Journal of Official Statistics
Vol. 2, No. 2, 1986, pp. 117–134

The Status of Computer-Assisted Telephone Interviewing: Part II – Data Quality Issues

Robert M. Groves and William L. Nicholls II[1]

Abstract: This is the second of a two-paper series presenting an overview of the status of computer-assisted telephone interviewing (CATI) based primarily on published articles and unpublished reports describing experiences with and evaluations of CATI in the U.S. and other developed nations. This second paper examines current knowledge about CATI's consequences for data quality, including effects both on nonresponse and measurement error. It also reviews CATI's new contributions to data collection methodology and suggests needed areas of research to assess better CATI's data collection characteristics.

Key words: CATI; computer-assisted telephone interviewing; telephone interviewing; data quality; nonresponse; measurement error.

1. Introduction

The first paper in this series reviewed those aspects of computer assisted telephone interviewing (CATI) that can affect the costs of data collection and the speed with which surveys are completed. Costs and timeliness form one set of criteria to assess CATI's desirability as a data collection method, but they ignore changes in data quality that may result from its use. This paper focusses on data quality issues in CATI. It is structured about two sources of error: (1) nonresponse, including both loss of entire sample cases (unit nonresponse) and item missing data; and (2) measurement error, that is, errors associated with question wording, interviewer actions, and respondent failures to provide accurate answers.

This paper is critical of the current state of our knowledge about CATI's impact on data quality. The reader will learn that, with a few exceptions, there is little reliable empirical evidence that CATI affects data quality. This absence is especially noteworthy in the context of the frequent expectation of data quality improvement from CATI. Groves (1983) said "... it seems that most of the changes in moving from non-CATI work to CATI work will arise not in the reduction of costs but in changes ... in the error structure." Nicholls (1978) said that instead of cost reasons guiding the choice of CATI, "CATI is to be

[1] Robert M. Groves is a Senior Study Director at the Survey Research Center, The University of Michigan, Ann Arbor, MI. 48106, USA. William L. Nicholls II is Director of the CATI Project, Center for Survey Methods Research, U.S. Census Bureau. The first paper in this series acknowledges colleagues who made important contributions to the two-paper set.

recommended for other reasons, such as its survey management capabilities, its improved standardization of survey practice, its opportunities for field and sampling quality control, and its ability to handle survey instruments of unusual complexity." Dozens of quotations could be given with similar sentiments (e.g., Fink (1983), Dekker and Dorn (1984)). Despite the abundant descriptions of potential quality improvement with CATI, there are also some concerns about lower data quality. Presser (1983) has noted that while CATI improves data quality in some respects, it also "can produce data problems that are unlikely to occur in paper-and-pencil interviewing." It is difficult, however, to move beyond these statements to empirically based estimates of data quality.

The lack of empirical results on data quality stem mainly, we believe, from the large amount of energy devoted to software development to the detriment of experiments in survey design and measurement using CATI. Indeed, the attention of survey practitioners has been dominated by the problems of adapting existing paper questionnaires to CATI form. Much of the day-to-day usage of CATI systems still involves the transition from a paper technology to a computer-assisted one. This concentration on how to make a terminal screen behave like a paper questionnaire, has, we believe, been a necessary first step in adapting to the new technology, but one that has left basic questions about how CATI can affect data quality unanswered.

2. Nonresponse Error

CATI can affect survey nonresponse directly through its special features and indirectly through interviewers' and respondents' reactions to using the medium. The call scheduling feature of CATI systems is the most direct attempt to use CATI to reduce unit nonresponse, the failure to contact sample units or the failure to persuade those contacted to grant the survey interview. Call scheduling software implements algorithms that determine when a sample number will be called (and sometimes which interviewer will make the call). Given a fixed survey period, if a call scheduling procedure can obtain interviews with sample units more quickly, it will often obtain higher response rates for the entire survey.

Table 1 presents the contact rates for three telephone surveys using CATI and non-CATI methods for different portions of the sample. All three show significantly higher proportions of the CATI sample cases were contacted by interviewers (whether or not they yielded interviews). The largest difference in contact rates (72 percent with CATI, 57 percent without CATI) appears for a CATI system without computer-assisted call scheduling (House (1984)). This test used four CATI stations, with interviewers supplied cases to dial by a supervisor standing behind them. Thus, the CATI version resembled in some features the process of computer assisted call scheduling. Non-CATI interviewers made their own decisions on when to dial and what numbers to dial. The two studies reported by Ferrari (1984, 1986) were telephone followup studies to mail questionnaire nonresponse. In both tests, a larger proportion of the CATI sample was contacted. In these surveys, current telephone numbers of sampled persons frequently were not available, and the interviewers were asked to obtain telephone numbers from directory assistance and other sources. These procedures were incorporated into the CATI instrument as a mandatory part of the interviewers' work. The non-CATI interviewers followed guidelines containing the same rules, but independently selected cases to call within their assignments. Since fully controlled experimental designs were not employed in these comparisons, other explanations of the

Table 1. Contact and Refusal Rates in Several Surveys with CATI and Non-CATI Components

Survey	Contact Rate (%)		Refusal Rate (%)	
	CATI	Non-CATI	CATI	Non-CATI
Colo-Rectal Cancer Survey (Harlow et al. (1985))			8	8
Nebraska Hog Survey (Coulter (1985))			9	8
Cattle Dual Frame Survey (House (1984))	72*	57*	8	8
Survey of Scientists and Engineers (Ferrari (1984))	50.2*	44.4	21.7	21.2
Census of Agriculture (Ferrari (1986))	84.3*	79.0	5.2*	12.5

* Statistically significant difference between CATI and Non-CATI at 0.05 level.

results are possible, but they suggest that CATI may more fully enforce interviewing procedures affecting response rates.

Call scheduling algorithms now in use appear to vary on several points:

a) Whether they are deterministic or assign priorities to all active sample numbers. Some call scheduling algorithms use information about the timing of past unsuccessful calls to determine which subset among all active numbers should be called during a given time period. Only those numbers are eligible for calls during the time period. Other algorithms identify priority groups or assign priority scores to all active numbers for a given time period, and numbers are dialed within the time period in order of their priorities. With this approach it is theoretically possible to call all numbers in the sample if the time period fails to yield interviews.
b) Whether the algorithm attempts to maximize the cumulative probability of a number being answered over the survey period or attempts to maximize the productivity of interviewers at each moment in the survey period. Some algorithms attempt to identify and schedule for each interviewing shift those numbers that have the highest probability of contact at that time. Others attempt to assure that all numbers are scheduled for calling in a way that they are individually given the best chance of being answered *sometime* during the survey period.
c) Whether the call scheduling procedures require human intervention for certain types of activities. Some call scheduling systems have the supervisors assign appointments and refusal conversion cases to interviewers, after a review of the case. Others have complete machine control over all cases, except some that are placed into a group needing supervisory review (e.g., unusual sampling problems or respondent difficulties).

Despite these rather large differences among call scheduling procedures, there is very little written documentation on the rationale for the differences. Further, there appears to have been little experimentation regarding the effects on nonresponse rates of differences in call scheduling procedures. Once a procedure is chosen for a CATI system, it generally is used without evaluation. What is needed at this point is an experimentally controlled test of the value of automated call scheduling on unit nonresponse rates (by comparing to a paper scheduling procedure) and experimental variation of the three major distinguishing features above.

Refusals to grant an interview can also be affected it either respondents or interviewers find the cognitive or affective burdens of CATI different from those of traditional methods. In early use of CATI by university survey groups respondent acceptance was a major concern. Researchers debated the desirability of informing respondents that answers were being entered into a computer and of using terminals with silent keyboards. Within the United States these concerns have all but vanished among those conducting CATI surveys. No evidence of respondent resistance to CATI (as distinct from telephone interviewing) has been reported. Shared experiences suggest that it makes little difference whether or not respondents are informed about the computer. Anecdotal evidence suggests that at least some respondents take the survey more seriously when they learn their answers are being entered directly to a computer (Morton and House (1983)).

Published studies of interviewer reactions to CATI, based on staff debriefing sessions and questionnaires, have reported that interviewers often compare CATI favorably to paper-and-pencil methods of telephone interviewing. For example, both Groves and Mathiowetz (1984) and Coulter (1985) present results from a postsurvey questionnaire administered to interviewers who had conducted both CATI and non-CATI interviews on the same survey. They argue that there are few differences in reported interviewer training difficulties between CATI and paper-and-pencil interviewing (see Table 2). The Groves and Mathiowetz data show some suggestion of greater tension among interviewers on CATI, although the numbers are very small. Coulter (1985) also presents interviewers' relative assessment of individual features of the two modes of interviewing. These show large preferences for paper-and-pencil methods for changing answers to previous questions and some preference for CATI in arranging callbacks, but little or no preference for other aspects (i.e., question branching, entering numeric answers, and entering notes). These studies are based on small samples of interviewers, who were novices on CATI. The impressions of most organizations who have conducted large numbers of CATI surveys is that attitudes towards CATI become even more positive with time.

Interviewer acceptance of CATI appears to depend on the reliability and speed of the CATI system utilized. Frequent hardware, software, or instrument failures can undermine confidence and raise frustration levels (Nicholls (1978), Morton and House (1983)). System response times (the length of time between the entry of an answer and the display of the next question) are especially critical. Response times exceeding two seconds appear to make it difficult for the interviewer to maintain an appropriate pace. One controlled study which compared CATI and non-CATI telephone interviewing found a significant difference in response rates only in the first of three replicates, during a period of CATI hardware problems (Groves and Mathiowetz (1984)). In well functioning CATI systems, there is little evidence that CATI affects refusal rates. Four studies summarized in Table 1 found no appreciable difference in CATI and non-CATI refusal rates. The one exception reported a significantly lower refusal rate for CATI which may be attributable to uncontrolled factors in this comparison study.

Item missing data arise both from the interviewer's failure to ask a question or enter a response *and* respondent failure to provide a substantive answer. Perhaps the most frequently cited advantage of CATI is rigid control over question flow and recording of responses, forcing the interviewer through question sequences appropriate to the respondent, and demanding data entry at all questions presented. This feature can theoret-

Table 2. Responses of Interviewers to Postsurvey Questionnaire from Two Experimental Studies

Response	Number Responding			
	CATI		Non-CATI	
	Groves and Mathiowetz (1984)	Coulter (1985)	Groves and Mathiowetz (1984)	Coulter (1985)
Difficulty of learning to interview in mode				
Very difficult	1	0	0	0
Somewhat difficult	7	4	5	1
Somewhat easy	14	7	14	8
Very easy	9	2	12	4
Fatigue due to interviewing				
Very tiring	3	1	1	2
Somewhat tiring	7	5	9	4
Not very tiring[1]	18	3	19	4
Not tiring at all	2	4	2	3
Not ascertained	1		0	
Tension due to interviewing				
Very tense	2		0	
Somewhat tense	9		6	
Not very tense	13		14	
Not tense at all	6		10	
Not ascertained	1		1	
Total	31	13	31	13

[1] Instead of the phrasing "Not very tiring", the Coulter work used "Only a little tired".

ically eliminate interviewer errors of skipping questions. It is also possible in CATI for the researcher to prohibit an interviewer from entering a "don't know" answer by limiting acceptable entries, but in practice many organizations permit interviewers to enter "don't know" and refusal codes to most or all questions, as with paper questionnaires. Forced data entry at each question does not ensure entry of a substantively meaningful value.

When the same questionnaire is used in CATI and on paper, CATI interviews have been found to experience about the same levels of item nonresponse on demographic and income items as paper-and-pencil telephone interviews (Groves and Mathiowetz (1984)). CATI *does* enforce the proper skip patterns in contingent questioning sequences, removing the burden from the interviewer of exercising skip logic. On 28 questions with complex skip patterns, 1.8 percent of the CATI entries were found to have consistency errors (errors that involve item missing data) compared to 8.8 percent of the paper-and-pencil entries.[2] Fielder (1985) compiled item missing data rates for a repeated cross-section RDD survey on community response to earthquake threat in California, which was conducted using paper methods on the first two waves and CATI on the third and fourth. Table 3 shows a reduction in item missing data rates (combined "don't know," "refused," and "not ascertained") for demographic, income, and opinion items. None of these differences exceed common levels of statistical significance.

[2] This study illustrates the difficulties of obtaining precise comparative estimates of such apparently simple effects on data quality. If paper-and-pencil interviewers are permitted to edit their forms after completion of the interview and if these forms also pass through clerical editing and key entry before the comparisons are made, these later stages of paper-and-pencil processing may increase or decrease errors or mask those committed in interviewing. Ideally, both gross errors existing at the end of the interview and net errors at the completion of processing should be examined in the comparisons.

Table 3. Mean Percentage of Cases with Item Missing Data

Item Type	Mean Percentage of Cases with Item Missing Data	
	CATI	Non-CATI
Health Survey Test (Groves and Mathiowetz (1984))		
Sex, Age, Education	0.7	0.9
Race, Income, Marital Status	14.6	15.0
Community Response to Earthquakes (Fielder (1985))		
15 Demographic Items	0.3	0.5
7 Income Items	1.7	2.7
11 Opinion Items	2.3	5.3
Survey of Scientists and Engineers[1] Ferrari (1986))		
14 Items Asked of All Respondents	7.1*	24.6
12 Items Asked of Most Respondents	7.5*	26.6

[1] Excludes imputed and recoded items whose missing data rates could not be ascertained.
* Statistically significant difference between CATI and non-CATI at the 0.05 level, edit and imputation.

The comparison of item missing data rates between CATI and non-CATI methods is clearly dependent on the nature of questionnaire formatting and training procedures in each of the modes. In cases where the number of items asked is partly left to the interviewers' judgment, CATI can demonstrate large reductions in item nonresponse. This can occur in telephone followup to mail questionnaire nonresponse, where the interviewer must reconcile the conflicting goals of obtaining as much information as possible and not antagonizing possibly reluctant respondents. The first test of the U.S. Census Bureau's CATI system in the Survey of Scientists and Engineers was for this type of application. As shown in Table 3, rates of item nonresponse were substantially lower for the CATI than the non-CATI staff. The greater difficulty of omitting applicable questions in CATI may have contributed to this difference, but since the results are not based on a fully controlled experimental design, they remain only suggestive.[3] Some systems record item missing data frequencies by interviewer for interviewer performance reports.

If the researcher anticipates item missing data problems on an individual question, CATI may also reduce item nonresponse by enforcement of probing on inadequate answers. In recent CATI surveys undertaken by the U.S. Census Bureau, information on the age of household members is first requested by asking for each person's exact date of birth. If this is not known by the respondent, the person's age in years is asked. If this also is unknown, age is asked for in broad categories necessary for later stages of the interview or for tabulations. Each successive probe appears on the interviewer's screen only when the interviewer fails to enter the required information to the prior question. In paper questionnaires interviewer instructions for such probing in difficult cases typically are provided in a separate document, and the prescribed procedure is followed only if the interviewer remembers. Systematic evidence on the efficacy of enforced probing with CATI currently does not exist, so its possible effects on item nonresponse remain speculative.

[3] An alternative interpretation is that CATI pressures the interviewer to enter some response, even in cases where none is warranted.

Despite the above positive effects of CATI on nonresponse and missing data, some features suggest potential increases in item missing data. The most serious are the result of undetected mistakes in the design of CATI instruments. Presser has argued "that the complexity of CATI technology exceeds our present ability to check it....Typographical errors in paper questionnaires rarely affect data quality. Logic errors in CATI applications may frequently impair data quality" (Presser (1983)). Presser cites two types of setup errors with severe consequences. The first type are mistakes in branching or skipping instructions which result in questions not being asked of applicable respondents. The second type are mistakes in assigning entries to data fields which may result in overwriting earlier information from another question. In most current CATI systems, both mistakes result in permanent loss of data, and both occur more frequently than those unfamiliar with CATI data collection would anticipate.[4] Others strongly disagree with this observation and note that errors are much easier to correct in one CATI application than in thousands of paper copies of questionnaires (Palit and Sharp (1985)). Furthermore, they note that a typographical error in a paper questionnaire can be as fatal as one on CATI.

Protection from such setup errors in complex surveys begins with careful design work (including modular design of sections and sometimes flowcharting of the application) but also requires thorough debugging of each CATI questionnaire and output file prior to production interviewing. Proof-reading CATI instruments and trials of the system with 20 or 30 test cases are rarely sufficient, even when the check follows each case through to the output file. Several techniques are being used by different organizations to supplement such routine checks:

a) Display of the answers from critical earlier questions in later questions to more easily identify branching errors during debugging.
b) Use of screens which may be accessed from anywhere in the interview, to summarize large sets of answers to prior questions.
c) Similar screens that appear at the end of each section of the interview to summarize entries obtained in that section. (These screens are displayed only during the debugging phase not during production interviewing.)
d) Utility programs that summarize the logical structure of the interview, display the assignment of questions to data locations, and branching to and from each question, including those that cannot be reached from any other question.

Despite these aids, the debugging process remains one of the most vulnerable areas of CATI data collection, and further development in this area is clearly needed. Documentation of practices across survey organizations might be productive of real advances.

Item missing data can also arise from hardware or CATI software errors. Researchers unfamiliar with CATI are often concerned that hardware crashes or software errors will erase hours or days of interviewing work. Relative to the early days of CATI, data losses from these sources are rare. Some systems preserve all but the last entry or two of interviews in progress if a malfunction occurs in the CATI software but the operating system is still functioning. An operating system or cpu crash, however, may lose interviews then in progress. In either event, completed interviews are not affected, and backups to magnetic tape or another medium are often routinely scheduled for added protection. Losses of entire interviews appear to be extremely

[4] It is noteworthy that early CATI systems were not as vulnerable to overwriting errors. Often they made entries to a "transaction" file which recorded every interviewer entry. The final data file was then constructed from the transaction file. Most current systems now write entries to fixed-field records, and if the answer is changed, that field is overwritten. Despite their disadvantages, transaction file systems provided a better means of reconstructing interviews when problems resulted from setup errors.

rare once production systems have stabilized. The U.S. Census Bureau has experienced no such loss in approximately 10 000 completed CATI interviews conducted over a three year period.

3. Measurement Error

Measurement errors exist in surveys when recorded answers to questions do not accurately represent the respondents' characteristics. Measurement errors may be attributable to one or more of the entities in the data collection process – the respondent, the interviewer, the questionnaire, supervisory or clerical staff who edit or transcribe completed forms, and key entry staff. CATI eliminates the last two as *independent* sources of error, although in most studies it is likely that they contribute little to total survey error and may aid in the detection and correction of errors from earlier steps. CATI acts most clearly on the questionnaire as a method to affect measurement errors, and in doing so has the potential of changing the effects that interviewers, supervisors, and respondents have on measurement errors.

When contemplating how CATI can alter the effects of interviewers on measurement errors, it is useful to examine how CATI *changes* the activities performed by the interviewer. The largest change with CATI is the use of a terminal to display questions and enter survey responses.

3.1. Reading the Survey Questions

When survey questions are transferred without change from paper forms to CATI displays, there seems little reason to expect that CATI will either help or hinder the interviewer's ability to read them correctly. The ergonomic concerns about video intensity, size of characters, and position of terminal screens are legitimate, but appear to have solutions common to all situations of terminal use. When a CATI application alters the text of a question to tailor it to a respondent's prior answers, it may ease the burden of the interviewer and thereby improve the survey measure. For example, in a paper version of the U.S. Health Interview Survey, the following question appears:

4a. For what condition did – – – see or talk to the [doctor/(entry in 3c)] on (date in 1)?

The interviewer is instructed to read the question while entering the correct name or pronoun in the first blank and determining what kind of medical person was seen and on what date the visit took place. *If the CATI application designer chose to utilize the "fill" capability most systems have,* the question could appear as:

4a. For what condition did Frank see or talk to the podiatrist on July 5, 1985?

This form of the question clearly reduces the burden on the interviewer to deliver the desired wording of the question. The relevant issues in terms of measurement error are, however: a) how often are errors in wording made in the paper version and how damaging are those to respondent answers, and b) how often are there delays in the interviewer asking the question because of the necessity to assemble the proper words and what effects do these have on respondent answers? In short, there is no empirical research on the effects of easing interviewer burden through use of "fills." It *is* clear in all CATI systems, however, that taking advantage of such CATI capabilities requires more design time than ignoring the option. Errors in question delivery because of the burden of entering the appropriate words can be tested using small scale laboratory studies with interviewer subjects. These would be low cost experiments designed to measure delays in question

delivery without "fills" and errors in "filling" that interviewers make. If these errors had a nonnegligible frequency, later experiments could be embedded in ongoing surveys to measure the effect on respondent answers of the fill capability. Any ongoing CATI facility could easily implement such experimentation.

Another question related to wording issues and CATI is the enforcement of probing, a procedure left to interviewers to implement in most paper questionnaires. Harris (1952) reported that probing failures appear to be the most common "invisible" interviewer error among personal interviewers in the U.K. Social Survey. The work of Cannell and his associates (see Cannell, Miller, and Oksenberg (1981)) has explored enforced probing by placement of followup questions in the questionnaire (removing them from interviewer discretion). Again, there is little reason to suspect inherent differences in probing behavior associated with the use of CATI. The general approach of Cannell to programming probes is, however, a natural use of CATI capabilities. For example, instead of the following form on paper,

> During that two-week period, that is, from _____ to _____, how many days did illness or injury keep you from work?
>
> _____ days

the programmed feedback would have the interviewer characterize the respondent's answer, then give an appropriate followup feedback or probe:

> For this question, we'd like to get the number as exact as you can report it. During that two-week period, that is from April 1 to April 14, how many days did illness or injury keep you from work?
>
> 1. EXACT NUMBER
> 3. RANGE; "ALL WEEK"
> 8. DON'T KNOW
> 9. Not Ascertained
>
> If an exact answer is given (answer number 1 above),
>
> RECORD NUMBER OF DAYS
> I see, we're interested in that.
>
> If a range answer is given,
> Could you be more exact about the number of days?
>
> If a "don't know" answer is given,
> Would you think for a minute and give me your best estimate?

Even greater dependence on computer capabilities can enhance survey measurement. One CATI survey of economists, in addition to having valid ranges for answers to a question regarding estimated inflation rates in the future, asked the interviewer to verify answers that were beyond those reasonably offered, given prior answers to similar questions (Survey Research Center (1985)). Fink (1983) describes a CATI measurement of tolerance to a doubling or tripling of cigarette prices in a survey of smokers. The respondent first reported what he/she currently paid per pack (e.g., "$0.65 per pack") and then about their likelihood of continued smoking if the price increased (e.g., "to $1.30 per pack," "to $1.95 per pack"). A CATI survey of large fleets of commercial vehicles first asked the respondent to report how many vehicles of certain types (e.g., small trucks and vans) were in the fleet (Survey Research Center (1983)). Followup questions asked for further categorization of the vehicles (e.g., "How many of the 27 small trucks and vans are typically driven less than 60 miles per day?", "How many are driven less than 30 miles per day?"), with checks that the number of vehicles reported did not conflict with the answers previously given.

Alterations to question text, such as fills, enforced probing, and arithmetic feedback to respondents, depart substantially from paper-and-pencil questionnaire design. This increases setup time and risk of errors in the CATI application. The CATI community clearly

needs guidance concerning which of these changes act to reduce measurement error sufficiently to justify their costs.

3.2. *Recording of Textual Material*

Most speculations about interviewer entry errors with CATI concern open questions, where interviewers must type in the words of the respondent. Entry errors can occur in this task through:

a) the inability to enter the respondent's answer quickly enough or fully enough, and possible effects on respondent behavior when entry is slow, and
b) typographical errors that make the recorded response unintelligible.

Although CATI interviewers generally are required to have minimal typing skills, usually at least 20 words a minute, most can write faster than they can type. If they cannot enter the answers as quickly as respondents deliver them, they are often taught to tell the respondent that they need more time for full recording of the answers, the same guidance given to telephone interviewers for paper questionnaires. Nevertheless, the slower rate of recording in CATI may reduce the completeness of records of open question responses and perhaps discourage some respondents from making full replies. Morton and House (1983) summarize field staff impressions on two CATI surveys which suggest that this is not the case. They say: "A common objection to the use of CATI for collecting free response material has been the belief that typing is slower than recording verbatim by hand ... we found that lack of typing speed did not seem to be an irritant to the respondent, and speed did improve as interviewers felt more comfortable with the keyboard." One of the studies Morton and House describe made extensive use of free answer questions and employed interviewers previously experienced with paper-and-pencil interviewing methods. After this staff was trained in CATI interviewing, they were given a choice of recording free responses on CATI or writing them on paper for later entry to CATI. "Although only one of the interviewers was a trained typist, all chose to record responses directly on the CRT ... The unanimous judgment of the field staff was that recording was easier and at least as complete with answers entered directly."

Harlow et al. (1985) do not compare CATI and non-CATI performance on open questions, but do examine the entry of respondent comments and the recording of probes used by the interviewer. Interviewers in both modes were instructed to record comments and probes. On CATI they used keyboard function keys to initiate the recording. Fewer respondent comments (4.1 per interview on CATI; 5.5 on non-CATI) and fewer probes (8.3 per interview on CATI; 10.2 on non-CATI) were recorded using the CATI keyboard than using paper questionnaires. The authors speculate that one explanation for the reduced documentation on probes for CATI was the fact that interviewers could not indicate a probe once the final answer for a question was entered.

The recording of textual material is an area ripe for careful study. Comparisons of open responses recorded by CATI and by paper-and-pencil methods in the same survey, or mock interview studies, are needed. Visual monitoring in CATI also may be used to assess such deficiencies when combined with an objective coding system like that proposed by Mathiowetz and Cannell (1980). Again, the type of research that is required here first is low cost experimentation with interviewers as subjects in a controlled survey laboratory.

3.3. *Recording Responses to Closed Questions*

Recording errors to closed questions in paper-and-pencil methods have not been frequently studied, and the field thus lacks a standard by which CATI recording errors can be evaluated. There are, however, important changes in the interviewers' task on CATI when indicating

responses. With paper forms, the interviewer checks a box or circles a number adjacent to the response category that is chosen. In CATI, the interviewer depress a number on the keyboard corresponding to the respondent's answer.[5]

CATI systems attempt to minimize recording errors in questions with fixed or other numeric answers by limiting entries to permissible ranges. If an entry is made outside this range, the interviewer is alerted to the error with a message. Without the enhancements in question wording and skip logic described above, a case can be made that recording errors are more frequent with CATI than with paper-and-pencil methods. With paper forms, the interviewer's hand and eye are both concentrated on the box to check or the number to circle which lies immediately adjacent to the chosen answer.[6] Recording a closed response in CATI requires more hand-eye coordination. The chosen response number is read from the screen, entered on the keyboard, and verified at a different place on the screen. Since studies comparing recording errors in CATI with those on paper forms have not been undertaken, it remains unknown whether on balance CATI increases or decreases their frequency.

3.4. Corrections to Previous Answers

Another error associated with interviewers occurs when respondents change their answers to an earlier question. This appears to be a rare phenomenon in most paper questionnaire surveys, but there is little evidence of its actual frequency. When this occurs the interviewer is often instructed to locate the question, erase or mark out the prior answer, and replace it with the new one. The interviewer is trained to follow the question flow appropriate to the changed answer.

On form-based CATI systems, a return to prior data items is sometimes accomplished by cursor movement to a field displayed on the same screen. These systems more closely resemble paper questionnaires on this attribute than do item-based systems.[7] To move to a prior question or screen in these systems, the interviewer uses special function keys or commands. Most CATI systems permit repeated use of this function to move backward, one question at a time. In one test by the National Opinion Research Center, interviewers backed up to review answers once in every 50 questions. Once in every 20 questions an interviewer returned and changed an answer (Press (1985)). Such findings will obviously vary by the nature of the application and the interviewers' familiarity with it.

Problems sometimes arise when it is necessary for the interviewer to return to a question too far back in the interview to be reached efficiently by repeated single backing. Some systems allow the interviewer to back to the beginning of a section of questions (sections being defined by the researcher). Other systems permit the interviewer to jump back to items identified by their item label; but means must then be found providing the interviewers with ready access to the item labels of questions they may need to reach. Use of this option is not frequent. In the National Crime Survey CATI test at the U.S. Census Bureau, CATI interviewers appeared to do this about

[5] In most CATI systems, the interviewer must then depress the return key to make the entry. In a few, the entry is made as soon as an appropriate number key is depressed, and the interviewer cannot review the entry before the next question is displayed. (It is uncertain, however, how often interviewers actually review the response code in two-action systems before moving to the next question. Frequently, the number key and return key are depressed almost concurrently.)

[6] CATI systems which use light pens rather than keyboards to select and record response categories more closely parallel paper-and-pencil forms in concentration of hand and eye at the same point. Of course, they must employ other methods for entering responses to open questions.

[7] The distinction between "item-based" and "form-based" CATI systems is described in Section 2.4 of Part I.

one case in fifty.[8] Other systems attempt to identify questions to which the interviewer might want to return and provide jump back options at that point, such as simple entries that will make the jumps automatically.[9] None of these CATI options, however, provide interviewers with as fast and flexible a method of reaching former questions as paper questionnaires do.

Although the backing up may be more laborious, most CATI systems provide a faster and more accurate method of continuing the interview after changes are made than do paper-and-pencil methods. If the changed answer does not affect the following questions to be asked, and if all edit checks are passed, the interviewer will be returned to the display where the backward movement began. But if the revised answer changes the questions to be asked or makes a later answer inconsistent with the previous one, the system will stop at following items which now need to be asked or changed. The computer performs these tasks far faster and more accurately than an interviewer could with paper-and-pencil methods.

3.5. *Supervisory Oversight*

Such common forms of serious interviewer error as altering the scope of the question and failure to probe, typically are not detectable from review of completed forms. Even without CATI, centralized telephone interviewing permits unobtrusive telephone monitoring of any interviewer at any time. This represents a major advance in the control of previously "invisible" interviewer errors by placing them under supervisory observation and corrective action. CATI permits even more thorough supervision by adding visual monitoring of the interviewer's screen to the audio monitoring of the respondent-interviewer conversation. This allows a monitor to see each screen as the interviewer reads it, hear the respondent's answer, and see the answer entered by the interviewer. In surveys where the wordings of questions vary with prior information obtained, visual monitoring may be required to ensure that the monitor is alerted to the correct versions of questions. Finally, through video monitoring, CATI can provide greater protection than any other form of survey interviewing from the most damaging interviewer errors and misbehavior, such as interviewing the wrong respondent, falsification of entire interviews ("curbstoning"), or omitting questions and guessing the answers. No systematic evidence is available on the frequency of such gross interviewer errors in conventional surveys, although most large data collection agencies have encountered them at least occasionally. They appear to be more common among interviewing staffs in large metropolitan areas where the recruitment and retention of well qualified interviewers often proves difficult.

Cannell et al. (1983) used audio and video based monitoring and structured coding of interviewer behavior to assess compliance with interviewer training guidelines. In a survey employing both CATI and non-CATI interviewing, they measured whether the interviewer read the questions as worded, whether they probed for a complete answer when necessary, whether they delivered the questions at the specific pace, etc. They found variation across questions in the proportion read correctly (from about 0.80 to 0.96), but they do not cite differences between CATI and non-CATI cases.

3.6. *Editing for Consistency*

Interview data may be checked for internal consistency during the interview, by sub-

[8] This low frequency of occurrence may reflect the difficulty of the action.

[9] The previously given count of jump back actions in the National Crime Survey does not include this type of automatic jumpback from simple entries to specific items.

Table 4. Summary of Edit Errors in Cattle and Hog Inventory Survey

Method of Interviewing	Number of Edit Errors		
	Total	Critical	Non-Critical
Non-CATI	245	53	192
CATI	190	12	187
Percent Relative Difference[1]	22	77	3

[1] 100 (Non-CATI – CATI)/(Non-CATI)
Source: Tortora (1985) or for a fuller description, see House (1985).

sequent clerical review, by key entry edits, or by batch computer edits. With paper-and-pencil methods, the cross-checking of items during the interview is usually quite limited. Interviewers may be asked to see that reported percentages add to 100 percent or to identify and correct obvious inconsistencies, such as children reported as older than their natural parents. More detailed consistency checking is usually postponed until after the interview, especially if it involves complicated arithmetic. Consistency errors discovered after the interview during clerical review, key entry, or batch computer edits are usually resolved by recontacting the respondent (a costly procedure), setting one or more variables to missing data values, or altering the values of one or more variables. Very little is known about the magnitude of error introduced into final estimates by these procedures.

CATI can introduce online data evaluation through various uses of consistency checking. In addition to use of range edits, acceptable entries may be further limited by checks based on the answers to prior questions. For example, multiple-answer questions may employ checks to ensure that the same answer is not entered more than once; or a precoded question on stolen property recovered after a crime may restrict entries to property previously reported stolen. Alternatively, entries inconsistent with prior answers or other previously entered information may prompt additional questions to reconcile the discrepancy. These probes may identify errors in recording responses as well as inconsistencies among the respondent's answers. Finally, answers from former questions (or text insertions based on them) may be displayed in later questions providing the interviewer with additional opportunities to identify and correct earlier recording errors. Unfortunately, there are few documented effects of these enhancements, and much of the practice in CATI applications at this time merely adapts paper questionnaires to CATI usage or does not permit comparison of errors with and without such enhancements. Online consistency editing with CATI has costs in setup time, interviewing time, and computing resources. Thus, batch computer editing after the interview sometimes may continue for some inconsistencies, while online editing may be restricted to edits which: (1) would be performed by clerical or key entry staff with paper forms;[10] (2) would require a recall of the respondent if an edit failure occurred; or (3) are of special importance to the survey.

Tortora (1985) presents a comparison of edit failures from data collected by CATI and those collected using paper methods for a

[10] Not all forms of clerical review are eliminated by CATI, even when online edit checks are extensively employed. Clerical review is frequently necessary to examine changed or added name and address fields, to review "other specify" responses to ensure that they cannot be coded to existing categories, and to review interviewer notes which may require changes in respondents' recorded answers. By providing print-outs or displays which focus the clerical staff's activities on these specific tasks, CATI can expedite this review.

cattle inventory survey of farms. These data were obtained by using the same edit software on the two data sets. Table 4 on the previous page shows that CATI reduced the number of critical errors remaining after interviewing by 77 percent in comparison with non-CATI interviewing. Critical errors were those which, upon discovery, required another contact with the sample case. The reduction in noncritical errors was only three percent. On CATI, however, the vast majority of these noncritical "errors" had been verified correct by the respondent during the interview.

3.7. *Measurement of Interviewer Effects*

There are reasons to suspect that measurement errors associated with the interviewer may increase with CATI, and reasons to suspect that they may decrease. The computer has released the interviewer from burdens of implementing skip logic, but it has given the interviewer new burdens of entering data in a different way. CATI has replaced writing with typing. Item-based CATI systems focus the interviewer's attention on the next question to be asked but remove the context of prior and following questions provided by large paper questionnaires. Unfortunately there have been few studies that compare interviewer behavior during the questioning of a respondent with and without CATI.

Early descriptions of centralized telephone interviewing noted the possibility of measuring interviewer variance through randomized assignment of case to interviewers. Some examples of that facility exist (see, for example, Tucker (1983)). CATI offers a further advantage of such measurement, since the computer can handle the randomization of case assignment to interviewers. There was further speculation that the use of CATI applications would itself reduce inter-interviewer variability through forcing interviewers to adhere to a questioning sequence specified in the application. A contrasting hypothesis was that interviewer effects arise in the delivery of the question and self-initiated probing by interviewers. CATI can affect these behaviors chiefly through the implementation of controlled probing specified by the application. One test of CATI and non-CATI interviewing on the same questionnaire did not include those enhancements in the comparison. Groves and Magilavy (1986) found small tendencies to lower interviewer effects in the CATI-based estimates, but the differences were not statistically significant in a survey employing 33 interviewers.

3.8. *Gross Differences Between Survey Estimates from CATI and Non-CATI Methods*

The activities involved in collecting survey data can be disaggregated into hundreds of separate actions, each with its own potential for introducing error in the resultant data. There is little or no empirical evidence regarding errors for most of the steps involved in a CATI survey. Two controlled experiments found few if any differences in statistics calculated from a telephone survey using CATI and one done simultaneously with a paper questionnaire (Groves and Mathiowetz (1984), Tortora (1985)). It is to be expected that the results of such experiments are dependent on the nature of the questionnaire, the survey procedures (e.g., respondent rule, call scheduling), the experience of the interviewers with CATI and with the questionnaire, and the complexity of the questionnaire. Further, it is most likely, that findings of such experiments may change over time, as the field moves from the use of CATI to implement questionnaires designed for a paper technology, to those designed especially for CATI. At that point, the comparisons of CATI and non-CATI surveys on measure-

ments errors will include both the effects of flow control and edit checks *and* the effects of altered question format using the computer assistance.

4. Unique Features of CATI

There *do* exist some CATI features that have no simple paper analogue but on the surface offer increased accuracy of survey estimates or increased knowledge about their error properties. Dutka and Frankel (1980) describe uses of sequential sampling aided by CATI to achieve desired sampling precision of estimates. Such a design would use the CATI computer to calculate survey estimates from the set of interviews obtained from initial replicate subsamples in the survey data collection. Only with the existence of a machine readable data file could the designer compute up-to-the-minute survey estimates in order to guide the decision about adding more replicate samples. The addition of replicate samples would be guided by the error properties of early replicates and the targeted magnitude of error for the final estimates.

On the side of measurement error, survey organizations with CATI are increasingly using sequential pretesting procedures. These include video and audio monitoring of all cases by the questionnaire designers, examination of response distributions, immediate change of the CATI application when a problem is detected, further waves of testing, until they converge on an instrument that seems to be free of such problems. This method, in contrast to one or two pretests, each followed by a debriefing of pretest interviewers by the project staff, seems to offer more rapid development of question wording. It depends on the ability to make rapid changes to instruments.

Further, there are examples of CATI applications that have no feasible paper analogues. Smith and Smith (1980) and Fink (1983) describe the use of open questions containing dozens to hundreds of possible answers (e.g., "What is the make and model of your car(s)?") with answers that lead to different sequences of questions. The CATI application compares the text answer entered to large sets of prestored possible answers (using a "table lookup" procedure). This comparison is used to clarify imprecise answers (e.g., "a CHEVY but I don't know what model") by presenting examples of a general class (e.g., "CAPRICE, NOVA, BEL AIR"). The comparison can also be used to direct routing to specific questions on the set or an individual entity. It can also be used for consistency checks (e.g., "THERE IS NO FOUR CYLINDER MODEL IN A CHEVROLET CAPRICE. SEEK VERIFICATION OF RESPONSES."). It has been noted that this methodology has the general promise of improving the quality of open questions that are now coded after the survey. For example, occupation and industry measures in paper questions are compromises between the need to ask a large set of contingent questions and the cost and interviewer error associated with such sets. The commonly used two to four question sequence, for some percentage of workers, yields large difficulties for post survey coding. Instead, with an appropriate CATI application, each respondent might receive the set of questions necessary to identify his/her appropriate final occupational category.

Some panel surveys using paper questionnaires alert the interviewer or respondent to answers given in an earlier wave, but it is a rarely used design because of costs. Panel surveys on CATI are accessing prior wave data for verification, prompting of respondents regarding prior responses, or routing of the questionnaire to questions to be asked only of some respondents (Waksberg (1984)). The ability to use prior data in such a way offers to researchers measurement opportunities that were not seriously considered before (e.g.,

calculation and presentation of change in values for reassessment by the respondent – "Compared to what you mentioned last time we talked, it appears your housing costs have increased by $5 200 in the last year? Does that sound correct or would you think the change is different?")

5. Summary and Conclusions

At this time an evaluation of CATI's capability to affect data quality is limited by the tendency of most CATI applications to be simple translations of questionnaires that could have used or formerly used paper methods. In reviewing the speculations of CATI designers and the observations of CATI users, it seems clear that positive effects of CATI on data quality are most often unambiguous when changes to the questionnaire or survey procedures can be implemented with CATI. This generalization applies to the gains in call scheduling, video monitoring, improvement in reporting through enforced probing or feedback of arithmetic implications of prior answers, and the tailoring of questions to prior answers. Without such CATI features the inherent changes to the survey procedures from CATI are sometimes even viewed as having potentially harmful effects. For example, it is easier to strike a wrong key than to check a box erroneously, most interviewers can write faster than they can type, and it is easier to correct prior answers by turning a page than stepping back using function keys. Further, there remains great uncertainty about whether some of the CATI features that increase the time of application setup (and hence the costs of the CATI survey) but *do* change the survey procedure act to improve data quality sufficiently to justify their costs. These features include the use of fills for appropriate pronouns and verb cases in question wording.

Even for features for which strong prior arguments can be made about their potential to improve data quality, the field by and large has produced little evidence that those potentials have been achieved. While funding and energy has in the past been devoted to software design and hardware acquisition, CATI's future should contain careful experimentation regarding its unique contribution to control survey errors. This work, we believe, should use several methodologies:

a) *small scale, laboratory experiments* – Many of the questions regarding styles of screen and question presentation can be addressed in small laboratory studies in which interviewers are the subjects and the dependent variables are outcomes like proper question delivery, facility with moving backward in the questionnaire, etc. This is relatively cheap research and could be replicated across systems and organizations.
b) *trial and error development using ongoing surveys* – Other remaining questions regarding CATI are not sufficiently well framed to justify experimental designs. Call scheduling algorithms, for example, deserve trial and error adjustments until a small set of clearly distinct options emerge. Other topics like this include administrative procedures for the use of video monitoring, use of alternative probing procedures, procedures for resolving inconsistencies in respondent answers discovered by the CATI application.
c) *fully controlled experiments within large scale surveys* – If small scale experimentation cannot yield answers applicable to real survey situations, then controlled experiments within ongoing surveys are required. It is likely that such designs will be required for all issues involving the effects of CATI on the interviewer-respondent interaction. These tests would be run after initial trial and error evaluation yields a set of alternatives that show equal promise.

It is time, we believe, that CATI capture the interest of researchers devoted to learning how to alter survey procedures to measure and enhance data quality. Such efforts would mark a clear departure from the efforts in CATI thus far, concentrating on faithful representation of paper questionnaire procedures in a computer assisted medium. Only with such efforts will the survey community be

given evidence that the potential of CATI to enhance data quality can, in practice, be achieved.

6. References

Cannell, C.F., et al. (1983): An Experimental Comparison of Telephone and Personal Interview Health Surveys. National Center for Health Statistics, Washington, D.C.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981): Research on Interviewing Techniques. In Sociological Methodology, edited by S. Leinhardt, Jossey-Bass, pp. 389–437.

Coulter, R. (1985): A Comparison of CATI and Non-CATI on a Nebraska Hog Survey. Staff Report No. 85, Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

Dekker, F. and Dorn, P. (1984): Computer Assisted Telephonic Interviewing: A Research Project in the Netherlands. Paper presented at the Conference of the Institute of British Geographers, Durham, United Kingdom.

Dutka, S. and Frankel, L.R. (1980): Sequential Survey Design Through the Use of Computer Assisted Telephone Interviewing. Proceedings of the American Statistical Association, Section on Business and Economic Statistics, pp. 73–76.

Ferrari, P.W. (1984): Preliminary Results from the Evaluation of the CATI Test for the 1982 National Survey of Natural Scientists and Engineers. Unpublished research report. U.S. Bureau of the Census, Washington, D.C.

Ferrari, P.W. (1986): An Evaluation of Computer-Assisted Telephone Interviewing Used During the 1982 Census of Agriculture. Unpublished evaluation report, U.S. Census Bureau.

Fielder, E. (1985): Personal communication with the authors.

Fink, J.C. (1983): CATI's First Decade: The Chilton Experience. Sociological Methods & Research, 12(2), pp. 153–168.

Groves, R.M. (1983): Implications of CATI: Costs, Errors, and Organization of Telephone Survey Research. Sociological Methods & Research, 12(2), November, pp. 199–215.

Groves, R.M. and Mathiowetz, N.A. (1984): Computer Assisted Telephone Interviewing: Effects on Interviewers and Respondents. Public Opinion Quarterly, 48(1), pp. 356–369.

Groves, R.M. and Magilavy, L.J. (1986): Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys. Public Opinion Quarterly, 50(2), pp. 251–266.

Harlow, B.L., Rosenthal, J.F., and Ziegler, R.G. (1985): A Comparison of Computer-Assisted and Hard Copy Telephone Interviewing. American Journal of Epidemiology, 122(2), pp. 335–340.

Harris, M. (1952): Documents Used During the Selection and Training of Social Survey Interviewers and Selected Papers on Interviewers and Interviewing. Social Survey Division, Central Office of Information, Great Britain.

House, C.C. (1984): Computer Assisted Telephone Interviewing on Cattle Multiple Frame Survey. Staff Report No. 82, Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

House, C.C. (1985): Questionnaire Design with Computer Assisted Telephone Interviewing. Journal of Official Statistics, 1(2), pp. 209–219.

Mathiowetz, N. and Cannell, C. (1980): Coding Interviewer Behavior as a Method of Evaluating Performance. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 525–528.

Morton, B. and House, C.C. (1983): Training Interviewers for Computer-Assisted Telephone Interviewing. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 129–133.

Nicholls II, W.L. (1978): Experiences with CATI in a Large-Scale Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 9–17.

Palit, C.D. and Sharp, H. (1985): Personal correspondence to the authors in review of an earlier draft.

Press, G. (1985): Personal communication with the authors.

Presser, S. (1983): Discussion of papers by House and Morton and Morton and House. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 142–143.

Smith, R. and Smith, R. (1980): Evaluation and Enhancements of Computer Controlled Telephone Interviewing. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 513–515.

Survey Research Center (1983): CATI Application for Commercial Fleet Managers Survey. Institute for Social Research, The University of Michigan, Ann Arbor, MI.

Survey Research Center (1985): CATI Application for National Association of Business Economists Survey. Institute for Social Research, The University of Michigan, Ann Arbor, MI.

Tortora, R. (1985): CATI in an Agricultural Statistical Agency. Journal of Official Statistics, 1(3), pp. 301–314.

Tucker, C. (1983): Interviewer Effects in Telephone Surveys. Public Opinion Quarterly, 47, pp. 84–95.

Waksberg, J. (1984): Discussion of Nicholls and Groves paper. Proceedings of the International Statistical Institute, Amsterdam.

Received January 1986
Revised April 1986