expect from the rest of the data. As a result, there is not a positive main effect of training on rho. Rather, the ANOVA picks up the importance of the combination of training and supervision interviewers received.

The influence of supervision appeared mainly in two training groups: those who received the least training and those who received the most. In both cases, when interviews were tape recorded, the quality of interviewing approached the best levels that were generated. However, less thorough supervision produced less satisfactory interviewer performance as measured by rho. In particular, for those interviewers who received ten days of training, their performance was very poor from the standpoint of reliability when the actual question and answer process was not supervised.

Finally, it should be noted that while the two-way ANOVA did not indicate a statistically significant effect of supervision, we also did a direct comparison of taping versus not taping overall. When we corrected the standard errors to adjust for the fact that there are 54 measures per training-supervision group, and hence all observations are not independent, the value of t is 1.5.

Interviewer Bias

For many kinds of survey items, an analysis which shows that interviewers obtain different answers does not provide any information about which answers are best. In the preceding analysis, the sole focus is on reliability. Accuracy is not assessed. It usually is impossible to evaluate the quality of survey data without an objective criterion. However, it is known that there are certain systematic biases that recur in social science research.

TABLE 1

Average Rho (x1000) by Level of Training and Supervision

	Supervision			
Training	Level 1	Level 2	Level 3	Average
<1 day	14	10	8	11
2 days	12	6	11	9
5 days	9	9	8	9
10 days	15	20	7	12
Average	12	10	8	10

Analysis includes only items for which interviewers affected the answers ($p < .10$ by F test). Analysis of variance results: Training effect ($F = 1.94$, $df = 3$, $p > .10$); supervision effect ($F = 1.92$, $df = 2$, $p > .10$); interaction ($F = 2.09$; $df = 6$; $p < .05$).

Six members of our research team independently evaluated every item in the questionnaire. One aspect of the evaluation was whether or not a direction of a "better answer" in aggregate could be specified. Some questions were so rated because it was thought that underreporting was likely. For example, Cannell (1977) has used number of chronic conditions and number of visits to doctors as indicators of reporting quality. Other items were so rated because some of the response alternatives were judged to be sensitive either because they were socially valued or because they were asking respondents to admit to characteristics that might lead to being critized or under valued. Obviously, there were some questions which had both of those properties. Altogether there were 57 items for which the research staff felt that there was a direction in which a "better" answer could be specified. Even though a different set of raters might not agree completely, we think this set of items gives us one way of evaluating the likelihood of interviewer bias and another way to assess interviewer quality.

It will be recalled that each interviewer's sample was a random subsample of the total sample. On average, an interviewer interviewed twenty-five respondents. On each of the 57 items, the mean of each interviewer's responses was calculated. This mean was then translated into a signed standard score, based on the standard deviation of the total distribution of the sample. The procedure meant that an interviewer whose respondents gave answers on average that were more in the direction of the difficult answer or sensitive answer had positive scores; interviewers whose respondents reports were in the direction judged to be less accurate were given negative scores.

Once those calculations were made, the 3249 observations (57 interviewers scores on 57 items) were subjected to a two-way analysis of variance to test the hypothesis that training, supervision or a combination thereof would lead to interviewers obtaining, on average, less biased responses. The results and the actual average standard scores are in Table 2. Table 3 presents some specific contrasts emanating from Table 2.

We hesitate to put precise probability estimates on the contrasts observed for at least two reasons. First, although a one-tailed test is probably reasonable, since there is little basis for predicting more training or supervision would make interviewers worse, there are a few examples in our analyses where opposite effects may be occurring. Second, although the data presented in Table 3 are completely consistent with our original analysis plan, the selection was informed by looking at Table 2. Nonetheless, for the exploratory purposes at hand, we think the t values in Table 3 a provide useful indication of the probability of various observed differences.

As was the case for rho, neither training nor supervision alone enabled us to predict well how biased the answers interviewers obtained would be, though for all items combined there is almost a significant relationship to training. However, for all analyses, when one takes into account the combined pattern of training and supervision, there is a significant relationship to the amount of bias that interviewers obtained.

The main effect of training is not monotonic. Specifically, those interviewers who received five days of training did not fall into line. However, in contrast to the data with respect to rhos, those who received ten days of training tended to achieve the best reporting.

With respect to supervision, there is not a significant positive effect of taping overall because of a complex interaction with training. For three of the groups, those who received two days, five days or ten days of training, there was a positive influence of supervision on interviewer performance. The patterns are roughly monotonic; interviewers who were tape recorded tended to perform better.

However, for interviewers who received only one day of training, tape recording may have had a negative effect on the amount of bias in answers. Those who were tape recorded appeared to perform worse, in the sense that they obtained more biased data, than interviewers with one day of training who were not tape recorded.

Finally, it should be noted that those who received the full treatment - 10 days of training plus tape recording - appeared to significantly out perform all other interviewer groups.

DISCUSSION

The most important implication of the analysis presented thus far is that the level of training and supervision does matter. How interviewers are trained and supervised has significant potential to effect the extent to which interviewers effect the answers they get, as well as the potential to reduce bias in the answers that interviewers obtain. A corollary is that what we try to teach interviewers to do reduces error if we can get them to do it. The data also points to the importance of the combination of supervision and trainning in affecting interviewer performance.

Moreover, it is important to understand that the order of magnitude of the impacts observed here are very substantial. For example, the range of rhos observed in Table 2 were from .006 to .020. If interviewers had an average assignment of 30 interviews and performed at the level of the best group, they would increase the standard error of

TABLE 2

Average Standard Score (x1000)* on Questions
Judged Most Likely Subject to Systematic Bias
By Level of Training and Supervision

	Supervision			
Training	Level 1	Level 2	Level 3	Average
1 day	18	11	-26	1
2 days	7	-5	43	15
5 days	-8	-5	7	-2
10 days	0	21	52	27
Average	4	5	20	10

* A positive score is a score judged to be less biased. Analysis of variance results: Training effect ($F = 2.44$, $df = 3$, $p > .10$); supervision effect ($F = 1.36$, $df = 2$, $p < .10$); interaction ($F = 2.64$; $df = 6$; $p < .05$).

TABLE 3

Contrasts of Average Standard Scores
on Questions Judged Most Likely Subject to Systematic
Bias for Selected Training and Supervison Groups

Comparisons	Effect Coefficient	Adjusted* Standard Error	Value of t
Training Groups			
10 days vs. Less (All supervision levels)	.064	.051	1.25
Supervision Groups			
Taped vs. not (All training groups)	.015	.016	1.0
Taped vs. not (Exclude 1-day training)	.099	.056	1.77
Taped vs. not (1-day training only)	.04	.03	1.33
Training-Supervision			
10 days training and taped vs. rest	.50	.26	1.92

* Standard errors were multiplied by $1 + p(b-1)$, where p = the intraclass correlation, b = the number of questions per interviewer, to take into account the fact that there were multiple measures per interviewer that were not independent.

designs which had no interviewer effects. In contrast, if interviewers with the same size estimates by less than 10 percent over assignment performed in the manner of the worst group, the standard errors would be more than 25 percent higher then for a simple random sample with no interviewer effects.

The data point to the value of close supervision of field interviewers, particularly the value of tape recording and supervising the question and answer process. Although it is somewhat puzzling why the patterns are not completely consistent for all groups, it is quite striking that none of the groups that were tape recorded were very different from the very best group. On the other hand, those interviewers who did not receive systematic feedback about their interviewing were consistently worse than the average. Although statistical significance of taping did not show up in all the data presented here, the persistance of findings in these and other analyses that taping is positive has convinced us that the value of taping will be well supported when our data analyses are complete.

The one negative indication is that taping may produce bias if interviewers are not well trained. Our hypothesis in this respect is that the tape recorder may inhibit respondents if interviewers do not handle the situation comfortably and professionally. We are carrying out further analyses to explore that hypothesis.

Perhaps the most interesting data pertain to the effect of increasingly extensive training on standardization. The interviewers who received ten days of training did learn how to ask questions more exactly and probe non-directively better than others; we have tests to show that. However, those who were not tape recorded produced higher values of rho than they should have.

To understand this finding, it is worthwhile recalling what rho measures: it measures the consistency or comparability of interviewers. We have theorized that interviewers are subject to two sets of pressures: those from the researcher and those from the respondent. A key feature of untaped or unobserved field interviews is that the quality of interviewer performance in the question and answer process is unmonitored. Interviewers can do what they want. Our working hypothesis is that our 10-day training may have exacerbated between interviewer differences because the conscientious interviewers could perform so well (as indicated by the data on reduced bias) while others changed their behavior from what they learned in training to respond to situational pressures when they were not tape recorded. In this way, we think the potential for variable interviewer performance may be enhanced by extensive training when there is inadequate supervision.

Fortunately, assessment of the validity of these and other hypotheses will not rest solely on the data presented here. The project involved reinterviewing respondents, tests of interviewers and analysis of taped interviews. These additional data will help us to understand why the relationships observed occur, as well as helping to strengthen the power of our analyses where our current tests provide suggestive but not convincing evidence. However, the analysis presented here provides important evidence on two unsurprising but heretofore undocumented points:

1) The training and supervision interviewers receive singly and, most important, in combination do affect the error in survey data.

2) If we can understand how to train and supervise effectively, there is potential for important improvements in the precision of survey based estimates.

REFERENCES

Bailar, B. A., Bailey, L., & Stevens, J. "Measures of interviewer bias and variance", Journal of Marketing Research, 1977, 14, 337-343.

Cannell, C. F. and Fowler, F. J. A note on interviewer effect in self-enumerative proceudres, American Sociological Review, 1964, 29, 276.

Cannell, C. F., et al. A summary of studies. In Vital and Health Statistics, Series 2, No. 69, 1977, Washington, D.C.: U.S. Government Printing Office.

Freeman, J., and Butler, E. W. Some sources of interviewer variance in surveys. Public Opinion Quarterly, 1976, 40, 79-91.

Groves, R. M., and Kahn, R. L. Surveys by Telephone, New York. Academic Press, 1979.

Groves, R. M., and Magilavy, L. J. Estimates of interviewer variance in telephone surveys. Paper delivered to American Statistical Association, 1980.

Hanson, R. N., and Marks, E. S. Influence of the interviewer on the accuracy of survey results. Journal of American Statistical Association, 1958, 53, 635-655.

Kish, L. Studies of interviewers variance for attitudinal variables. Journal of the American Statistical Association, 1962, 57, 92-115.