INTERVIEWER VARIANCE IN TWO TELEPHONE SURVEYS

J. Michael Brick, Richard McGuinness, Susan J. Lapham, Margaret Cahalan, Dedrick Owens, and Lucinda Gray
Contact Person: Richard McGuinness, U.S. Bureau of the Census, Washington, DC 20233

Key Words: Behavior Coding, Interviewer Effects, Intra-class Correlation, Measurement Error

Introduction

Nonsampling errors are often acknowledged as an important contributor to errors in estimates from surveys, but measures of the size and direction for important sources of nonsampling errors are seldom produced. In the 1993 National Survey of College Graduates (NSCG) and the National Survey of Recent College Graduates (NSRCG), the contribution of interviewers to nonsampling error was an important source of error that was directly evaluated for these telephone surveys. The NSCG used a mixed mode of data collection, but the results reported here only pertain to the collection done by telephone. The NSRCG was conducted entirely by telephone.

Nonsampling errors due to interviewers, sometimes called interviewer effects, arise because interviewers may not always conduct the interviews in exactly the same way and these differences may impact on the respondents' answers. Interviewer effects are important in these surveys because they can be used to inform decisions about procedures to be used in future surveys and to improve inferences made from the surveys already conducted.

The first important role of measuring interviewer effects is to help improve question wording and interviewer training. Concerning question wording, if an item is well-constructed and understood by all interviewers and respondents, then it is unlikely that interviewer effects will be large. However, for items that are not well-constructed, estimates of the size of interviewer effects will provide feedback to survey designers about which items need improvement in how the questions are asked. In addition, estimates of interviewer effects may also be used to support revisions in other survey procedures, such as the training of interviewers.

The second important role of measuring interviewer effects in these surveys is to obtain better estimates of the precision of the estimates from the surveys. Even if the interviewer effects or systematic biases associated with interviewers cancel out when aggregated across interviewers, the variances of the estimates are larger because the differences are an additional source of variation. Ordinary estimates of the variance of an estimate do not account for interviewer variability. As a result, the precision of the estimates is often overestimated and confidence intervals and tests of significance may be lower than the nominal significance level. If interviewer effects are estimated, the estimates of the standard errors of the estimates can be adjusted so that inferences are closer to the nominal level.

The interviewer effects are best estimated by using an interpenetrating sample design in which respondents are randomly assigned to the interviewers. Even though an interpenetrating design was not used in either the NSCG or the NSRCG because of its data collection cost, both the Census Bureau and Westat assumed that assignment of the respondents to the interviewers was random after taking steps to eliminate from the analysis those cases that clearly violated the assumption.

The analyses from the two surveys generally identified the same items as having large interviewer effects, but the estimated sizes of the effects were different. Thus, for the objective of redesigning the specific wording of items, the results were consistent. However, the findings on the size of the effects raise other questions.

An important question was the extent to which the different models used to estimate the interviewer effects led to different estimates. The fourth section of this paper examines this question by applying both models to both surveys. These results indicate that the differences in the size of the interviewer effects is not totally due to the models and that other factors must be considered.

A different view of the interviewer effects is presented in the fifth section by examining the relationship of the size of the interviewer effects to measures from a behavioral coding of the interviews. A small number of interviews in the NSRCG were tape-recorded and coded using a behavioral coding scheme. The analysis indicates that the results from behavioral coding may be measuring a different aspect of the interviewing process than simple interviewer effects. This may have implications for assessing the quality of interviews.

NSRCG Interviewer Variance Study

The NSRCG had a two-stage sample. In the first stage a sample of institutions offering bachelor's and master's degrees in science and engineering was selected. A sample of bachelor's and master's degree

recipients in science and engineering for three cohorts (graduation years 1990, 1991, and 1992) was selected from the sampled schools. After an advance letter, the sampled students were then interviewed by telephone using CATI from one of two central telephone centers at Westat.

The data used for the analysis of interviewer effects in the NSRCG included interviews completed by both bachelor's and master's degree recipients for all three cohorts. Interviews assigned to specific interviewers or groups of interviewers with special training or skills were deleted from this analysis. Also, cases missing certain key information were dropped. The data set used in the analysis contained about 18,000 completed interviews. If a case was missing a specific item, then the case was eliminated for that particular analysis, but not for the analysis of other items.

For the NSRCG, cases were assigned systematically to the next available interviewer according to a priority scheme that was independent of the interviewer. In other words, the scheduling might have depended upon the calling history of the case, but the characteristics of the interviewer were not used in the assignment procedure. In all, 215 responses were analyzed.

One method to estimate the interviewer component of variance is to use an ANOVA model as suggested by Kish (1962). One of the problems with that approach for the NSRCG is the lack of full randomization of the cases assigned to the interviewers. To account for non-random factors the ANOVA model was revised to include fixed effects. The resulting mixed model can be written as

$$y_{ij(k)} = \alpha_{(k)} + \beta_j + \tau_{ij(k)}$$

where the α term is a general fixed effect, k is a subscript for the fixed effects, and β is the random error associated with interviewer j. (This terminology is used to avoid writing each fixed effect and is appropriate because the estimates of the specific fixed effects are not important in this context.) The error term (τ) accounts for all the deviations from the fixed and random effects in the model. Despite the fact that this is a mixed model, we will refer to it as the ANOVA in much of the subsequent discussion.

Since the weights of the graduates were highly variable, degree and major field were included as fixed effects in addition to the following features associated with the interviewing process: telephone center location, season, respondent's time of day of interview, and respondent's time zone.

The VARCOMP procedure in SAS was used to implement the estimation of the model's parameters using a restricted maximum likelihood method of estimation. The ratio of the estimated variance component for the random interviewer effect and for the error was defined as the intra-interviewer correlation coefficient. For the analysis, all of the variables with response categories were structured so that they were dichotomous. This structure raises concerns about the model assumptions of the homogeneity of the variance and the normality of the effects. As a result of this concern, estimates close to zero or 100 percent may not be well-suited to be estimated by the procedures employed. The same restriction also applies to the normality assumption.

Findings from the NSRCG Study The intra-interviewer correlations across nearly all the questions examined were very small. The median correlation was 0.002 and the mean correlation was 0.007. The mean was much larger than the median due to a few items with very large correlations. To assess the statistical significance of the correlations, the estimates were compared to critical values of ρ: 0.03 for questions asked in about 700 interviews and 0.001 for questions asked in all the interviews.

Another measure of statistical significance that is relatively constant across the different sample sizes is $1+\rho(m-1)$, where ρ is the intra-interviewer correlation and m is the average number of interviews completed by the interviewers. As is shown in Brick (1994), when the value of this factor, which we refer to as the variance inflation factor, exceeds about 1.2 (a 20 percent increase in the variance due to the interviewer effect) then the interviewer effect is statistically significant for this problem. With 215 items being examined, we would expect about 5 percent of them to be statistically significant under the null hypothesis that there is no interviewer effect. However, somewhat more than the expected number of effects were statistically significant.

The most common feature of the items with higher correlations was that they were primarily open-ended questions that the interviewer had to code the respondent's replies into one or more categories. The interviewer was required to record all the responses given. In addition, many of the items with larger correlations were often only asked of subsets of the graduates. In general, the impact on the standard errors of the estimates are substantially reduced when the items are asked for a subset of the respondents because the value of m is much smaller.

NSCG Interviewer Variance Study

The NSCG is a national survey conducted every two years by the U.S. Bureau of the Census. Using a list compiled from the 1990 decennial census, the Census Bureau randomly sampled from a list of persons who

had received at least a bachelor's degree in 1990 or before. The Census Bureau conducted a study in order to identify which NSCG response categories had high interviewer effects, and identification of problem categories produced recommendations for questionnaire rewording and further training.

About 23,000 completed telephone center (CATI) cases were used for this analysis, as it was expected that this type of case would have fewer uncontrolled effects than field cases. At the same time, because CATI interviewers were more closely monitored and supervised than field representatives, it was expected that interviewer effects for telephone center cases would provide a lower limit on interviewer effects for field cases. As with the NSRCG, cases were assigned systematically to the next available interviewer according to a priority scheme that was independent of the interviewer. However, the priority schemes were different for the two surveys.

Since the survey did not have an interpenetrating design, the completed CATI cases had a number of uncontrolled effects, including telephone center location, time of day, day of the week, time zone, phase of the survey, as well as possible mode interaction effects.

Questions that were considered likely to show interviewer effects were included for analysis. All of the questions were categorical, and for multiple response questions dichotomous responses were created for each category (i.e., whether the response was in or out of the category). A total of 29 questions, with 203 response categories, was chosen.

Interviewer effects were estimated using a method based on the beta-binomial distribution. In the model used, each interviewer's probability of success in asking a dichotomous question is itself considered a stochastic variable with a beta distribution. This method was used to deal with the homoscedasticity and normality problems that arise from employing the ANOVA method with categorical data. The beta-binomial model used for NSCG satisfies the above requirements for estimating variance components of proportions.

The chi-square type Wald statistic was used to test whether to accept the null hypothesis that the data fit the beta-binomial model. Given that the data passed the model test, a standard Z statistic was used to test for the presence of significant interviewer effects.

In the NSCG study, the percent increase in the standard error of the estimate caused by interviewer error—called the standard error inflation factor—was used as a measure of the interviewer effect. This percent increase is

$$100[1+(m-1)\rho]^{1/2}-100.$$

Since the beta-binomial model did not include fixed effects, the NSCG analysis could not remove the effects of such factors as telephone center and time of day. These effects probably increased estimates of interviewer effects.

Weighted data were not used in the analysis. Although it is possible that unweighted data could underestimate interviewer effects, it was thought that this would not affect the determination of which questions had high between-interviewer variance.

Findings from the NSCG Study Twelve questions displayed large interviewer effects. For these questions, it was found that 26 out of 180 categories had large values of ρ or substantial standard error inflation factors or both. Twenty-three of the original 203 response categories failed the beta-binomial model test. The requirement for a large value of ρ was that it be greater than 0.015. A substantial standard error inflation factor was defined to be greater than 20 percent.

Evaluation of Differences between the Surveys

The mean and median estimated values of ρ reported for the two surveys were quite different despite the fact that the items and responses identified as having larger values were relatively consistent. The mean and median estimates of ρ reported in the NSRCG were consistently lower than those in the NSCG. These differences pose some interesting questions about whether the differences might be explained by the different models or whether there are other factors that might be reflected in the estimates of interviewer variance. This section attempts to explore the differences and some of the potential reasons they have occurred. We begin by selecting a subset of items for the analysis and then apply the beta-binomial and analysis of variance with fixed effect models to each data set to obtain estimates of correlation. (See Table 1 for differences in the designs of the two surveys.)

Methods A subset of 11 questions that were asked in both studies were selected and then restructured, producing 61 dichotomous items for analysis. The beta-binomial model and the ANOVA models described in the previous section were then fitted to these items, resulting in four estimates of ρ for each question (one for each survey and model combination. The same fixed effects were included in the ANOVA part of the combined study model.

Results of the Evaluation Summary statistics for the values of the estimated ρ's under the two models and for each survey are shown in Table 2. We evaluated the estimated median correlations to avoid some of the problems associated with outlying estimates. Using the median as a measure, the estimated correlations from the

NSCG appear to be about three to six times larger than those of the NSRCG, depending upon the model. Another way of looking at the size of the estimated correlations is by counting the number of estimates that are greater than or equal to 0.01. The number of items with ρ's greater than or equal to 0.01 is about four times larger for the NSCG than for the NSRCG. This suggests that the models used to estimate the correlations were not primarily responsible for the differences in the magnitude of the estimates.

To more directly examine the differences in the estimates from the two models, summary statistics for the beta-binomial and the ANOVA model estimates for the same survey in Table 2 can be compared. In general, the mean and median values for the beta-binomial model are larger than the estimates for the ANOVA model, but the magnitude of the difference depends on the study. Another way of evaluating the differences is to compare the estimates for the same item for the two models (after first converting the beta-binomial negative estimates to zero, equivalent to the conversion for the ANOVA model). For the NSRCG, the beta-binomial model estimates for 65 percent of the items were larger than the ANOVA estimates, while for the NSCG the beta-binomial estimates were larger than the ANOVA model estimates for 93 percent of the items. The larger estimates from the beta-binomial model are probably the result of excluding the fixed effects that were included in the ANOVA model.

Other Factors That Might Explain Differences The differences in the magnitude of the estimated correlations for the two studies remain after accounting for the different models. Thus, other factors that differ for the two studies and that might account for different interviewer effects must be considered. Below, we address three potential factors: differences in respondent populations, differences in training, and differences in administrative methods.

Respondent populations One of the major differences between the NSRCG and the NSCG that could impact on the interviewer effects was in the target populations. The NSRCG included recent graduates—those who graduated from college in 1991, 1992, or 1993. The NSCG included persons who graduated from college in 1990 and before, with an upper age limit of 75 years.

To examine the age difference, the beta-binomial model was fit to the subgroup of NSCG respondents age 30 and under. The estimates from the NSCG respondents age 30 were consistent with the full NSCG sample. This suggests that the age of the respondents (and the other items that vary with age) does not appear to account for the differences between the surveys.

Another major difference in the two surveys was that the NSRCG was done completely by telephone while the NSCG used a mixed mode. Only those cases that did not respond to the mail were sent to CATI in the NSCG, and this difference could have influenced the estimates. The beta-binomial model was applied to 2,955 NSCG cases that were part of a mode study that went straight to CATI. The estimates of ρ's for this subgroup were consistent with the full sample rather than being like the NSRCG. Thus, no evidence exists that the population differences were a major factor in the differences in the size of the estimates.

Training procedures In addition to informing interviewers about the specific procedures for the survey, a goal of training is to reduce the amount of variation that interviewers bring to the study. Interviewers are generally instructed to follow the procedures and read the questions as written without leading the respondents. The amount of training that interviewers received was very different for the two studies. Interviewers for the NSRCG received 4 hours of general telephone training (new interviewers) and about 16 hours of survey-specific training (all interviewers). Interviewers for the NSCG received 4 hours of general telephone training and about 4 hours of survey-specific training. Thus, the interviewers for the NSRCG had significantly more training in the specific survey than the NSCG. This difference could account for some of the difference, but there is no way to test this hypothesis because the interviewer training procedures were consistent within survey.

Administrative procedures Training is only the most obvious of a set of administrative procedures that were different in the two surveys. Westat and the Census Bureau did not strive to make their environments and procedures similar for these studies, so "house effects" are possible. It is possible that the interviewers recruited by the organizations are different (e.g., there may be differences in education levels and incomes). Other "house effects" such as differences in supervision and monitoring might also be present. These other administrative procedures, like the training differences noted above, cannot be evaluated.

Behavior Coding Study

In addition to the interviewer effects study, a study of the interaction between interviewers and respondents was conducted in the NSRCG. This was a small study involving only 19 of the 105 interviewers employed in the NSRCG. Approximately 100 interviews were tape-recorded with the permission of the respondent. The taped interviews were then coded to assess 8 different elements of the behavior of the interviewers (4 of the elements) and the respondents (the other 4 elements

coded). The full use of these data were included in a report on the study that is available from the NSF (Gray et al., 1994).

The behavior coding study was examined in conjunction with the interviewer effects study to examine if the types of effects noted in the interviewer study were the same as those found in the behavior coding study. In particular, two hypotheses were posed: 1) Are large interviewer effects (ρ) correlated with items that require a great deal of probing by the interviewers or items for which the respondents ask the interviewers to clarify; and 2) Do items that require a more than average amount of probing or clarification lead to large interviewer effects?

A simple metric from the behavior study was used to examine these hypotheses. Most of the items in the survey were asked by the interviewers and then answered by the respondent without any additional discussion. This type of interchange was coded as asked and answered (A&A) in the behavior coding of the interviews.

In general, the relationship between the behavior coding (as measured by the A&A metric) and the interviewer effects was not apparent from this study. The estimated correlation between the proportion A&A and the interviewer effects was negative, as postulated, but small (-0.08). Measured a different way, of the 16 items with statistically significant interviewer effects, 10 (or 62.5 percent) of the items were A&A at least 85 percent of the time. About 67 percent of all the items were A&A at least 85 percent of the time. Thus, the items with significant interviewer effects were not more likely to have low A&A percents.

Looking at the converse, again the relationship was not apparent. Of the 9 items with values of A&A of less than 75 percent (those requiring more than average probing and clarification), only 5 had statistically significant interviewer effects. This is not statistically different from the entire set of items, where the set of items was restricted to those items that had 10 or more responses in the behavior coding study.

Thus, the results from this small evaluation suggest that the relationship between the interviewer effects as measured in the NSRCG and the behavior coding are not highly correlated. The two types of studies may be measuring different phenomena and have very different implications for questionnaire design.

Conclusion

The contribution of interviewers to nonsampling error was an important source of error that was directly evaluated for the NSCG and the NSRCG. The NSCG and NSRCG studies used similar questionnaires, with many of the items being identical in both surveys.

Despite differences in the models used in the two studies, the analyses generally identified the same items as having large interviewer effects. Thus, for the purpose of revising the questionnaire for future surveys, the two methodological studies gave consistent results.

The size of the interviewer effects, as measured by the correlation, did differ by survey, with the estimated correlations from the NSCG being generally larger than the estimates from the NSRCG. (Nonparametric sign tests were used to evaluate whether sets of correlations were the same.) In this analysis, the NSCG estimates were found to be generally larger than those from the NSRCG for each model, indicating that the differences in the size of the interviewer effects was not primarily due to the models. The correlations estimated from the beta-binomial model were also generally larger than those from the mixed model, probably because the mixed model contained fixed effects to try to control for the lack of randomization. However, these differences appeared to be much less than those between the surveys.

Other factors that might affect the size of the interviewer effects were also considered. The study populations were different, with the NSCG respondents older than the NSRCG respondents. Another difference was that the NSCG respondents were typically already nonrespondents to the mailed question, while the NSRCG respondents were not asked to respond by mail. By restricting the analysis to subsets of the NSCG respondents, no evidence was found that these factors were related to larger interviewer effects.

Additional factors that could not be empirically evaluated might also have influenced the size of the interviewer effects. One major difference in the surveys was the time allocated to training. The NSCG used many fewer hours of training than the NSRCG. Other "house effects" associated with the administrative procedures used by the Census Bureau and Westat for their telephone surveys could have affected the estimates, but these are, like the training, not measurable.

In addition to the comparative analysis of the two surveys, this paper also discusses an attempt to relate interviewer effects to measures from a behavioral coding of the interviews. A small number of interviews in the NSRCG were tape-recorded and coded using a behavioral coding scheme. The lack of a consistent relationship between the behavioral coding and the interviewer effects suggests the two methodologies may be measuring different aspects of the interviewing process.

References

Brick, M., "1993 National Survey of Recent College Graduates: Interviewer Variance Study," preliminary draft of a report to the National Science Foundation, October 1994.

Gray, C., Cahalan, M., and Chen, S., "Recorded Interviewer Behavior Coding Report," preliminary draft of a report to the National Science Foundation, May 1994.

Kish, L., "Studies of Interviewer Variance for Attitudinal Variables," *Journal of the American Statistical Association*, Vol. 57, No. 297, 1962, pp. 92-115.

Lapham, S., "Inch by Inch, Rho by Rho: An Attempt to Understand Interviewer Variance and Interviewer Variance Models," a paper written for the Joint Statistical Methodology Program at the University of Maryland, December 7, 1994.

Pannekoek, J., "Interviewer Variance in a Telephone Survey," *Journal of Official Statistics*, Vol. 4, 1988, pp. 375-384.

Ringstrom, D., Owens, D., and McGuinness, R., "Interviewer Variance in the 1993 National Survey of College Graduates," preliminary draft of a paper submitted to the National Science Foundation, November 1994.

Stokes, S.L., and Hill, J.R., "Modelling Interviewer Variability for Dichotomous Variables," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1985, pp. 344-348.

Table 1. Descriptions of the NSRCG and the NSCG

Study Description	NSRCG	NSCG
Frame	Administrative records from universities of 1991, 1992, and 1993 college graduates	1990 decennial census list of college graduates—CATI interview sample
Mode	CATI	Mail, CATI, personal interview
Target population	Science and engineering college graduates who received their degree in 1991, 1992, or 1993	Science and engineering college graduates under 75 years who received their degree prior to 1991
Field period	April-December 1993	April-December 1993
Response rate	85.0%*	79.0%*
Number of interviewers	105	316
Number of completed interviews	17,586	22,960
Average work load per interviewer	185.1	71.6
Range of completed interviews	2-619	1-272

* Noncontacts (those without telephones, no bachelor's degree, deceased, etc.) included in denominator.

Table 2. Summary Statistics for Intra-interviewer Correlation, By Estimator and Survey

Model	NSRCG		NSCG	
	Beta-binomial	ANOVA	Beta-binomial	ANOVA
Number of variables analyzed	61	61	61	61
Median ρ	0.00136	0.00061	0.00480	0.00369
Mean ρ	0.00209	0.00172	0.01217	0.00824
Number of ρ's 0.01 or greater	5	3	21	11