Journal of Official Statistics
Vol. 8. No. 1, 1992, pp. 63–76

Developing Systematic Procedures for Monitoring in a Centralized Telephone Facility

Mick P. Couper[1], Lisa Holland[2], and Robert M. Groves[3]

Abstract: The ability to monitor the behavior of interviewers is a highly valued feature of centralized telephone interviewing. Despite the ubiquity of monitoring, current practices tend to be left to the discretion of supervisors and to be rather subjective observations. This paper describes one approach to systematizing the practice of monitoring. Key features of this monitoring scheme are the use of probability sampling methods and the collection of monitoring data at the question level. The results of one implementation of the scheme are presented.

Key words: Monitoring; interviewers; centralized telephone facilities; CATI.

1. Introduction

Monitoring in centralized telephone interviewing is the observation by a third party of the interaction between the survey interviewer and the respondent. At a minimum, the monitor listens to the audio communication through a telephone intercept of the interview interaction. In computer-assisted telephone interviewing (CATI) systems, monitoring also commonly includes the observation of a duplicate of the interviewer's terminal screen image, containing the current question and the respondent's answer as entered by the interviewer.

Monitoring has been described as one of the most important quality enhancement features of centralized interviewing (Dillman 1978; Groves and Kahn 1979; Mathiowetz and Cannell 1980; Oksenberg and Cannell 1988). When appropriately used, it permits reinforcement of training guidelines, presumably leading to a reduction in interviewer-induced error, both the bias and variance components. Monitoring is also believed to improve costs, productivity and efficiency of centralized telephone facilities (Lensen 1988).

Monitoring is part of standard practice in most centralized telephone facilities. In a study by Haggerty, Nicholls, and Dull

[1] Mick Couper is a Special Assistant, U.S. Bureau of the Census and Research Investigator, Survey Research Center, University of Michigan, Ann Arbor, MI, U.S.A.
[2] Lisa Holland is a Research Associate in the Field Section, Survey Research Center, University of Michigan, Ann Arbor, MI, U.S.A.
[3] Robert Groves is the Associate Director for Statistical Design, Methodology and Standards, U.S. Bureau of the Census and Professor and Research Scientist, Survey Research Center, University of Michigan, Ann Arbor, MI, U.S.A. Address correspondence to first author at: U.S. Bureau of the Census, Washington, DC 20233, U.S.A.

Acknowledgements: This research was conducted while all three authors were at the Survey Research Center (SRC) at the University of Michigan. The development of monitoring procedures was supported in part by the Bureau of Labor Statistics. The authors wish to thank Tim Beebe of SRC for his assistance in analyzing the monitoring data. The reviewers are also thanked for their valuable comments. Views expressed are those of the authors and do not necessarily reflect those of the U.S. Bureau of the Census.

(1989), 83% of survey organizations reported using some form of monitoring. However, practices and procedures varied widely across organizations. Current monitoring practices also tend to be unsystematic and subjective. Although many organizations monitor at a standard rate (a fraction of interviews or interviewing hours), the scheduling of monitoring shifts and the decisions of whom to monitor at what times are largely left to the discretion of the supervisors doing the monitoring. In addition, supervisors often record only general impressions of the interaction, rather than using objective measures defined to meet the intended purpose of monitoring. These variations within and across organizations may be one reason for the current dearth of evidence on the effectiveness of monitoring.

Such unstandardized and unsystematic practices do not adequately meet the stated goals of monitoring (the measurement and reduction of interviewer error). Some of the shortcomings of existing approaches are:

a. These procedures allow no generalizations to be made from the monitoring data. In any single monitoring session, there is no guarantee that what is being heard is representative of an interviewer's usual behavior, of all interviews in a particular study, or of all occurrences of a particular survey question.
b. Unstandardized procedures do not ensure that all interviewers are given an opportunity to be monitored. When interviewer selection is left to the discretion of the monitors, the potential for systematic biases is introduced.
c. Without a systematic selection procedure, it may be difficult to maintain control over when monitoring occurs. Supervisors may monitor only when they are not occupied with other tasks. If certain departures from prescribed interviewing behavior occur more frequently at these times, monitoring would disproportionately miss them.
d. Many organizations have different interviewers working on different studies. If monitoring procedures vary by study, then evaluations of interviewer behavior or performance across studies would be limited.
e. Evaluating interviewers on the basis of general impressions obtained by supervisors during periods of monitoring may lead to claims of discrimination or favoritism. Such general impressions do not provide evidence that can be queried or contested by an interviewer.
f. Monitoring based on general impressions of an interaction rather than question-level data focuses on interviewer behavior only. Interviewer behavior that results from a mistake in the instrument or difficulties on the part of the respondent might erroneously be attributed to deficiencies on the part of the interviewer.

The development of systematic monitoring procedures to overcome these deficiencies has two key elements: (a) the use of probability sampling to determine when and whom to monitor, and (b) the use of forms to facilitate the objective evaluation of interviewer behavior at the question level.

Much attention has been devoted to the measurement of interviewer behavior at the question level (see Cannell, Lawson, and Hausser 1975; Cannell and Oksenberg 1988; Sykes and Morton-Williams 1987). With few exceptions (see, for example, Mathiowetz and Cannell 1980) the coding schemes are designed for interaction coding of tape-recorded interviews rather than on-line monitoring of live interviews. The scheme

described here borrows heavily from the contributions of Cannell and others in terms of what behaviors ought to be monitored and how they should be measured.

By contrast, little attention has been paid to the development of sampling schemes to permit systematic monitoring. Chapman and Weinstein (1990) describe a sampling design tested in the U.S. Census Bureau's telephone facility. Two drawbacks of this design are (a) it assumes a constant rate of selection across shifts and interviewers and (b) interviewers are monitored for a set time regardless of their activity. This system was found to be too hard to administer operationally (Ferraiuolo, personal communication). For a sampling scheme to be effective, it should be easy enough to be used by facility staff and supervisors, yet flexible enough to meet multiple needs on a variety of studies.

This paper describes one approach at systematizing the practice of monitoring. This approach combines the use of probability sampling with the development of objective measures of interviewer behavior. Following a description of the design of the monitoring scheme, the results of an early implementation in the Survey Research Center's (SRC) telephone facility are presented. Interviewers working on a single study were monitored for a period of one month using the scheme described here. Although a number of design options are mentioned, this paper reports on only one particular implementation of the scheme. The system can be easily adapted to meet a variety of needs.

2. Design of Systematic Monitoring

Monitoring designs consist of a number of separate steps. These are: (a) selection of shifts, interviewers and interactions to monitor, (b) data collection by the monitor, and (c) provision of feedback to interviewers and facility management. Each of these will be discussed in turn.

2.1. Selection

One conceptualization of sampling for monitoring is that of selection from a three-dimensional space. The three dimensions are: (a) when to monitor (shifts during which monitors will work), (b) whom to monitor (which interviewers should be monitored), and (c) what to monitor (what interviewer activities are eligible for monitoring). The three dimensions of time, interviewer and activity may represent strata in the selection scheme. Different probabilities of selection can be assigned to any or all of these dimensions to meet various needs.

To illustrate these selection dimensions, monitors are often directed to give greater attention to new interviewers, interviewers who exhibited problems in an earlier monitoring session, or interviewers with low response rates. Managers also have attempted to devote more monitoring resources to the earlier parts of a survey period in order to have time to correct errors that are found. At the extreme, this may eliminate monitoring of certain interviewers or certain times.

The design outlined here uses disproportionate selection of interviewers, reflecting current practices of concentrating monitoring efforts on less experienced interviewers. Three groups of interviewers are identified, based on performance ratings and experience. The first group (Strong) consists of those interviewers who have more than six months interviewing experience at the SRC facility and have been rated above the median on a scale of performance scores. The second group (Average) consists of the remainder of those with at least six months experience. The last group (New) consists of those having less than six months experience at the facility.

In addition to these aspects of the sample design, decisions need to be made regarding such issues as the number of monitoring shifts to schedule, the length of each shift and session, and the determination of eligible behaviors.

a. Number of monitoring shifts

A monitoring "shift" refers to a period of time that one person is scheduled to monitor interviewers. Many factors influence the number of monitoring shifts to schedule during an interviewing period. These include costs and staffing considerations, the minimum number of times desired for an interviewer to be monitored, and the length of time the interviewer is monitored.

To determine the number of shifts to schedule, assumptions also need to be made about monitor productivity (how much time is spent actually listening to an interview) and interviewer productivity (how much time is spent interacting with respondents). In this test, shifts were selected to reflect the time supervisors currently spend monitoring (approximately 5% of all interviewer hours).

b. Length of monitoring shifts

The appropriate length of each scheduled monitoring shift is largely a function of setup time and monitoring burden. "Setup time" refers to the time it takes a monitor to prepare for a shift. "Monitoring burden" is the amount and complexity of monitoring activity required of the monitor during the shift.

The longer the time a monitor needs to prepare for a shift, the more cost effective it is to have longer shifts. However, the longer the shift, the more tiring and burdensome the task becomes. Judgments from experienced monitors suggest that shifts of less than one hour are too short, but that monitoring for more than two hours at a time is too fatiguing. For this reason, monitoring shifts of one hour were used.

c. Selection of monitoring shifts

Monitoring shifts were selected from weekly schedules of interviewing shifts using probability sampling. Monitoring shifts were selected proportional to the weighted number of interviewers scheduled to work during each hour. Interviewers were stratified into the above three groups based on their past performance. Relative selection probabilities in the ratio of 1 : 1.5 : 2 were assigned to the Strong : Average : New interviewer groups. In this way, New interviewers were to be selected at twice the rate of Strong interviewers. The monitoring shifts selected were administered by the supervisor working in the facility at the designated times.

d. Length of monitoring sessions

A monitoring "session" is defined as the period of time that a monitor listens to a particular interaction between interviewer and respondent. A session could be a full interview or a partial interview. It may be a prespecified period of time, a given number of questions, or may vary according to the length of the interview.

Determining optimal session length depends on various factors. Long monitoring sessions allow the monitor to experience more of the interaction between the interviewer and the respondent. With such sessions, the monitor may gain evidence that interviewer behavior not consistent with training guidelines may actually be caused by weaknesses in the instrument or unusual behavior on the part of the respondent. Conversely, short monitoring sessions permit larger numbers of different interviews to be monitored, albeit with less detail on each interview. Monitoring results based on many cases of each interviewer provide a more stable mix of difficult and easy cases. Monitoring a portion of an interview also allows the monitor to select any interview in progress, rather than wait for one that is

about to begin. This is expected to reduce the amount of time spent determining eligibility, and thus increase the productivity of the monitor.

In earlier tests of the monitoring scheme, a time limit of five minutes per session was used. However, in practice, monitors found it difficult to keep track of the time. It was thus decided to monitor a fixed number of questions. A maximum of 20 questions were monitored in each session.

e. Determination of eligible behaviors

Candidates for what behaviors to monitor include interviewer activity between dialings, dialing behavior, initial contact behavior, interviewing behavior, and post-interview editing behavior. Which behaviors should be monitored depends on the goals of the monitoring. Targeting some of the behaviors requires the ability to know what different interviewers are doing at a specific point in time. This seems to vary according to the telephone technology available to the facility.

The rules for determining eligibility will also affect the frequency with which certain portions of the interaction (such as introductions and conclusions) are monitored, as well as the efficiency of the monitor.

In an effort to target interviewers' deliveries of introductions to the survey, cases in which the respondent's telephone was still ringing were considered eligible. This obviously meant that a number of selections resulted in ineligible behaviors (no answer, wrong number, etc.). In this case, the monitor simply made another selection, and thus experienced lower productivity. Although introductions were monitored in this test, the results will not be reported here.

f. Selection of sessions

The selection of sessions was also done using probability sampling. For the current test, equal probabilities of selection were given to all interviewers within shifts.

Monitors selected sessions using a computer program which generated a random selection of eligible interviewer stations with each key press. Once a station was selected, the monitor determined whether the interviewer was engaged in an eligible activity. If so, the case was monitored for the required 20 questions (or until the end of that interview). If the interviewer's activity was deemed ineligible for monitoring, further selections were made until an eligible case was obtained. Under this scheme an interviewer could be selected more than once in a monitoring shift.

g. Selection probabilities

Using the scheme outlined above, the overall probability of a particular case being monitored can be determined. This is approximately

$$S_\alpha \cdot \frac{NM_\beta}{M_\alpha} \cdot \frac{A_\beta}{T_\beta} \cdot \frac{R_\gamma}{I_\beta}$$

where

S_α = the number of monitoring shifts scheduled for the αth selection period (week)

NM_β = The measure of monitoring need for the βth monitoring shift (weighted number of interviewers in that shift)

M_α = the cumulative measure of monitoring need in the αth selection period

A_β = the number of interviewer selection attempts made during the βth shift

T_β = the number of potential interviewer selections during the βth shift

R_γ = the weight (based on the performance rating) assigned to the γth interviewer in the βth monitoring shift

I_β = the total of the interviewer weights of those working during the βth shift.

An example will illustrate the probability of selection for an interviewer in a particular shift. Suppose that 25 monitoring shifts were scheduled for a particular week (S_α). The measure of monitoring need (NM_β) for a particular interviewing shift is calculated by taking the weighted sum of the interviewers scheduled to work that shift:

Number of interviewers	Weight	
4 Strong interviewers	× 1.0	= 4.0
4 Average interviewers	× 1.5	= 6.0
3 New interviewers	× 2.0	= 6.0
		16.0

This measure of need was calculated for each interviewing shift in the week. The sum of these measures (M_α) is 612.5, and is the denominator of this term. This term is the probability that the particular shift would be selected for monitoring. The next term is the ratio of selection attempts made in the shift over possible selections. This can be approximated by assuming that one selection per second is made. Therefore, if 35 minutes (35 × 60 = 2100 seconds) were spent monitoring out of the 60 minutes allotted, then 1500 seconds (3600–2100) were spent selecting cases. The final term in the equation is based on the interviewers who actually worked that particular shift, and reflects the probability of a single interviewer being selected. Assuming that all scheduled interviewers worked the shift, I_β is 16. The probability of a specific Average interviewer being selected is then 1.5/16 = 0.094. The weights used in this last expression could be the same as or different to those used to calculate the measure of monitoring need. The full expression thus becomes

$$25 \cdot \frac{16}{612.5} \cdot \frac{1500}{3600} \cdot \frac{1.5}{16} = 0.0255$$

In this test, equal probabilities of selection were assigned to all interviewers within a particular shift. Thus R_γ/I_β is constant within shifts. The inverse of the selection probabilities (in the above example, 1/0.0255 = 39.2) are then assigned as selection weights for the analysis of the monitoring data.

2.2. *Measurement*

Measurement procedures used in monitoring vary across three dimensions. First, the measurement unit can vary. Monitors can record data at the level of a single set of utterances made by a participant in the interaction, at the level of a question and set of response categories, or at the level of the entire interview. Second, data recorded by the monitor may be subjective or objective in nature. That is, monitors may record ratings of the interviewer's performance, about their tone of voice, or about the clarification the interviewer provides the respondent. In contrast, the monitor may record the behavior of the participants using only objective criteria, such as whether the respondent interrupts the interviewer or whether the interviewer reads the question exactly as it is worded. Finally, some monitoring designs record only the behavior of the interviewer while others collect information about both interviewer and respondent behavior.

Multiple iterations of informal tests were run in order to refine monitoring measurement specifications and to develop the forms on which monitoring data would be recorded. The collection of data at the question level raises the amount of information collected, while still maintaining a feasible task for the monitors. Furthermore, the present design primarily collects objective measures of behavior which are more easily quantified. Informal tests of measurement procedures demonstrated the monitors' difficulty in

simultaneously recording a large number of interviewer and respondent behaviors. The number of codes was therefore reduced by excluding all respondent behaviors and those interviewer behaviors that were found to be rare. Part of the form used in the present study is reproduced in Appendix A.

The measurement specifications determine what codes will be included in the monitoring form. If the focus of monitoring is evaluation of interviewer performance, the form should be designed to record interviewer behaviors including question-asking, probing and feedback. Other considerations also play a role in the development of monitoring forms. The pace of survey interviews prohibits the use of complicated forms. The form should have a minimal number of pages, be easy to read, and include all necessary definitions of the codes. Once the monitor is trained, he or she should not have to rely on references other than the form while monitoring. The form should be versatile enough to monitor multiple studies (whether CATI or paper-and-pencil) in a single facility. Finally, the form should be organized such that data collection can be done directly from the form without any additional coding or editing, to permit the most timely use of monitoring data.

2.3. Feedback and reporting

A key factor in the administrative utility of monitoring lies in the feedback that is given to interviewers. Such feedback can take two forms: (a) immediate feedback during or after a monitoring shift, and (b) routine periodic feedback using a standardized report form. Immediate feedback after every monitoring session would provide the monitor with the opportunity to explain all errors and praise all successes to the interviewer in detail, and allow for timely intervention in the case of critical errors or inappropriate behavior. Immediate feedback may be verbal or written. However, such feedback is costly (in terms of both interviewer and supervisor time). In addition, it is based on single case samples and thus tends to be highly variable, with atypical results given as much weight as typical. Nevertheless, both supervisors and interviewers value this form of feedback, and it was provided where appropriate. In addition, routine feedback based on cumulated results of multiple monitoring sessions provides the interviewer with a record of his or her performance over time and relative to others in the facility.

3. Implementation

A test of the monitoring scheme was conducted in the SRC telephone facility during September 1990. A single study, the Survey of Consumer Attitudes, was used for this test. This study was chosen for several reasons. It was a paper-and-pencil study, which was necessary as the facility was not then equipped to monitor all CATI interviewing from a single monitoring station. Furthermore, the design needed to be versatile enough to be effective for both CATI and non-CATI studies. The study is also conducted every month. This provides a basis for comparing monitoring data collected in subsequent months. The survey instrument was also familiar to both interviewers and supervisors.

A total of 25 interviewers worked on this study in September. These interviewers were classified into three groups for monitoring purposes, resulting in 8 Strong, 10 Average and 8 New interviewers. Six telephone facility supervisors and three methodologists monitored selected shifts over a period of four weeks. All were fully trained in the use of the monitoring forms and the selection software prior to the start of the test.

There were certain difficulties associated with the implementation of this procedure. Initially, some supervisors had difficulty with the notion of probability sampling, and did not always see the importance of monitoring during the selected shifts. As a result, they would often substitute a more convenient hour for that selected. Monitors also expressed frustration when the same interviewer was selected two or three times in succession within a shift. This was particularly evident when the number of interviewers in a shift was small.

It took several iterations of testing to adequately define eligible interviewing shifts. In the current design, the first hour of interviewing each day and shift changes were omitted because supervisors were occupied with checking in interviewers and assigning work. The last hour of each day was also omitted because it did not allow sufficient time for supervisors to provide feedback to interviewers at the completion of the monitoring shift. One way to reduce the number of ineligible times would be to start monitoring shifts on the half-hour. Shifts with three or fewer interviewers scheduled to work were also excluded because they were an inefficient use of supervisors' time.

There were initial objections to the fact that the monitoring form recorded only instances of negative behaviors or errors. Assuming that interviewer errors occur infrequently, recording negative behaviors is less burdensome on monitors, allowing more detailed data to be collected. However, it is important that the results of monitoring be presented to interviewers in a positive manner.

Despite these difficulties, the monitoring scheme has been accepted by supervisors and interviewers. The scheme described here has been used (with minor modifications) on multiple studies in the SRC facility continuously since this test.

4. Results

In examining the success of the monitoring scheme applied to this study, both monitoring productivity and interviewer performance are of interest. Each of these will be discussed in turn. The question-level data presented below are weighted to reflect the differential probabilities of selection of the three interviewer groups across weeks and shifts. The productivity data are unweighted.

4.1. Monitor productivity

A total of 49 hours were spent monitoring over the course of this test. In this time a total of 201 monitoring sessions were conducted, with an average of 18.6 questions monitored per session. A summary of the work done over the four weeks is presented in Table 1.

Monitoring productivity declined from the first to the last week of this study. This may reflect the fact that fewer interviews were completed towards the end of the data collection period. Furthermore, fewer interviewers worked in the last week than in the first, thus increasing the likelihood of having no interviewers engaged in eligible behaviors at certain times during the monitoring shift and reducing the productivity of the monitors.

The success of the stratified selection scheme in ensuring that New interviewers are monitored more than Strong interviewers should also be evaluated. Data to address this issue are presented in Table 2. The average number of hours worked are based on all interviewer hours charged to this project, and are an overestimate of the time spent interviewing.

Average interviewers were monitored at a lower rate than Strong interviewers. This is

Table 1. Monitor productivity by week

	Week 1	Week 2	Week 3	Week 4	Total
Number of hours monitored	3	18	20	8	49
Number of sessions monitored	14	71	88	28	201
Number of questions monitored	272	1,304	1,651	523	3,750
Average number of sessions per hour monitored	4.7	3.9	4.4	3.5	4.1
Average number of questions per hour monitored	90.7	72.4	82.6	65.4	76.6

contrary to expectation given the differential rates of selection. One reason for this discrepancy may be different levels of productivity across the three groups. If Strong interviewers are more productive than Average interviewers (i.e., they have a higher proportion of time engaged in eligible activities), they will be selected at a higher rate. In fact, Strong interviewers completed an average of 1.08 interviews per hour worked, compared to 0.80 for Average interviewers and 0.76 for New interviewers. It is obvious that the relative selection probabilities for the three groups need to be adjusted to take differences in productivity into account. These rates were adjusted in later applications to yield desired rates of monitoring.

How successful was the monitoring scheme in ensuring that each interviewer was monitored a sufficient number of times over the course of the study? It was seen in Table 2 that interviewers were monitored an average of 8.1 times each. Each of the 25 interviewers was monitored at least once. On average, Strong interviewers were monitored once for every 5.2 hours they worked, Average interviewers once for every 5.5 hours worked and New interviewers once for every 4.0 hours worked.

Using the current selection scheme and level of monitoring effort, this suggests that to have an interviewer monitored five times in a reporting period (e.g., a month), he or she would need to work at least 27 hours in that time. These numbers can be used to determine the level of monitoring that is required to ensure that sufficient data are obtained for each interviewer.

4.2. Behaviors monitored

It was noted earlier that a total of 3,750 questions were monitored during this study. The number and percentage of each type of error recorded are presented in Table 3. The

Table 2. Interviewer hours and monitored sessions

	Strong	Average	New	Total
Number of interviewers	8	9	8	25
Average number of hours worked	47.7	34.5	35.1	38.9
Average number of sessions monitored	8.9	6.2	9.1	8.1
Average number of questions monitored	165.3	115.6	170.1	151.3
Average number of questions monitored per hour worked	3.46	3.35	4.85	3.83

Table 3. Error rates by interviewer category in percent

	Strong	Average	New	Total
Number of questions monitored	(1,349)	(1,040)	(1,361)	(3,750)
Question asking				
Minor wording changes	3.3	6.8	3.4	4.2
Major wording changes	1.3	0.7	2.3	1.6
Incomplete questions	0.2	0.8	3.2	1.6
Skip errors	0.5	1.0	0.4	0.6
Repetition of question				
Inappropriate	0.4	1.3	1.0	0.9
Failures to repeat	0.1	0.2	0.4	0.3
Definitions/clarifications				
Inappropriate	0.2	0.5	0.8	0.5
Failures to provide	0.0	0.2	0.0	0.1
Probing				
Inappropriate	1.2	1.5	1.7	1.5
Directive probing	0.2	0.4	1.2	0.7
Failure to probe	1.2	1.2	1.0	1.1
Over-probing	0.0	0.2	0.9	0.4
Feedback				
(Feedback provided)	(39.9)	(39.3)	(30.2)	(35.7)
Inappropriate feedback	1.6	2.9	1.9	2.0
Directive feedback	1.1	2.6	0.6	1.2

error rates are simply the number of questions with errors expressed as a percentage of all questions monitored. Minor wording changes contribute the greatest proportion of errors detected. The rates at which other errors are committed are low.

How effective is the monitoring scheme in distinguishing among interviewer categories? It appears that New interviewers are more likely to make major wording changes, less likely to read each question in its entirety, and less likely to provide feedback to respondents. However, Average interviewers appear to commit more minor wording errors than either New or Strong interviewers. It should be noted that the allocation of interviewers to these three groups was a somewhat arbitrary process based on supervisors' subjective judgements of interviewer performance. One advantage of this scheme is that the results of initial monitoring can be used to obtain a more objective classification of interviewers in subsequent months, rather than relying on such subjective classification by supervisors.

An examination of error rates over the four weeks of the study reveals no consistent trends. It appears that feedback is given more in weeks 1 and 4, more minor wording changes occur in week 2 than the other three weeks, and major wording changes decline over the four weeks (from 2.2% in week 1 to 1.3% in week 4). However, none of these differences are statistically significant. (Simple random sampling variance estimators were used throughout; later analyses will account for clustering effects.) The lack of differences in error rates over time may be due to the fact that this is a stable, ongoing survey. Interviewers are familiar with the

instrument, and do not change their behavior from week to week. Furthermore, feedback was provided to interviewers during the course of this test. This may have prevented any increases in errors made over time. The examination of trends in error rates from one month to another for an ongoing study of this nature may be more revealing.

Question-level monitoring also provides some information about which questions may be causing problems for interviewers and respondents. Four questions are selected to illustrate this point.

The wording of these four questions is as follows:

Question A2A: "Now looking ahead – do you think that *a year from now* you (and your family living there) will be *better off* financially, or *worse off*, or just about the same as now?

Question B1: Do you (and your family living there) own your home, pay rent, or what?

Question C1: Because automobiles are an important purchase for individual families and an important part of the entire economy, I would like to ask some specific questions about them. First, do you (or anyone in your family living there) own a car, pickup, van, jeep, suburban, blazer-type vehicle or motorhome?

Question E8: Would you mind telling me your race or ethnic background. Are you white, black, Hispanic, American Indian or Alaskan native, Asian or Pacific Islander?

Selected error rates for these questions are presented in Table 4.

It is clear that the types of errors made by interviewers vary according to the questions being asked. For the race question (E8), minor or major wording changes are made 20% of the time, compared to only 3% of the time for the rent question. Question A2A produced a number of problems of wording, clarification, probing and feedback. These data may be used to suggest changes to question wording, the provision of appropriate definitions or suggestions for probing, or further training focusing on

Table 4. Error rates for four questions (percent)

	Question A2A	Question B1	Question C1	Question E8
Number of times monitored	(28)	(29)	(30)	(29)
Question asking				
Minor wording changes	4.2	3.3	7.7	8.2
Major wording changes	11.6	0.0	2.8	0.0
Incomplete questions	0.0	0.0	5.5	12.3
Definitions/clarifications				
Incorrect	11.6	0.0	0.0	0.0
Probing				
Inappropriate	7.4	0.0	0.0	5.9
Directive probing	8.8	0.0	0.0	1.1
Feedback				
(Feedback provided)	(60.2)	(33.3)	(44.2)	(27.2)
Inappropriate feedback	5.1	0.0	6.1	5.1
Directive feedback	4.2	0.0	0.0	0.0

particular questions or specific areas of interviewer behavior.

5. Conclusion and Discussion

It is difficult to evaluate the effectiveness of this monitoring scheme relative to existing approaches, as such approaches tend to be subjective and unsystematic in nature, and do not produce quantifiable results. However, discussions with all levels of facility staff suggest that the system described here is positively received. Interviewers generally feel that it introduces greater fairness in monitoring and evaluation. Supervisors feel that it improves their efficiency and reduces subjectivity. Managers value the data produced for evaluating performance in the facility.

Based on the results of this test, it appears that a systematic monitoring scheme is both feasible and useful. The benefits of such a system may lie not only in the improved quality of the data obtained from monitoring, but also in increased monitoring efficiency resulting from computer assistance and the use of systematic procedures. As shown here, the data produced by systematic monitoring may be used for a number of different purposes.

A number of enhancements are being made to the scheme. Work is being done to refine the selection process, particularly the selection probabilities across groups and the number of monitoring shifts to schedule. The development of summary monthly and quarterly reports, and the integration of the monitoring process into the overall evaluation of interviewers is also planned.

One of the keys to the successful implementation of a systematic monitoring scheme is the use of computers to assist in the various monitoring tasks (particularly shift and session selection), and thereby reduce the burden on supervisors. Current work is focusing on the feasibility of using computers to facilitate other stages of the monitoring task, such as the increased automation of shift and session selection, the recording of monitoring data online, the output of standardized reports, and the integration of monitoring data into the routine evaluation of facility and interviewer performance.

As a result of the success of this test and the positive feedback received by facility staff, these procedures have been incorporated into the ongoing activity of the SRC telephone facility, and expanded to include all studies, both CATI and non-CATI. The goal is the application of statistical quality control principles and practices to the process of centralized telephone interviewing.

6. References

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). A Technique for Evaluating Interviewer Performance. Ann Arbor, MI: Survey Research Center.

Cannell, C.F. and Oksenberg, L. (1988). Observation of Behaviors in Telephone Interviews. In Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls II, W.L. and Waksberg, J. (eds.), Telephone Survey Methodology. New York: John Wiley.

Chapman, D.W. and Weinstein, R.B. (1990). Sampling Design for a Monitoring Plan for CATI Interviewing. Journal of Official Statistics, 6, 205–211.

Dillman, D.A. (1978). Mail and Telephone Surveys. New York: John Wiley.

Groves, R.M. and Kahn, R.L. (1979). Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Haggerty, C., Nicholls II, W.L., and Dull,

V. (1989). Monitoring Practice for Telephone Surveys. Chicago: NORC (unpublished paper).

Lensen, S. (1988). Telephone Monitoring. Ottawa: Statistics Canada (unpublished paper).

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

Oksenberg, L. and Cannell, C.F. (1988). Monitoring Behavior in a Survey Pretest Interview as a Technique for Identifying Problems with Questions and a Guide for Improvement. Ann Arbor, MI: Survey Research Center (unpublished paper).

Skyes, W. and Morton-Williams, J. (1987). Evaluating Survey Questions. Journal of Official Statistics, 3, 191–207.

Received January 1991
Revised January 1992

APPENDIX A
EXAMPLE OF MONITORING FORM

QUESTION #	INITIAL QUESTION ASKING				REPEAT QUESTION		DEFINE/ CLARIFY		PROBING				FEEDBACK GIVEN			CATI RECORD	COMMENTS
1.	Minor	Major	Incomp.	Skip	I	F	I	F	I	D	F	O	YES	I	D	R	
2.	Minor	Major	Incomp.	Skip	I	F	I	F	I	D	F	O	YES	I	D	R	
3.	Minor	Major	Incomp.	Skip	I	F	I	F	I	D	F	O	YES	I	D	R	
4.	Minor	Major	Incomp.	Skip	I	F	I	F	I	D	F	O	YES	I	D	R	
5.	Minor	Major	Incomp.	Skip	I	F	I	F	I	D	F	O	YES	I	D	R	

KEY TO MONITORED BEHAVIOR:

Initial question asking:
- Minor wording change
- Major wording change
- Incomplete questions
- Incorrect skip

Repeat question:
- Inappropriate repetition
- Failure to repeat question

Define/clarify:
- Inappropriate definition/clarification
- Failure to define/clarify

Probing:
- Inappropriate probing
- Directive/evaluative probing
- Failure to probe
- Over-probing

Feedback given:
- Was feedback provided
- Inappropriate feedback
- Directive/evaluative feedback

CATI record:
- Was response recorded correctly