A Comparison of Recording Errors Between CATI and Paper-and-Pencil Data Collection Modes

James M. Lepkowski, Sally A. Sadosky, Mick P. Couper, Stephanie Chardoul, Lisa Carn, Lesli Jo Scott
Survey Research Center, Institute for Social Research, University of Michigan

Key Words: Question types, Coding Reliability

Anecdotal reports indicate that keying errors in computer assisted data entry may, for some types of data, be an important source of error in survey data. For example, at the Survey Research Center, University of Michigan, observations on the Panel Study of Income Dynamics (PSID) revealed that interviewers using computer assisted data entry make decimal location errors in recording dollar amounts for income, asset, and house value questions. Interviewers were observed, for instance, entering $10,000 when a $100,000 house value was reported. Such "order of magnitude errors" will seriously bias cross-sectional results and substantially increase measurement error in longitudinal analyses.

Recording errors may also have occurred during previous rounds of paper-and-pencil data collection. Thus, it is inadequate to assess only errors in one mode to assess the importance of recording errors in computer assisted data entry. To our knowledge, though, no study has explicitly compared recording or keying errors between computer assisted and paper-and-pencil modes.

Rustemayer (1977) examined recording errors in paper-and-pencil data collection, administering mock interviews to experienced, end-of-training, and new interviewers. She found that experienced interviewers made an entry inconsistent with respondent verbal reports in 4.5% of the entries. Kennedy, Lengacher, and Demerth (1990) studied keying errors in computer assisted interviews by monitoring computer assisted interviews across four studies. They found only 16 keying errors in 2,583 entries, or an error rate of 0.6%. Dielman and Couper (1995) used tape recordings for computer assisted interviews to assess keying error rates in 16,778 closed-ended questions from 116 interviewers. They observed an error rate less than 0.1%.

Besides keying or manual recording errors, Nicholls and Groves (1986) suggested that computer assisted interviewing would lead to fewer responses to open-ended questions. However, Catlin and Ingram (1988) found no difference in the length of open-ended responses (as measured by the number of characters or words) in a randomized comparison of computer assisted and paper-and-pencil modes. Bernard (1988) also found no difference in the length of open-ended responses when comparing a computer assisted interview with earlier data from paper-and-pencil interviews.

One area where differences have been found is in following skip logic and failing to complete items. Groves and Mathiowetz (1984) observed five times more skip errors for paper-and-pencil interviewers compared to computer assisted interviewers. Sebestik *et al.* (1988) reported that more than 90% of errors made by paper-and-pencil interviewers were failures to record an answer; computer assisted interviewers made no such errors. Olsen (1991) examined skip errors in the 12th round of the National Longitudinal Survey/Youth comparison of computer assisted and paper-and-pencil modes; about one in 100 skips were incorrectly followed in paper-and-pencil compared to none in the computer assisted mode. Tortora (1985) and Catlin and Ingram (1988) found fewer edit failures in computer assisted compared to paper-and-pencil mode, an indication that there were fewer skip logic errors or incomplete items.

We report on an experimental study imbedded in a national panel study of income dynamics. Data were collected from a subsample of 400 families using CATI or paper-and-pencil mode of data collection. All interviews were tape recorded and subsequently coded to determine if there were discrepancies between answers given by respondents and data recorded by interviewers.

Methods

The 1994 PSID was the 27th wave of a large-scale annual data collection project. Sample subjects were the original panel families who have been participating since the original 1968 wave and a Latino panel added in 1990. For the experiment, we used only the Core panel subjects. The principal data collection mode was computer assisted telephone interviewing (CATI) from a centralized facility, with a small number of face-to-face interviews by regional field interviewers. PSID staff trained approximately 150 interviewers (30 with 1-2 years experience on the project, the remainder newly hired) in 3-day training sessions on study specific concepts and procedures. (All newly hired interviewers first received extensive training in basic interviewing techniques.) Interviewers were trained only in CATI data collection. A total of 8665 interviews were completed from mid-March through December, 1994, with a response rate of 96%.

Interviewers were assigned to nine data collection "teams" (approximately eight interviewers on each) which shared sample and were led by a "specialist." Approximately six weeks into the data collection period, six interviewing teams were randomly selected for the mode experiment. In order to equalize training experience between CATI and paper-and-pencil groups, all interviewers in these six teams were trained to conduct

paper-and-pencil data collection. Interviewers received four hours of training that included a general overview of the experiment, basic paper-and-pencil data collection instructions, and a practice interview using the paper questionnaire.

Three of six groups were then randomly assigned to the paper-and-pencil data collection mode, and three to CATI. Each team was instructed to take interviews in the respective mode, using standard sample management procedures. All 596 PSID interviews conducted by these six teams during a three week period (June 5-28) were tape-recorded. After extensive review of each tape to identify technical problems (inaudible respondents, cut-offs, etc.), 200 tapes were selected randomly from each mode for coding and data analysis.

Only the first seven sections (A - G) containing the primary substance of the PSID interview and asked every year were coded. Section A covers basic housing information (size and type of home, amount paid for rent or mortgage, plans to move); Sections B-C collects current and previous year employment history for the head of the household (detailed description of job(s) held, wage rates, efforts to find other work); Sections D-E collect current and previous year employment history for the female spouse, if the head of the household is male; Section F covers food costs for the family (including receipt of Food Stamps); and Section G collects income for every member of the family (including wages and salaries, self-employment income, and all types of government assistance, obtaining the amount of income from each source).

Five experienced PSID interviewers and one who was only familiar with the PSID were trained for recording error coding. All coders attended an intensive training, consisting of four hours of lecture on principles of coding, hands-on practice, a practice case that each coder did individually (half of the case was coded as a paper-pencil interview and half as a CATI interview), and a follow-up session to discuss coder reliability from the practice case and to review general procedures. All coders handled both paper-and-pencil and CATI interviews, with assignments made to ensure that no coder received her/his own interviews. The coders listened to taped interviews, viewed the completed interview, and directly entered codes into a computer assisted system. Weekly meetings were held with the coders throughout the coding period to review procedures and examine various code frequencies to assess inter-coder reliability.

A simple, replicable coding system was devised to capture recording errors reliably across numerous question types (see the Appendix for a description of the codes). Questions were grouped into three types: closed-ended, open-ended, and checkpoints. (Checkpoints are not examined in this presentation.) In addition to standard closed-ended questions with fixed response sets, closed-ended questions in this investigation include short answer questions in which a one or two word response may be recorded, as well as amounts, counts, and time intervals during which income may be received. Open-ended questions are those which require more extensive answers.

For closed-ended questions, two types of interviewer-respondent exchanges were identified: simple straightforward and complex. A simple straightforward exchange is one in which the interviewer reads the question and the respondent gives an appropriate response option. A highly reliable coding of differences between respondent report and interviewer recorded data is readily made. In complex exchanges, interviewers probed a response or the respondent qualified an answer, gave an uncodable answer, or elaborated on an answer. This distinction of types of exchanges has not been made previously in the analysis of recording errors.

Close-ended simple straightforward answer recordings were classified as "accurate" or "inaccurate." Closed-ended questions with complex exchanges were classified as "accurate" or, because of the ambiguities in complex exchanges, "may be misrecorded." For open-ended questions, the recording was "verbatim," captured the "essence" of the answer, and "clearly misrepresented" the respondent's answer.

Closed- and open-ended questions were occasionally skipped during question reading. Interviewers may or may not recorded an answer for skipped questions. Most often question skipping occurred when the interviewer simply recorded an answer they believed the respondent gave at a previous question. From a standardized interviewing perspective, such events are errors since every question must be read to the respondent. Since the questions were not asked, coders could not ascertain whether a recording error occurred.

Coded data were carefully reviewed by study staff prior to data analysis. Each question was assigned a type and a response set. In addition, each error was classified into one of several broad categories to aid in understanding the nature of the errors observed. The question types, types of errors, and response sets are described in the Results section.

Analysis consisted of simple two-group comparisons between CATI and paper-and-pencil results. Statistical tests did not account for suspected effects of clustering of questions within interviews, interviewers, or coders. Inferences based on test statistics shown here must be adjusted for suspected losses in precision due to this clustering. The low frequencies of many events examined here, and cross-tabulations by interview and interviewer indicate that the effects of clustering are not severe.

Results

A total of 53,948 closed- and open-ended question askings were tape recorded and checked for accuracy across both modes for the 400 completed and coded interviews. A

total of 51,136 (94.8%) are closed-ended questions, consisting mainly of Yes/No responses, multiple choice categories, and short-answers. Most (73.8%) of the closed-ended question asking exchanges were simple and straightforward. Other closed-ended questions included skipped questions and those that the coder was unable to code due to tape or other difficulties. The open-ended questions mainly concerned occupation and industry and account for 2,395 question askings. There are no differences between modes in the relative frequency of these various types of questions and exchanges.

Closed-Ended Questions

Table 2 presents the frequency of recording errors for closed-ended questions. Only a handful of the simple and straightforward exchanges in each mode were recorded in error: 20 in CATI and 26 in paper-and-pencil. The difference between the rates in the two modes is not statistically significant. These 46 errors were principally of three types: (1) 16 or 35% were at Yes/No questions with the incorrect or opposite response recorded, or a "don't know" recorded when an answer was given; (2) 15 or 33% occurred when the interviewer incorrectly indicated the time reference, such as payment per month, day, or year; and (3) 11 or 24% the interviewer chose an incorrect response in a multiple response question. Eight of the 20 errors in CATI occurred at multiple choice questions, while 12 of the paper-and-pencil errors involved Yes/No questions. No interview had more than two of these recording errors, and no one of the 40 interviewers for whom interviews were coded had more than two of these types of recording errors. Thus, there does not appear to be any clustering by interview or interviewer. No question had more than one of these recording errors, and there was no difference in the error rates across sections of the questionnaire. As a result, no sequencing of errors was detected. Thus, it appears that the recording errors for closed-ended simple straightforward exchanges occur haphazardly at worst, and probably randomly.

A total of 785 or 6% of the closed-ended complex exchange question askings may have been misrecorded. The rate is significantly higher in paper-and-pencil (7.3%) than in CATI (5.0%). In order to characterize these errors more completely, they were classified as follows:

Implied response: interviewer inferred a response from respondent's answer when a valid response was not given. For example, interviewer assumed a year when the time period was not stated by the respondent.

Qualified response: interviewer failed to probe for a best estimate when the respondent qualified the answer. For example, respondent said "Somewhat more than $100," and the interviewer, without further probing, recorded $100.

Failure to record answer: interviewer left the response blank, marked two or more response options when only one was possible, or wrote illegibly.

Failure to repeat question: interviewer did not adequately probe or repeat question when required. For example, interviewer did not probe when a respondent laid off in February also reported that they were on vacation in that same month, and marked both laid off and on vacation.

Simple misrecording: during a complex exchange, interviewer entered incorrect response option. For example, interviewer marked "Yes" when respondent said "No."

Decimal shift: interviewer misplaced the decimal. For example, respondent reported $1,500 income from a source, but the interviewer recorded $150.

Numeric error: interviewer wrote incorrect numeric response (other than a decimal shift). For example, respondent reported $128, but the interviewer recorded $125.

Table 3 presents the frequency of these types of errors. The distributions for each mode are significantly different. While the implied response is the most frequent error in both modes, the relative frequency of implied responses is much higher in CATI (41.2% v 27.6%) primarily because there are many more failures to record errors in paper-and-pencil (123 in paper-and-pencil v only 17 in CATI). Further, paper-and-pencil had more than twice as many simple misrecording errors. Thus, the higher frequency of recording errors among complex exchanges in paper-and-pencil mode is due to a higher frequency of failure to record or simple recording errors.

These failure to record or simple misrecording errors may occur with higher frequency for certain types of questions. Each question in the PSID questionnaire was assigned a response set based on the nature of information recorded:

Short answer, simple: number or other information, other than dollar amount or time. *A16. How many rooms do you have, not counting bathrooms?*

Short answer, amount: dollar amount. *G17a. How much was from bonuses?*

Short answer, time: amount of time for specified time reference. *A26. How many years have you been paying on it?*

Multiple choice, simple: multiple responses. *A15. How is your home heated? With gas, electricity, oil, or what?*

Multiple choice, time: a time amount and period reference. *D99. About how much did she make at this? _____ per / 1. Hour / 2. Week / 3. Two-weeks / 4. Month / 5. Year /*

Multiple choice, month: all months that apply. *G42a. During which months of 1993 did you get this income? / Jan / Feb / Mar / . . . / Dec /*

Yes/No: yes or no response option. *A40. Do you have air conditioning?*

Table 4 presents the frequency of errors by mode across these response sets. Error rates are higher for all types of response sets in paper-and-pencil than in CATI. The difference in rates between the modes is statistically significant only for the two response sets involving time, due to the fact that they are the second and third most frequent response sets, after Yes/No questions. There is no one response set, though, where paper-and-pencil errors are concentrated.

There are only 6 interviews that have 10 or more errors in closed-ended complex question askings. While there were no interviews that were error free, the majority had four or fewer errors on complex exchanges, with 39% having only 1 or 2. At the question level, the only question which shows a significant difference between modes is B78: *Then, how many weeks did you actually work on your main job*? Recording errors occurred 24 times in paper-and-pencil at this question compared to just twice CATI. This is not surprising since in paper-and-pencil B78 requires the interviewer to use a marginal worksheet to calculate the number of weeks worked, sick, on vacation, unemployed, *etc.*, and then transfer the information to the response field, while in CATI the calculations are done automatically. Finally, there is no evidence that the recording errors for closed-ended complex exchanges clustered by interviewer.

Open-Ended Questions

Table 5 presents the results for open-ended questions. The number of recorded answers that clearly misrepresents the respondent's meaning is statistically higher in paper-and-pencil than CATI ($p<0.10$). However, paper-and-pencil also has a significantly higher proportion of verbatim recordings, and thus a lower rate of capturing merely the essence of the response. The "industry and occupation" questions, *"What kind of work did he/she usually do? What was his/her occupation?"* constitute 95% open-ended question asking errors. No clear clustering by interview or interviewer was observed.

Skipped Questions

There are two types of skipped questions for closed- and open-ended questions in Tables 2 and 5. For closed-ended questions in CATI, it is not possible to continue the interview without entering a response, and therefore skipped questions with no answer are impossible. The "Skips Question/No Answer Recorded" errors that do arise in CATI are structural in the occupation and industry questions. There two separate questions are asked, but they share a single response field. Interviewers are trained ask both questions, and record answers to both in the single field. When skipped questions errors occurred in CATI, interviewers only asked first question, and, assuming that they captured the respondent's most important activities and duties, skipped the second question entirely. This situation also occurs in paper-and-pencil, but far less frequently since these two questions do not share a single response field.

The "Skips Question/Records Answer" is a behavioral error in which the interviewer records an answer based on a response obtained at a previous question. It occurs more often in paper-and-pencil presumably because multiple questions are presented on the same page allowing interviewers to answer several questions at once. In CATI, since the interviewer is presented with each question one at a time, skipping a question but recording an answer is less likely to occur.

Discussion

The results of this investigation confirm findings of others: the frequency of recording errors for closed-ended questions in simple straightforward exchanges is extremely small and nearly random, and there is no difference between modes. Recording errors are much more common for closed-ended questions involving complex exchanges and for open-ended questions, and there are significant differences between modes. Paper-and-pencil has higher error rates among closed-ended complex exchange question askings, and higher rates of skipped questions with answer recorded. The higher rates in paper-and-pencil among complex question askings is due to failure to record an answer and simple misrecording errors. Recording errors occur at the same rate for both modes for open-ended questions, but there is a slightly higher frequency of clearly misrepresented answers in paper-and-pencil. At the same time, the paper-and-pencil mode also produces more verbatim recordings of answers, probably due to greater interviewer difficulty using a keyboard.

These results are tentative until a formal treatment of coder reliability and clustering of errors by coder can be completed. There is no evidence that errors clustered by interview or interviewer; only one question (B78) appears to have a larger than expected share of errors. Work continues to estimate coder reliability.

The kinds of observational data collected in this survey do not allow us to determine what types of errors occur at which steps. It appears that slips are infrequent in both modes, and not a fruitful area for further research. Laboratory-based studies are needed to determine the nature of the mistakes being made, where in the process they occur, and how to train interviewers to reduce their frequency.

Support for this investigation was received from NSF grant and NIA supplemental award no. SES 9022891, and the Survey Research Center, University of Michigan. The authors wish to thank Greg Duncan, Northwestern University, for stimulating this research and assistance acquiring funding; the Panel Study of Income Dynamics Board of Overseers for providing support; Sandy Hofferth, Frank Stafford, and Bill Shay for continuing support and advice during the research; Steve Blixt for assistance developing codes and computer applications and assistance training coders; Marshall Cummings for assistance with programming; Tom Gonzales for assistance with PSID coding procedures; and Beth-Ellen Pennell and the Survey Research Center interviewing staff for advice and implementation of data collection and coding.

References

Catlin, G. and Ingram, S. (1988), "The effects of CATI on costs and data quality: A comparison of CATI and paper methods in centralized interviewing." Chapter 27 in Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. and Waksberg, J. (eds.), *Telephone Survey Methodology*. New York: Wiley.

Dielman, L. and Couper, M.P. (1995), "Data quality in a CAPI survey: Keying errors." *Journal of Official Statistics*, forthcoming.

Groves, R.M. and Mathiowetz, N.A. (1984), "Computer assisted telephone interviewing: Effect on interviewers and respondents." *Public Opinion Quarterly*, 48: 356-369.

Kennedy, J.M., Lengacher, J.E., and Demerath, L. (1990), "Interviewer entry error in CATI interviews." Paper presented at the International Conference on Measurement Errors in Surveys.

Nicholls II, W.L. and Groves, R.M. (1986), "The status of computer-assisted telephone interviewing: Part I -- Introduction and impact on cost and timeliness of survey data." *Journal of Official Statistics*, 2 (2): 93-115.

Olsen, R. (1991), "Mode effects on data quality -- CAPI versus pencil and paper." Ohio State University, unpublished paper.

Rustemeyer, A. (1977), "Measuring interviewer performance in mock interviews." *Proceedings of the American Statistical Association, Social Statistics Section*, 341-346.

Sebestik, J., Zelon, H., Dewitt, D. and O'Reilly, J.O. (1988), "Initial experiences with CAPI." *Proceedings of the Bureau of the Census Fourth Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, pp. 357-365.

Tortora, R.D. (1985), "CATI in an agricultural statistical agency." *Journal of Official Statistics*, 1 (3): 301-314.

Table 1. Number and Percentage of Question Askings by Type of Question and Mode

Type of Question	CATI		Paper & Pencil	
	n	%	n	%
Total	26,286	100.0	27,662	100.0
Closed-Ended	25,063	95.3	26,413	95.4
Simple	18,543	70.5	19,210	69.4
Complex	6,286	23.9	6,478	23.4
Other	234	0.1	725	2.6
Open-Ended	1,223	4.6	1,249	4.5

Table 2. Closed-Ended Questions by Type of Interaction, Recording Error, and Mode

Type of Question and Recording Error	CATI		Paper & Pencil	
	n	%	n	%
Total	25063		26413	
Simple strght-frwrd	18543	100.0	19210	100.0
Accurate	18523	99.9	19184	99.9
Inaccurate	20	0.1	26	0.1
Complex	6286	100.0	6478	100.0
Accurate	5975	95.0	6004	92.7
May be misrecorded[a]	311	5.0	474	7.3
Other	234	100.0	725	100.0
Skipped, no answer	0	0.0	267	36.8
Skipped, answered	123	52.6	229	31.6
Uncodable	111	47.4	229	31.6

[a] Mode difference statistically significant at $p < 0.001$.

Table 3. Number and Percentage of Complex Interaction Errors by Type of Error

Type of Error	CATI		Paper & Pencil	
	n	%	n	%
Total	311	100.0	474	100.0
Implied response	128	41.2	131	27.6
Qualified resp.	76	24.4	76	16.0
Failure to record	17	5.5	123	25.9
Failure to repeat/probe	52	16.7	69	14.6
Simple misrecording	28	9.0	67	14.1
Decimal shift	1	0.3	2	0.4
Numeric error	9	3	6	1

Table 4. Misrecording Errors in Complex Interactions by Response Set

Resp. set	CATI			Paper & Pencil		
	Total	Err.	%	Total	Err.	%
Total[a]	6181	308	4.9	6449	472	*7.3*
Short answer						
Simple	342	19	5.6	388	35	*9.0*
Amount	781	48	6.2	858	56	*6.5*
Time[b]	1119	58	5.2	1179	100	*8.5*
Mult. choice						
Simple	790	47	6.0	788	65	*8.3*
Time	1053	62	5.9	1062	88	*8.3*
Month[c]	284	12	4.2	325	19	*5.9*
Yes/No	1812	62	3.4	1849	109	*5.9*

Mode difference statistically significant at [a] $p < 0.001$, [b] $p < 0.01$, or [c] $p < 0.05$.

Table 5. Number and Percentage of Open-Ended Questions by Type of Recording Behavior and Mode

Recording Behavior	CATI		Paper & Pencil	
	n	%	n	%
Total	1,223	100.0	1,249	100.0
Verbatim	201	16.4	260	20.8
Essence	892	72.9	853	68.3
Misrepresents	72	5.9	100	8.0
Other				
Skipped, answer not recorded	40	3.3	12	1.0
Skipped, answer recorded	15	1.2	10	0.8
Non-codable	3	0.3	14	1.1

Appendix: Recording Error Codes

Description	Code
Closed-ended questions	
Simple straightforward exchange	
(No probing and the answer was codable, not qualified, or elaborated on. Verification of answer by the interviewer are acceptable.)	
ACCURATE	0
INACCURATE	1
(Recorded answer does not accurately represent respondent answer, including failure to record an answer when given.)	
Complex exchange	
(Interviewer probed response or respondent qualified answer, gave uncodable answer, sought clarification, or elaborated the answer.)	
ACCURATE	3
MAY BE MISRECORDED	4
(Recorded answer may distort what respondent said.)	
Other	
SKIPPED, NO ANSWER	6
(Interviewer skipped question reading and no answer is recorded.)	
SKIPPED, ANSWER	7
(Interviewer skipped question reading and recorded an answer.)	
UNCODABLE	9
(Coder unable to assign a code due to tape or other difficulties.)	
Open-ended question	
VERBATIM	0
(Recorded respondent's words except for insignificant deletions, additions, or substitutions.)	
ESSENCE	1
(Recorded answer represents respondent's answer despite significant difference between respondent statement and interviewer's recorded statement.)	
MISREPRESENTED	3
(Recorded answer clearly misrepresents respondent answer.)	
SKIPPED, NO ANSWER	6
SKIPPED, ANSWER	7
UNCODABLE	9