HOW INTERVIEW MODE AFFECTS DATA RELIABILITY

John M. Bushery, J. Michael Brick, Jacqueline Severynse, Richard A. McGuinness
John M. Bushery, U.S. Bureau of the Census, Washington, D.C. 20233 jbushery@ccmail.census.gov

Key Words: Data reliability, Response error, Interview mode, Reinterview

1. INTRODUCTION

When Dillman (1978) addressed the question "Which mode is better?" he considered many criteria, which can be grouped under cost, quality, and timeliness. Computer assisted interviewing technologies, such as CATI (computer assisted telephone interviewing), have changed some of the relative advantages among interview modes, but the basic principles Dillman considered still apply. Costs and timeliness are relatively easy to compare between modes. Evaluating quality takes more effort, but is crucial. Using a mode that compromises data quality may negate any advantages in cost or timeliness.

Completeness, accuracy (or validity), and reliability are all components of "quality." Completeness encompasses an unbiased sampling frame, good response rates, and low item nonresponse. Accuracy refers to unbiased responses to survey questions. Reliable responses are those that a respondent can provide consistently whenever asked the question.

Bishop (1988) hypothesized that respondents to self-administered questionnaires might provide better quality data because they have more time to think when answering the questions. Some researchers believe respondents are unable to grasp more than four or five response categories when questions and answer categories must be read aloud to them (Dillman 1978).

Research to date has produced mixed results concerning the quality of mail and interviewer-administered surveys. Mail surveys generally suffer from lower completeness. But they seem to have an advantage in accuracy for certain types of data. Interviews conducted with greater anonymity often have smaller social desirability effects (De Leeuw 1987). Krysan (1994) and Siemiatycki (1979) observed this phenomenon when comparing mail with interviewer-administered surveys. Cook (1993) and McHorney (1994) observed higher (and presumably more accurate) levels of reporting in mail surveys than in interviewer-administered surveys for health conditions. On the other hand, O'Toole (1986) found more underreporting of health conditions in the mail mode.

Administrative records, if available, can be used to evaluate the accuracy of survey data. Körmendi (1988) used income tax records to compare the quality of income data obtained in telephone and personal interviews. De Leeuw's meta-analysis discussed 28 studies devoted to mode comparisons between 1979 and 1986. About one-third of those studies addressed the issue of data quality by using administrative record checks. Siemiatycki used government health insurance records to validate responses to a household health survey in Montreal. Unfortunately, administrative records are seldom available to evaluate survey data quality. The reinterview study provides a relatively easy method to measure data reliability. Reinterview studies measure test-retest reliability (also called **simple response variance**). Few mode studies address reliability, particularly via a reinterview.

Cook used a reinterview, but did not address data reliability. McHorney did not use a reinterview, but found neither mail nor telephone interviews to have an advantage in terms of internal consistency reliability. O'Toole found few differences in the reliability of medical questions under three data collection modes: mail, telephone, and personal visit. That study did not meet all the requirements of the response variance model because the reinterviews for all modes were conducted in person. It also used simple agreement rates as the reliability estimator.

Besides allowing respondents time for more thoughtful answers, mail surveys have another advantage in reliability -- the correlated component of response error, or between-interviewer variance, is eliminated for cases completed by mail. This advantage played a major role in the Census Bureau's decision to conduct the 1960 and subsequent decennial enumerations by mail (U.S. Census Bureau 1985). Brick et al. (1995) discuss between-interviewer variance in the CATI portion of the 1993 National Survey of College Graduates (NSCG) and the 1993 National Survey of Recent College Graduates (NSRCG).

Computer assisted telephone interviews (CATI) have their own advantages. An interviewer can explain unclear questions if the respondent is confused. The centralized CATI sites offer the advantage of closer supervision, monitoring, and standardization, which can reduce correlated response error relative to decentralized interviewer-administered surveys.

Although O'Toole and McHorney found no overwhelming reliability differences between mail and telephone interviewing, the reinterview of the 1991 Schools and Staffing Survey (Bushery 1992) suggested that mail interviews produce more reliable data than telephone interviews. This result is consistent with Bishop's hypothesis. Cases interviewed and reinterviewed by mail displayed lower response variance than cases interviewed and reinterviewed by telephone.

However, that comparison was not a controlled experiment, so no definite statements about relative quality could be made. Schools not returning the mail questionnaire were interviewed by telephone. Mail respondents may have been the type who put more effort into their answers, while those contacted by telephone may have been more likely to give "top of the head" answers, whatever the interview mode.

This paper compares the simple response variance of data obtained for identical questions in the mixed-mode mail/CATI 1993 NSCG and the all-CATI 1995 NSRCG. This analysis still falls short of a rigorously designed experiment, but it eliminates the confounding effect of respondent cooperation determining interview mode.

2. METHODOLOGY

2.1 The 1993 NSCG and Reinterview

The 1993 NSCG was a mixed-mode survey, consisting of mail, CATI, and as a last resort, personal visit paper-and-pencil interviews. The reinterview covered only the mail and CATI phases of the survey. To produce an accurate measure of response variance, the reinterview replicated the original interview mode. Respondents who returned a mail questionnaire were reinterviewed by mail. If these respondents did not return a mail reinterview, they were sent one follow-up reinterview questionnaire, but no telephone or personal reinterviews were attempted. Respondents originally interviewed by CATI were reinterviewed by CATI. All but 15 percent of the cases in this study were interviewed and reinterviewed by mail.

The 1993 NSCG sample of 216,000 persons was selected from long-form Census respondents who had obtained a bachelor's degree or higher. The reinterview was designed to obtain 200 to 300 complete reinterviews in each of four broad science occupation groups: Engineering, Physical Science, Math/Computer Science, and Social Science/Psychology. The reinterview sample also included about 250 nonscientist respondents. Ultimately 1,685 complete reinterviews were obtained from the 2,506 cases eligible for the reinterview, yielding an unweighted response rate of 67 percent (66 percent for mail reinterviews and 77 percent for CATI reinterviews). To improve comparability between the NSCG and the NSRCG, this analysis uses only the 1,437 reinterviews completed by respondents in the four science groups.

2.2 The 1995 NSRCG and Reinterview

The 1995 NSRCG was conducted virtually completely by CATI. The reinterview reasked every item from the original interview and was conducted entirely by CATI. The NSRCG reinterview results may reflect a recall effect, in addition to simple response variance, because some households were reinterviewed as much as four months after the original interview. Only the question *"Were you working?"* is likely to be affected by recall in this comparison.

The 1995 NSRCG was a two-stage sample in which institutions were selected at the first stage and graduates selected at the second stage. The NSRCG sampled from lists of bachelor's and master's degree graduates in 1993 and 1994 provided by the sampled institutions. A sample of 275 institutions and 21,000 graduates was selected at different sampling rates, depending on major field of study, degree, and race/ethnicity.

An equal probability sample of responding graduates was selected for the reinterview. Bachelor's and master's degree graduates from both 1993 and 1994 were eligible for the reinterview sample. Graduates who previously refused the original interview but later agreed to complete it, and those who completed the interview late in the field period were excluded from the reinterview sample. Of the 800 graduates sampled, 658 completed the reinterview for an unweighted response rate of 82 percent.

2.3 Comparability of the NSCG and NSRCG

Some differences between these surveys may confound this analysis. The two surveys differed in target population and survey administrative procedures. Because 85 percent of the NSCG data used in this study were collected by mail, the training and survey administration differences should affect this comparison negligibly. Brick et al. (1995) discusses the comparability of the 1993 NSCG and the 1993 NSRCG. Except for question revisions, the 1995 NSRCG is similar in design and implementation to the 1993 NSRCG. Two factors may affect this comparison. The order of questions differs between the NSCG and NSRCG, and the NSRCG may have additional recall effects in the reinterview.

Differences in the survey populations also may affect these comparisons. The NSCG targeted all people holding a bachelor's degree or higher at the time of the 1990 Census. The NSRCG targeted only "recent" recipients of bachelor's or master's degrees in the fields of science and engineering. The NSCG respondents tend to be older than those in the NSRCG, and they have a broader range of educational backgrounds and occupations. Finally, the 15 percentage point lower reinterview response rate in the NSCG may confound this comparison. Respondents who did not return a mail reinterview also may have been less likely to provide thoughtful answers. If so, the response variance estimates for the NSCG may be understated. Section 2.4 describes the steps taken in this analysis to reduce confounding differences between the two surveys.

2.4 Analytic Methods Used

To improve comparability, this study used 1993 NSCG data only from the science groups. This helped equalize the populations with respect to educational background and occupation. Age differences between respondents in the two surveys may be problematic, but eliminating NSCG respondents whose ages were outside the NSRCG's age distribution would have left too little NSCG sample to measure response variance. Instead, all response categories with significantly different estimates between the NSCG and NSRCG (alpha = 0.05) were deleted from the comparisons. When the two estimates of percent in category are similar, meaningful comparisons of the response variance measures can be made. The authors consider it very unlikely that age differences alone would cause significant differences in the response variance measures.

Comparing the quality of reporting in the two modes using the reinterviews from these surveys is a reasonable approach because of the similarities of the populations and the interviews. Since some of the questions changed after the 1993 NSCG (sometimes based on the findings of the reinterview), only those questions that were the same in both administrations are included in the mode comparisons below. All of the questions are treated as binary variables. For the "mark all that apply" questions, each possible response category is treated as a separate variable. Similarly, items with k response categories are treated as k-1 dummy binary variables.

After eliminating the revised questions and answer categories with different estimates between the two surveys, only eight of the 79 questions in the 1993 NSCG were eligible for the comparison. From these eight questions, only 24 distinct answer categories could be compared.

Two statistics that assess the reliability of reporting in surveys are used for this analysis: the gross difference rate and the index of inconsistency. The gross difference rate is the percentage of items with different responses in the two interviews and is used to estimate the consistency of reporting or the simple response variance. The index of inconsistency is a relative measure of response variability, and in some circumstances, it is a measure of the proportion of the total variability due to random response error. Forsman and Schreiner (1991) give a more detailed discussion of these statistics and their uses.

Because of the variable weighting in both original surveys and the oversampling in the NSCG reinterview, weights were used to compute these statistics. The weights are the weights of the sampled respondents multiplied by the weight associated with the reinterview subsampling, although no additional adjustments were made to account for reinterview nonresponse. All observations with missing responses to either the original or the reinterview were excluded from the analysis. Items with too few observations to estimate the index of inconsistency reliably also were excluded from the comparisons. Sampling errors of the estimates for both surveys were computed using replication methods and the WesVarPC software.

Table 1 shows the general format of the possible reporting outcomes from the original and reinterviews. Because of the differential weighting, the values in the cells are actually weighted sums of the number of cases rather than the raw number of cases. The statistics described above can be formulated using the cells of this table. The gross difference rate and index of inconsistency, expressed as percentages, are

$$gdr(\%) = 100\,(b+c) / n \text{ and}$$

$$index(\%) = gdr(\%) / (P_o(1-P_r) + P_r(1-P_o)),$$

where $P_o = (a+c) / n$ and $P_r = (a+b) / n$.

Table 1. General Format of Interview-Reinterview Results

	Number of cases in Original Interview		
Reinterview	With characteristic	Without characteristic	Total
With characteristic	a	b	a+b
Without characteristic	c	d	c+d
Total	a+c	b+d	n = a+b+c+d

The individual estimates of the index and the gross difference rate were compared using the Z-test. The Wilcoxon matched-pairs signed-rank test was used to obtain a rough idea of which mode produced more reliable data overall. Although response errors may be correlated among the different categories of a question, it is assumed such errors are not correlated between questions. The average of the single-category estimates of response variance within each question were computed and the Wilcoxon matched-pairs signed-ranks test applied to the resulting eight pairs of estimates. All comparisons were tested for significance at the 0.05 level.

3. RESULTS

Of the 24 response categories compared, only two displayed significantly different indexes of inconsistency between the NSCG and the NSRCG. Those questions

were *"Were you working ...? Yes"* and *"Activity with most hours? Basic research."* The NSCG's gross difference rate was lower for three response categories: the two response categories just mentioned and *"Work activities...? Basic research."* The NSRCG enjoyed a lower gross difference rate for *"Reasons took college courses? To facilitate occupation change."* Table 2 lists abbreviated text from the questions, estimates of the index and gross difference rate for these questions, and estimates of percent in category.

The Wilcoxon matched-pairs signed-rank test showed the NSCG produced lower average values of both the index of inconsistency ($Z = -2.24$, $P = 0.03$) and the gross difference rate ($Z = -1.96$, $P = 0.05$) for the eight distinct questions. This result suggests that data are slightly more reliable under the primarily mail mode NSCG than the all-CATI NSRCG.

4. CONCLUSIONS

Despite the confounding factors, this study provides evidence that a mixed mode mail-CATI survey has a slight edge in reliability over an all-CATI survey. To learn more about the size of this advantage and to determine which questions work better with mail and which with CATI, a carefully controlled experiment would prove useful.

Finally, although data reliability is important, it is only one factor to consider in selecting an interview mode. Tradeoffs must be made among costs, timeliness, reliability, and other aspects of survey performance. These results suggest that reliability concerns should not play the primary role in selecting interview mode.

References:

Biemer, Paul P., (1988) "Measuring data quality," Telephone Survey Methodology, Groves, R. et al, editors, John Wiley & Sons, New York, pp 273-282.

Bishop, George F., Hippler, H., Schwarz, N., Strack, F., (1988) "A comparison of response effects in self-administered and telephone surveys," Telephone Survey Methodology, Groves, R. et al, editors, John Wiley & Sons, New York, pp 321-340.

Brick, J. Michael, McGuinness, R., Lapham, S., Cahalan, M., Owens, D., (1995) "Interviewer variance in two telephone surveys," American Statistical Association, Proceedings of the Section on Survey Research Methods, Orlando, pp 447-452.

Bushery, John M., Royce, D., Kasprzyk, D., (1992) "The Schools and Staffing Survey: How reinterview measures data quality," American Statistical Association, Proceedings of the Section on Survey Research Methods, Boston, pp 458-463.

Cook, Deborah J., Guyatt, G., Juniper, E., Griffith, L., McIlroy, W., Willan, A., Jaeschke, R., Epstein, R., (1993) "Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma," Journal of Clinical Epidemiology, Vol 46, No 6, pp 529-534.

de Leeuw, E. D., (1988) "Data quality in telephone and face to face surveys: A comparative meta-analysis," Telephone Survey Methodology, Groves, R. et al, editors, John Wiley & Sons, New York, pp 283-299.

Dillman, Don A., (1978) Mail and Telephone Surveys: the total design method, John Wiley & Sons, New York.

Dillman, Don A. and Tarnai, J., (1991) "Mode effects of cognitively designed recall questions: a comparison of answers to telephone and mail surveys," Measurement Errors in Surveys, Biemer, P. et al, editors, John Wiley & Sons, New York, pp 73-93.

Forsman, Gösta and Schreiner, I., (1991) "The design and analysis of reinterview: An overview," Measurement Errors in Surveys, Biemer, P. et al, editors, John Wiley & Sons, New York, pp 279-301.

Körmendi, Eszter, (1988) "The quality of income information in telephone and face to face surveys," Telephone Survey Methodology, Groves, R. et al, editors, John Wiley & Sons, New York, pp 341-356.

Krysan, Maria, Schuman, H., Scott, L. J., Beatty, P., (1994) "Response rates and response content in mail versus face-to-face surveys," Public Opinion Quarterly, Vol 58, pp 381-399.

McHorney, Colleen A., Kosinski, M., Ware, J. E., (1994) "Comparison of the costs and quality of norms for the SF-36 Health Survey collected by mail versus telephone interview: Results from a national survey," Medical Care, Vol 32, No 6, pp 551-567.

O'Toole, Brian I., Battistutta, D., Long, A., Crouch, K., (1986) "A comparison of costs and data quality of three health survey methods: Mail, telephone and personal home interview," American Journal of Epidemiology, Vol 124, No 2, pp 317-328.

Siemiatycki, Jack, (1979) "A comparison of mail, telephone, and home interview strategies for household health surveys," American Journal of Public Health, Vol 69, No 3, pp 238-245.

U.S. Census Bureau, (1985) Evaluating Census of Population and Housing, Statistical Training Document ISP-TR-5, Washington, D.C., September, pp 96-97.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Acknowledgments: Agnes Colbert, Deborah Keane, Patrick Flanagan of the Census Bureau.

Table 2. Reliability Measures for Mail/CATI NSCG and CATI NSRCG

QUESTION	Index of Inconsistency			Gross Difference Rate			Percent In Category	
	NSCG	NSRCG	Z(dif)	NSCG	NSRCG	Z(dif)	NSCG	NSRCG
Were you working? *(Yes)*	4.3	20.0	-3.17	1.0	4.4	-3.40	86.6	86.9
Extent job related to degree?								
Somewhat related	37.0	37.1	-0.00	15.5	14.8	0.33	29.6	26.9
Factors for working outside degree								
Working conditions	33.3	40.2	-0.48	16.5	18.3	-0.24	41.1	33.2
Job location	35.6	63.4	-1.88	17.8	31.9	-1.88	52.6	45.0
Career change	36.1	52.3	-0.89	17.9	20.6	-0.35	41.3	31.8
Job in degree not available	44.5	27.2	1.24	21.7	13.6	1.18	36.0	49.1
Work activities on primary job								
Basic research	35.9	45.6	-1.51	10.7	15.3	-2.11	18.5	23.4
Employee relations	38.0	46.0	-1.32	16.4	19.6	-1.22	31.6	30.8
Teaching	38.4	28.4	1.49	10.3	9.4	0.42	16.7	21.8
Activity with most hours								
Applied research	43.6	48.3	-0.58	8.4	8.1	0.17	10.2	9.3
Basic research	37.9	75.2	-2.77	2.2	4.7	-2.72	2.9	3.6
Computer applications	29.1	34.1	-0.77	9.1	8.8	0.17	19.2	15.0
Development	59.6	78.3	-1.82	8.0	6.6	0.83	8.2	5.7
Quality management	50.3	63.8	-1.02	3.0	4.7	-1.42	3.8	4.1
Extent 2nd job related to degree								
Closely related	12.1	23.6	-0.86	6.1	11.3	-0.79	47.4	42.7
Not related	10.8	14.7	-0.52	5.0	7.4	-0.63	37.0	43.5
Reasons took college courses								
Prepare for graduate school	31.2	45.5	-1.58	10.5	16.0	-1.80	21.2	26.3
Facilitate occupation change	45.0	31.7	1.56	21.3	13.4	2.17	38.2	30.5
Further skills in field	40.8	43.7	-0.23	9.5	13.1	-1.06	84.3	82.3
Promotion/advancement	45.8	48.7	-0.35	18.8	23.2	-1.10	71.1	65.1
Required by employer	37.2	55.5	-1.78	12.4	16.6	-1.22	20.2	19.9
Leisure/personal interest	36.6	52.3	-1.84	18.2	26.3	-1.88	53.8	51.0
Citizenship								
Native born	2.2	0.00	1.74	0.5	0.0	1.77	88.3	89.7
Permanent visa	6.9	5.7	0.29	0.3	0.3	0.00	2.5	3.0