

Session Number: Session 2B

Session Title: *Improving Estimates from Survey Data*

Session Organizer(s): Stephen Jenkins, University of Essex, Colchester, UK, and Holly Sutherland, University of Essex, Colchester, UK

Session Chair: Holly Sutherland, University of Essex, Colchester, UK

*Paper Prepared for the 29th General Conference of
The International Association for Research in Income and Wealth*

Joensuu, Finland, August 20 – 26, 2006

**WHO'S ASKING? INTERVIEWERS, THEIR INCENTIVES, AND DATA
QUALITY IN FIELD SURVEYS**

Arthur B. Kennickell

For additional information please contact:

Author Name(s) : Arthur B. Kennickell

Author Address(es) : Mail Stop 153, Federal Reserve Board, Washington, DC 20551,
USA

Author E-Mail(s) : Arthur.Kennickell@frb.gov

Author FAX(es) : (202) 452-5295

Author Telephone(s) : (202) 452-2247

This paper is posted on the following websites: <http://www.iariw.org>

Who's asking?
Interviewers, Their Incentives, and Data Quality in Field Surveys

Arthur B. Kennickell
Senior Economist and Project Director
Survey of Consumer Finances
Mail Stop 153
Federal Reserve Board
Washington, DC 20551
Phone: (202) 452-2247
Fax: (202) 452-5295
Email: Arthur.Kennickell@frb.gov

SCF Web Site: <http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html>

May 31, 2006

Views expressed in this paper are those of the author and do not necessarily represent those of the Board of Governors of the Federal Reserve System or its staff. The author thanks Leslie Athey and other Central Office staff at NORC and the field managers and interviewers for the 2004 Survey of Consumer Finances. The author is also grateful to his SCF colleagues at the Federal Reserve Board, particularly Brian Bucks, Gerhard Fries, and Kevin Moore.

In all surveys except self-administered ones, interviewers have a critical role in representing the study to respondents. Because most of their work is difficult or impossible to observe directly, historically research has largely been obliged to deal with “interviewer effects” in data collection as a type of statistical artifact, rather than addressing interviewers’ incentives and resulting behavior (see Groves [1989] for references). This paper focuses on two sets of behavioral decisions made by interviewers in field surveys that have serious implications for the quality of the data collected.

First, there is the decision to apply effort to convince respondents to participate in a survey. Because interviewers typically face strong pressures to complete interviews while minimizing their expenses in doing so, they are implicitly guided to apply effort to cases they believe are most likely to be completed. If interviewers’ expectations are unbiased and respondents are not entirely immune to persuasion, such behavior will tend to amplify the patterns of nonresponse that would have arisen from respondents who were faced by a more neutral approach. To the degree that an interviewer’s expectations are biased, such behavior may induce patterns of nonresponse that do not reflect respondents’ behavior at all. Variation in the level of persuasive skills possessed by interviewers is clearly also an important factor, but the continuing selection of interviewers according to their completion rates should tend to dampen such differences over time. There is no comparable “natural selection” that takes place to ensure an appropriate balance of effort for all types of cases.

Second, there is a potentially large set of decisions interviewers make during the actual administration of an interview. An interviewer’s decisions whether to follow the interview protocol, to probe ambiguous responses, to supply important auxiliary information, to signal to

the respondent the importance of coherent reporting, to support the respondent's confidence in the confidentiality of the information, etc. are critical determinants of data quality. But outside of experimental settings, it is rare that anything about data quality is known directly, aside from fairly coarse information. Most often, the only readily available information is whether a case was completed or it was not. In such cases, whatever selection over interviewers that does take place via performance evaluations over the field period is in a dimension not necessarily correlated with skill in collecting good information; indeed, earlier analysis of metadata from the U.S. Survey of Consumer Finances (SCF) suggests that there is little or no correlation between completion rates and data quality (Kennickell [2002]).

This paper discusses the incentive structure underlying interviewers' behavioral choices in these two key areas and it presents information on the structures created for the 2004 SCF in an attempt to direct interviewers toward performance more desirable from the standpoint of the ultimate data users. The first section of the paper discusses the general role of field interviewers in data collection and provides a very simple behavioral model. The second section provides brief background on the SCF. The third section considers the actions of interviewers during an interview and their implications for data quality. The next section examines efforts undertaken to distinguish potential selection effects in attempts to gain the cooperation of respondents to administer an interview. The final section concludes the paper and points to the next steps.

I. The Role of Interviewers

To clarify the incentives interviewers face in a field survey, consider the following very simple behavioral model. Suppose interviewer i gains positive utility (U) from the consumption of some good (x_i) and disutility from expending effort (E_i) in working a survey case (the case-

level subscripts are suppressed). Effort is the sum of work e_{ci} directed specifically toward areas that have an observable outcome—referred to as “case completion”—and work e_{qi} directed specifically areas that have no component observable to survey managers—referred to as “data quality.” Survey managers are assumed to desire a level of effort Q^* directed toward quality, but because only case completion is observed directly, compensation can only be based on case completion. To avoid needless technical complications, assume that interviewers are paid a fixed level of compensation (w_i) for completing a case, and that this compensation is used directly to purchase x_i . A case is completed when the persuasive input from interviewer j to the respondent (C_{ij}) exceeds a threshold value C_j^* . This input depends directly on the work directly aimed at completing the case and indirectly and less strongly on efforts toward data quality, which are perceived by the respondent only as a secondary motivating factor in enhancing the credibility of the interviewer. Thus, let $C_{ij} = C(e_{ij}^c, e_{ij}^q, Z_j)$, where

$$C(e_{ij}^c, 0, Z_j) \gg C(0, e_{ij}^q, Z_j) \text{ and } \frac{\partial C}{\partial e_{ij}^c} \geq \frac{\partial C}{\partial e_{ij}^q} \text{ for all } i \text{ and } j, \text{ and } Z_j \text{ is a set of characteristics of the}$$

respondent observed by the interviewer. Let $E_{ij}^* = \{(e_{ij}^c, e_{ij}^q) \mid C(e_{ij}^c, e_{ij}^q, Z_j) = C_j^*\}$, the set of pairs of types of work for which the total effort is exactly sufficient for the interview to be completed.

The formal problem an interviewer faces is: $\max_{x_i, -E_i} U(x_i, E_i)$ subject to $x_i = \sum_j w_i \tilde{c}_{ij}$, where

$\tilde{c}_{ij} = 1$ when $E_{ij} \in E_{ij}^*$ and equals zero otherwise; total effort is $E_i = \sum_j E_{ij}$.

If there are any values of E_i^* such that $U\left(x_i, \sum_j (e_{ij}^c + e_{ij}^q)\right) > U(0,0)$, then the interviewer will

choose the minimum values of $e_{ij}^c + e_{ij}^q$ in ascending order over cases until the gain from an

additional case is just offset by the amount of effort—that is, $\frac{\partial U(.)}{\partial x_i} = \frac{\partial U(.)}{\partial E_i}$. By construction,

effort directed toward quality always has less return for case completion than effort directed toward case completion. Thus, the rational self-interested interviewer in this models selects

$e_{ij}^q = 0$ and the minimum value of e_{ij}^c from the set $E_{ij}^*(e_{ij}^c, 0)$. In this model, one would have to

allow for an effective tradeoff between e_{ij}^c and e_{ij}^q so that an interviewer would be able to select a

non-corner solution for e_{ij}^q and e_{ij}^c , where $\frac{\partial C}{\partial e_{ij}^c} = \frac{\partial C}{\partial e_{ij}^q}$. Of course, there would still be no

mechanism to ensure that $\sum_{ij} e_{ij}^q = Q^*$. Note that even if the survey managers raised the amount

of w , the outcome would not be changed, because, again, there is no incentive mechanism based on case completion to measure and reward efforts directed toward quality.

This model omits many factors that are important in the practical work of interviewers. Some interviewers might gain pleasure from the act of producing high-quality data, and the comfortable style of interaction for some interviewers may include actions that tend to promote data quality more than is the case for other styles. The model also ignores that fact that in real interviews there is uncertainty about the outcome of a case, the fact that interviewers are normally paid for the time they work, not the number of completed cases, and the fact that some interviewers have a natural drive toward perfection, at least as they see it. Nonetheless, to the

degree that facilities for monitoring of quality of performance are absent, there is no means of reinforcing the quality-oriented behavior that is desired by survey managers or of punishing behavior that works against data quality. Thus, if effort directed toward quality is costly to interviewers, data quality is generally likely to be below the optimal level and there will be variations in quality across cases that reflect a mixture of problems related to interviewers and those more specific to respondents. Monitoring and feedback in some form are the only effective means of creating the proper incentives to steer interviewers to the desired outcome.

In many surveys, there is at least a limited ability to monitor some aspects of interviewers' behavior that reflect on the quality of the data. Because interviewers in social science research are not normally paid for their time, not for case completion, managers typically monitor the amount of time devoted to cases overall, and sometimes individual interviewers are asked to report on the details of the approaches they have taken to secure the cooperation of particular respondents. In instances where there are electronic call records of attempts to work a case, there is also the ability to monitor one dimension of effort at the level of individual cases. In such cases, managers have the opportunity to quiz interviewers about how they are distributing their work and to ascertain whether there might be systematic aversion on the part of interviewers to certain cases. However, even if a manager had sufficient time to worry about interviewers' assignments at the case level, it would be a major intellectual and technical achievement just to organize the available information on their own. As discussed later in this paper, there are tools that can make this job of monitoring easier and that also help with potential problems of systematic bias induced by interviewers' choices in the application of effort to cases.

In most serious field surveys, a selection of each interviewer's interviews is made and those cases are re-contacted with the aim of verifying that the interviews actually took place. Typically, such validation exercises do not enquire about values actually reported other than those for simple demographic variables necessary to ensure that the correct person is contacted. This is a very important quality requirement, but a minimal one.

The place where field interviewers are most on their own and data quality is most vulnerable is in the actual administration of an interview. Although it is technically feasible to make audio recordings of interviews, respondents' perceptions of the protection of their privacy might be altered. Moreover, it would be infeasible in all but the smallest surveys to listen to more than a very small fraction of interviews. As discussed later in the third section of this paper, an approach developed for the SCF may be useful in controlling behavior at this stage. In addition, the last section of the paper discusses a technique being developed for the 2007 SCF that it is hoped will bring more immediate attention to questionable responses recorded during the interview.

II. The Survey of Consumer Finances

The SCF is a triennial household survey sponsored by the U.S. Federal Reserve Board in cooperation with the Statistics of Income Division of the U.S. Internal Revenue Service.¹ Data collection for the survey is carried out by NORC at the University of Chicago. The mission of the survey is to provide detailed data on family finances for policy, descriptive purposes and modeling. The data used in this paper derive from the 2004 survey.

¹See Kennickell [2000] for discussion of the survey methodology and references to supporting research. See Bucks, Kennickell and Moore [2006] for a summary of data from the latest wave of the survey.

To provide adequate representation of both highly concentrated assets and more broadly held ones, the survey employs a dual-frame design. An area-probability sample is intended to provide a robust base for estimating characteristics that are broadly distributed in the population. A list sample is selected in a way that oversamples wealthy families, who hold a very large fraction of many portfolio items and who receive a very disproportionate share of total income; this sample is drawn from a set of statistical records derived from tax returns using a “wealth index” computed from observed income flows and other data to stratify the observations.

The questionnaire was implemented in the field as a computer program—computer-assisted personal interviewing, or CAPI—run on laptop computers used by the interviewers. The interview content includes detailed questions about families’ portfolios and the institutional relationships that underlie their holdings. In addition, the survey asks for information on current and past employment, pension coverage, demographic characteristics, and other topics. The questions are often viewed by respondents as both difficult and sensitive.

In addition to the main questionnaire administered to respondents, the data collection for the survey employed several other electronic instruments. A “screener” instrument provided detail on the mechanism by which the respondent within a household was selected. Another module was used to collect observational data from interviewers on neighborhood characteristics and on characteristics of respondents, particularly the things said by the respondent as the interviewer attempted to negotiate cooperation with the interview. A set of “call records” cataloged every action taken on each case; although parts of this system were automated, it required substantial input from interviewers in describing the actions. Finally, for each complete questionnaire, interviewers were required to complete a “debriefing instrument,” which was a

place interviewers could elaborate on problems that occurred or to provide a brief overview of the case to support the quality of the information collected; this instrument had sections with specific questions as well as open-ended field where the interviewer could provide any information that might be useful in understanding the interview. All of these supplemental modules existed in electronic form, but the screener and observational data were initially collected on a paper form and subsequently entered into the computer using a specially tailored program.

The question text in the main instrument was often tailored to the situation of the respondent by incorporating other information reported. Although the computer program handled the progress through the instrument in response to the answers entered, interviewers still had need of specific instructions for use in clarifying questions to respondents or in interpreting their answers. Where relevant, such instructions were included directly on the computer screen with the associated question. Where unresolvable questions arose during the interview or where there were important inconsistencies, the interviewers were instructed to make a comment at that time using a facility available at every point during the interview, or if that was not convenient, to report such problems in the required debriefing interview. Such comments have traditionally played a key role in supporting the editing of the SCF data, which has contributed strongly to the quality of the final information over time.

Unit nonresponse rates in the survey are high, compared with most other U.S. government surveys. For the area-probability sample in 2004, 68.7 percent of the eligible respondents participated (table 1). The situation for the list sample is a little more complicated. As was the case for the area-probability sample, cases in the list sample were contacted by letter before being approached by an interviewer, but unlike the situation for the area-probability

Table 1: Final case outcomes, by sample type, percent, 2004 SCF.

	<i>AP</i>	<i>LS</i>
Out of scope	18.0	0.6
Complete	56.3	30.0
Active nonresponse	19.8	30.0
Stopped work	5.9	39.5
<i>Memo items:</i>		
Postcard refusal	NA	12.9
Response rate	68.7	30.2

cases, they were offered an opportunity to refuse the interview definitively by returning a postcard—12.9 percent of the list sample cases did so. Although overall only 30.2 percent of the eligible selected list sample cases participated, the participation rate varies strongly over the wealth-

index strata.² For example, the rate in the stratum likely to be least wealthy was about 35 percent, and that in the stratum likely to be most wealthy was about 10 percent. Research indicates that, among other things, nonresponse is correlated with wealth (see Kennickell [2005] for a summary of recent research).

An attempt was made to approach all cases at least initially in person. Of the 4,522 completed cases in the 2004 survey, 55.3 percent were coded as having been at least begun by telephone; the remaining cases were at least begun in person. Interviewers were encouraged to use the telephone to interview the respondents when this mode was acceptable to the respondent; informal evidence suggests that respondents often strongly prefer to be interviewed by telephone once their confidence in the survey is established.

²Ineligible list sample cases were ones where the respondent had died without leaving a surviving spouse or partner and those where the respondent was abroad for at least the entire field period.

For the area-probability sample, ineligible units are ones that are currently inhabited, where units with temporarily absent residents are included as inhabited. For the list sample, ineligible units are ones where the respondent was abroad for at least the entire field period or the respondent is deceased and not survived by a spouse or partner. In practice, a substantial amount of effort is devoted to the determination of and elimination of ineligible units. For the 2004 area-probability sample, 18 percent of the sample was determined to be ineligible; the fraction for the list sample was less than 1 percent.

Item nonresponse rates vary greatly.³ Most ownership questions have very low missing data rates and rates tend to be higher for monetary variables. Monetary questions have an extremely important place in the survey. In waves of the SCF before 1995, the first year that CAPI was used, great reliance had been placed on interviewers to probe respondents who were uncertain about such responses or who were resistant to providing an answer. One option that interviewers had in this probing was to use a card that contained a number of dollar-denominated ranges, each of which was identifiable by a letter of the alphabet; for certain critical questions about income, a formal decision tree was used to negotiate a range with the respondent. The design of the computerized version of the instrument attempted to replicate the ideal structure of probing while offering more sophisticated options for reporting ranges for all dollar-denominated responses. Moreover, because the computer program itself initiated the question sequences appropriate to the respondent's initial answer, a more uniform application of probing effort was enforced.⁴ In the SCF, all monetary variables have the option of being answered as a range, rather than as a single value. The range may be directly volunteered or may be the result of the probing generated by the interviewing software when such a question is answered initially with either a "don't know" or "refuse" response. All missing data are imputed using a multiple imputation technique (see Kennickell [1998] for an overview).

Interviewers are the key to all surveys other than self-administered ones. A skillful, knowledgeable and highly motivated interviewer will be more persuasive with respondents, and

³See Kennickell [1998] for a discussion of missing data and multiple imputation in the SCF.

⁴As shown in Kennickell [1997], the very positive outcome in terms of collecting partial (range) information and the apparent lack of effect on the frequency of complete responses suggests that previously interviewers overall were not sufficiently vigorous in following the protocol for probing.

will do a better job of navigating complex interviews. Interviewer recruiting for the SCF has always focused finding people who are not intimidated by having to ask respondents to participate in a financial survey. For the 2004 survey, emphasis was also placed on the ability to collect high quality data. For SCF-experienced interviewers, information on data quality in the 2001 survey (see Kennickell [2002] for a description of the evaluation of data quality) was used to identify interviewers who were most likely to conduct coherent interviews. For the inevitable new hires, the selection criteria included attributes that were believed to be related to the ability to conduct good interviews: active listening skills, reasonable computer facility, etc. Area field managers were involved in hiring to the extent possible to give them a vested interest in the set of people hired.

For the 2004 SCF, 186 interviewers were trained. Of these, 45 were designated as “travelers” who were intended to be used intensively in area outside their home areas. Over the course of the approximately six-month field period, the number of interviewers working declined. Some interviewers left the project because they had completed what was viewed as all workable cases in their area and they did not wish to travel. Some were terminated because they were unable to complete interviews at a sufficiently high rate or within a range of costs per case or, in very rare instances, because they violated a critical part of the survey protocol. Others left because they prefer to work only in the part of the field period when relatively easy uncompleted cases are still more prevalent. Others left for a variety of personal reasons.

Table 2: Cumulative number of completed interview and cumulative percent of all completed interviews, by ranking in terms of interviewer productivity, 2004 SCF.

Interviewer rank: top	<i>Cumulative:</i>	
	Completed cases	% of all completed cases
10	939	20.8
20	1579	34.9
30	2164	47.9
40	2593	57.3
50	2944	65.1
60	3215	71.1
70	3454	76.4
80	3669	81.1
90	3845	85.0
100	3989	88.2
186	4522	100.0

The productivity of the interviewers varied greatly (table 2). Interviews were completed by only 181 of the 186 interviewers, 16 interviewers completed only one case. At the other end of the spectrum, the most productive interviewer completed 116 interviews. The most productive 30 accounted for nearly half of all completed cases.

It should be noted that the assignment of interviewers to a set of cases was not purely random.

Initial assignments of cases were driven primarily by geographic considerations, but there is always some degree of “matching” of interviewers and cases. Some interviewers devoted relatively large efforts to telephone interviewing, often with respondents who had been separately persuaded to participate either by a member of the traveling team of interviewers or by field staff who specialize in securing the cooperation of respondents.

III. Interviewers and Interview Quality

The SCF interview is focused very largely on factual information, rather than opinions. Despite a long series of attempts to optimize the survey language in light of information on respondents’ misunderstandings and on changes in the marketplace, many of the questions remain necessarily technical in ways that some respondents might find confusing or unintuitive. For that reason, interviewers on the project have always been asked to go beyond simply reading questions and recording responses. Among other things, they are instructed to probe potentially

ambiguous responses and to help respondents, either when they explicitly ask for help or when the interviewer senses there is confusion or misunderstanding.⁵ Nonetheless, it has been difficult to get all interviewers to practice such behavior uniformly and to the desired degree.

In waves of the SCF before 2004, signs of deteriorating quality of interview data had been detected (see Kennickell [2002]). Moreover, the common belief that interviewers who are good at getting people to agree to participate in interviews also collect good data turned out to be not well founded. Indeed, according to some measures, data quality and performance in persuading respondents appeared to be negatively correlated. Obviously, respondents are also an important factor in data quality, but even controlling for key respondent characteristics did not alter the findings.

With the goal of countering the decline in data quality, the 2004 SCF introduced a new mechanism to monitor the quality of the information collected and to feed back comments to the interviewers.⁶ This system had two parts, one intended to be generated very quickly by computer and the other generated at a longer lag through detailed review of case data by subject matter specialists.

In the part of the system designed to provide a rapidly available evaluation, the data for every ostensibly completed case were screened by computer to determine the proportion of missing values for monetary variables and the number of keystrokes stored in a set of important interviewer comment fields in the main questionnaire and the in the open-ended response fields

⁵Based on experimental evidence, Conrad and Schober [2005] and references cited therein provide a data quality rationale for interviewers to take an active role in defining the meaning of questions for respondents.

⁶See Athey and Kennickell [2005] for a discussion of the monitoring system and some preliminary analysis of the resulting data and Wang and Pedlow (2005) for a discussion of a part of the monitoring system designed for rapid turnaround.

in the debriefing interview.⁷ These two factors, which were viewed as the most basic indicators of data quality in the SCF, were observable as soon as the completed cases were received by electronic transmission, usually within a day of the completion of the cases.

Although the part of the CAPI program for the main instrument that probes missing monetary amounts should, in principle, result in uniform application of effort, it is obvious that interviewers still may have a great influence on the outcome. For example, interviewers who can figure out how to make a respondents comfortable about the interview process and the protections in place for the confidentiality of their data will tend to have lower missing data rates, conditional on the distribution of the characteristics of respondents in interviewers' assignments.

As a part of their training, interviewers were told repeatedly that it was very important to document unusual or difficult situations, either by using the comment facility within the main interview or by using the structured debriefing required for each completed case, as discussed earlier in this paper. Even when a case was completely problem free, interviewers were instructed to make a note to that effect in the debriefing. In the past, situations in which comments were scanty were overwhelmingly ones in which comments would have been of great use in editing the data. Although the total number of key strokes in such fields is not a certain indicator of the thoughtfulness of the text, it is at least a rough indicator of how seriously the interviewer took the task.

Outcomes of these measures for individual interviewers were compared with patterns for other interviewers. Interviewers who had substantially different patterns from those for other

⁷In addition to direct tallying of "don't know" and "refuse" responses to monetary questions, the measure also includes such responses to a selection of higher-order questions that might lead to key monetary questions.

interviewers, or whose performance fell below a critical level, were examined by their supervisor during regularly scheduled calls during the field period. To facilitate oversight by managers, the information was digested into a simple form.

The second part of the interviewer monitoring system entailed a review of the interview data for each completed observation along with all of its accompanying data by a subject matter expert who provided a case-specific data quality score and a written review. The case review was based on the comments provided by the interviewer in the main instrument and the debriefing, possible anomalies identified by computer programs, and direct examination of the data from the main interview. To support this editing process, all interviewer commentary and the results of the computer searches for each case were formatted together on an “edit sheet” along with key descriptive facts about the household; the main case data were formatted with sufficient labels to be read as a pseudo-questionnaire. The resulting evaluation was transmitted to the survey field managers for use in weekly reviews of interviewers’ progress.

New cases were available to the editors weekly. But because the review process was time consuming and the number of editors was small, it was not possible to keep pace with the interviewers in the field, particularly early in the field period when the rate of case completion was still high. Nonetheless, it was possible to keep up with a selection of cases most likely to be problematic and to ensure that the work of all interviewers was regularly reviewed. Occasionally, the editing work skipped a week of data in order to be able to return comments on cases most likely to be still fresh in the minds of the interviewers. By the close of the field period, over 85 percent of the cases had been edited and comments and quality scores on those cases had been returned to the field.

The field managers were instructed to make a review of both the rapidly-available results and the results from the intensive reviews a part of their weekly review of the progress of each interviewer. When the results of the intensive review indicated serious problems with a case, the interviewer could be required to attempt to recontact the respondent. In very serious cases where the respondent could not be recontacted, the case might be dropped from the analysis data set; when a case was dropped, the interviewer lost the “credit” for the completed case. Because interviewers’ retention was conditional on meeting their production goals, the threat of losing such credit should have been an important motivating factor.

This system was expected to have three effects. First, the feedback should serve the function of providing continuing education to interviewers on how to administer a questionnaire successfully. In the past, some interviewers had complained that they were never given feedback on how well they were performing in terms of their data quality. Second, the feedback signaled to the interviewers that coherent information is important, that the project staff were watching for deviations, and that there could be serious implications. Third, the overall effect should be to increase data quality.

It was learned in the project debriefing that the managers of the interviewers differed in the stringency with which they reviewed the two types of feedback with their interviewers. Although these differences should have led to uneven shifts in quality across management groups, the expected overall effect is still positive.⁸ Because the number of productive interviewers was already fairly thin relative to the number of sample areas and because the number of cases was high relative to the number of interviewers in some areas (such as New

⁸It would be interesting, in principle, to control for manager effects. However, complicated reassignments of areas during the field period make it a practical impossibility to make a sufficiently sharp association of interviewers with managers to support such modeling.

Table 3: Definition of subject matter expert case-level data quality score.

1: High priority problem with case
2: Medium priority problem with case
3: Minor problem with case
4: No important problem with case

York City), no interviewer was terminated purely for data quality reasons. Nonetheless, it is clear that all interviewers were aware that intensive monitoring was taking place and

they could be called upon to explain their work to their manager.

The case-specific scores assigned by the subject matter experts indicated the seriousness with which the field manager should review each of an interviewer’s cases with the interviewer (table 3). In the most serious cases (score=1), the interviewer could be asked to recontact the respondent to obtain clarifying or missing information, including in some cases entirely repeating the interview with a different (correct) respondent. A score at the other end of the spectrum (score=4) indicates either that a case had at most minor problems or that it had problems that were not ones for which the reviewer thought the interviewer bore any meaningful responsibility.

Although the intention was that the scores be free of respondent-level effects, this is unlikely to be purely so. Some interviews were done with families with more complex circumstance than others, and may, thus, have had more chances to experience problems. One sign that reinforces this notion is the higher frequency of low (bad) quality scores for cases in the higher strata of the list sample (table 4).

Table 4: Priority score by sample stratum, 2004 SCF.

Score	<i>Stratum</i>								
	All	0	1	2	3	4	5	6	7
1	6.0	5.0	6.0	5.5	5.7	9.1	8.5	9.4	13.5
2	15.5	12.2	12.0	20.0	17.6	22.8	21.0	26.9	32.7
3	49.4	49.1	52.0	45.5	52.4	46.7	51.7	51.7	40.4
4	29.1	33.7	30.0	29.1	24.3	21.4	18.8	12.0	13.5

Table 5: Distribution across interviews of the percent of missing dollar values, by sample type, 2004 SCF.

Percentile	All cases	AP cases	LS cases
25	0	0	0
50	6	4	9
75	18	15	24
90	37	33	41

The percent of missing dollar values in the rapidly-available feedback also shows considerable variation across cases (table 5).⁹ List sample cases, on average, have more “opportunities” to have missing data because on average they tend to have more complex arrangements than area-probability cases. Thus, it is not surprising that the distribution for the area-probability sample lies below that for the list sample.

Table 6: Joint percent distribution of priority score and percent of missing dollar values, all cases and area-probability cases only, 2004 SCF.

% missing \$ amounts.	Priority score			
	1	2	3	4
	<i>All cases</i>			
<5%	2.5	5.5	23.3	16.0
5%–9.9%	0.5	1.8	6.8	3.7
10%–24.9%	1.4	3.5	10.3	6.2
≥25%	1.6	4.7	9.0	3.2
	<i>All area-probability cases</i>			
<5%	2.5	5.3	25.0	18.9
5%–9.9%	0.4	1.3	7.1	4.1
10%–24.9%	1.0	2.9	9.2	6.9
≥25%	1.1	2.8	7.8	3.8

⁹Instances where a respondent provided a range instead of a single amount is not treated as a missing value here. The basis for the percentage calculation is the number of variables to which it was known that a dollar response should have been made and the number to which it was not known because a response to a higher-order question was missing.

Comparison of the rate of missing value with the priority score suggests that the two have fairly different patterns (table 6). Although these measures of quality are negatively correlated, the connection is weak—only -0.18 for all cases and -0.12 for the area-probability cases alone.¹⁰ This finding indicates that using both for feedback may be productive.

All of the level statistics reported are contaminated to at least a degree by the fact that the underlying data were used in monitoring and directing the interviewers who subsequently provided additional interviews.¹¹ To get a sense of the effect of monitoring, two pieces of information would be useful: comparable data for earlier surveys and the time series of results across the field period for the 2004 survey. The former would give a sense of the overall level shifts induced; the latter would show the effects of learning (and to some degree, selection) over the field period. Comparable data on missing dollar variables and data on the extent of interviewers comments are available for earlier years, but the previous surveys did not use a quality scoring system like that used in the 2004 survey.

Table 7 provides a summary of the available data for 2001 and 2004 across the set of biweekly intervals of the field period. The frequency of missing dollar values in 2004 is below that in 2001, and this result holds at almost every biweekly period, both for the full set of participants and for the area-probability sample alone. The clearest success was in terms of the amount of interviewer comments provided in the debriefing interview. The mean level is considerable higher in every period in 2004 than in 2001. The fact that the amount of comments

¹⁰Recall that a lower fraction of missing data and a higher priority score indicate higher quality.

¹¹Although monitoring should have reduced the levels of the two outcome measures relative to the unmonitored state, it is questionable whether monitoring would have affected the correlation between the two.

Table 7: Percent of all dollar values missing, mean priority score, and mean length of interviewer debriefing comments (number of characters); by biweekly interval of the field period; 2001 and 2004 SCF.

Biweekly period	% missing \$ values		Mean priority score		Mean length debriefing comments	
	2001	2004	2001	2004	2001	2004
	<i>Full sample</i>					
1	15.6	13.3	NA	3.01	162	310
2	15.9	11.4	NA	2.88	163	296
3	12.4	13.1	NA	2.99	179	324
4	15.9	13.1	NA	3.04	191	385
5	16.1	12.9	NA	2.94	173	429
6	17.4	12.6	NA	3.04	191	416
7	14.9	13.4	NA	2.95	198	391
8	13.7	13.2	NA	3.07	201	391
9	13.8	13.3	NA	3.09	196	416
10	16.6	13.9	NA	3.09	213	405
11	13.5	11.0	NA	3.10	227	377
12	16.2	11.3	NA	3.17	193	392
13	21.0	9.8	NA	2.83	192	436
14	.	11.2	NA	3.14	.	396
15	.	11.3	NA	3.14	.	322
16	.	14.0	NA	3.15	.	441
All						
	<i>AP Sample</i>					
1	15.6	12.6	NA	3.01	162	292
2	15.9	10.3	NA	2.97	163	278
3	12.4	11.4	NA	3.05	179	300
4	15.4	11.9	NA	3.11	163	350
5	15.7	11.4	NA	3.03	165	368
6	14.2	10.2	NA	3.10	179	398
7	14.5	11.2	NA	3.05	162	336
8	10.2	10.1	NA	3.17	166	415
9	11.2	12.4	NA	3.24	154	385
10	14.2	12.2	NA	3.26	158	311
11	11.0	10.9	NA	3.28	142	321
12	15.1	10.1	NA	3.32	179	311
13	25.2	9.0	NA	2.99	156	302
14	.	8.3	NA	3.33	.	357
15	.	8.9	NA	3.33	.	262
16	.	11.3	NA	3.54	.	332
All						

Note: The actual reporting periods for 2001 and 2004 extended beyond the 13 periods shown for 2001 and the 16 periods shown for 2004, but too few cases were obtained in the final weeks of the field period to support meaningful separate analysis here.

also rose more sharply in the first intervals in 2004 suggests that monitoring and feedback had an effect.

The pattern of the quality score over the field period shows some signs of increasing slightly toward the end of data collection—that is, data quality increased. Underlying the slight rise in the average quality of interviews over the field period is a bit stronger decline in the proportion of cases with the most serious problems, both for the sample overall and for the area-probability sample alone (table 8).

One would expect that cases would tend to become more difficult over the field period, and thus increase the likelihood of lower scores. To adjust for some key difference in the difficulty of cases over the course of field period, the quality scores were filtered using a regression technique. The full set of scores was regressed on

Table 8: Percent of cases with a priority score of 1, by biweekly interval of the field period, full sample and area-probability sample only, 2004 SCF.

Biweekly period	Full sample	AP sample
1	5.9	5.4
2	10.6	9.2
3	6.6	6.5
4	4.0	2.7
5	6.2	4.6
6	5.9	5.7
7	6.9	5.0
8	7.2	6.5
9	5.3	4.4
10	3.1	0.0
11	4.6	2.4
12	3.3	1.2
13	6.5	8.3
14	4.8	2.5
15	3.0	0.0
16	3.6	1.5
All	6.0	5.0

Table 9: Unadjusted and adjusted quality scores as a percent of mean score of each type, by biweekly interval of the field period, 2004 SCF.

Biweekly period	Unadjusted	Adjusted
1	99.87	98.01
2	95.56	94.42
3	99.16	98.63
4	100.74	100.92
5	97.35	97.46
6	100.69	101.58
7	97.74	98.99
8	101.88	102.43
9	102.39	103.88
10	102.42	102.92
11	102.71	101.76
12	105.02	104.81
13	93.88	94.88
14	104.06	104.09
15	104.14	104.56
16	104.32	105.95

dummy variables for sample stratum, age of the household head, marital status, region, population size class of local area, ownership of financial assets, presence of any type of debt, and the interviewer's perception of the respondent's interest in the interview, ability to express answers, and level of suspicion both before and after the interview; logarithms of the maximum of 1 and the values of income, assets, financial assets, and net worth; the ratio of total debt payments to total income; and the interaction of the age and region dummies with the logarithm of the maximum of 1 and income. The standardized score is calculated as the model evaluated for a fixed set of characteristics plus the observation-specific residual from the model. Despite the inclusion of this range of controls, the regression explains only about six percent of the variation in the scores. What remains unexplained is surely a function of a mixture of characteristics of the respondents and characteristics of interviewers as of each interval of the field period as well as possible variations in the scoring techniques of the subject matter experts.

Very similarly to the case with the unadjusted scores, the pattern of means of the adjusted scores shows a mild rise in data quality over the field period (table 9). The fact that the adjustment makes so little difference might reflect

Table 10: Distribution of cases over rounded mean standardized quality score and case completion group, both computed over interviewers, percent, 2004 SCF.

Mean strdzed quality score	Number of completed cases				
	1-10	11-25	26-50	≥51	All
1	2.8	0.0	0.0	0.0	2.8
2	6.6	2.8	0.6	0.6	10.5
3	23.8	21.6	21.0	15.5	81.8
4	5.0	0.0	0.0	0.0	5.0
All	38.1	24.3	21.6	16.0	100.0
<i>Memo items:</i>					
Std. dev. of standardized quality score over interviewers in group					
	0.695	0.334	0.216	0.216	0.484
Mean of standardized quality score over interviewers in group					
	2.84	2.94	2.95	3.12	2.93
Number of interviewers completing at least 1 interview					
	69	44	39	29	181

omission of other important variables in the model, including effects of matching of respondents and interviewers; interviewer fatigue over the field period; low levels of rigor in reviewing data quality with interviewers; or selection effects generated as the pool of interviewers decreased in size over the field period.

The patterns of the data quality scores across interviewers were varied. As noted earlier, work reported in Kennickell [2002] suggested that case completion and data quality were at best only loosely correlated.

The quality measures used in that study were not identical to the quality score discussed in this paper, but at least one of the core measures should be related. In the case of the 2004 survey, across all completion groups there is a distinct concentration of the standardized quality score at level 3, a level reflecting small and non-critical concerns with the data (table 10). In contrast to the conclusion of the earlier study, variability around this point was least for the two groups with more than 26 completed interviews. This result raises the question of whether these highly productive interviewers were simply more responsive to the structure of incentives from the beginning of the 2004 survey or whether monitoring and feedback improved their performance over the field period.

Table 11: Distribution across interviewers of parameter on biweekly reporting period in a regression of standardized quality score on reporting period, for interviewers completing various numbers of cases.

%ile	<i>Number of completed cases</i>		
	26-50	>50	>25
99	0.267	0.066	0.267
95	0.115	0.054	0.099
90	0.099	0.053	0.087
75	0.077	0.035	0.046
50	0.035	0.007	0.016
25	-0.004	-0.003	-0.004
10	-0.022	-0.021	-0.021
5	-0.060	-0.027	-0.039
1	-0.093	-0.040	-0.093
Mean	0.035	0.012	0.026
Std. dev.	0.061	0.027	0.050

If the standardized quality score for each case is regressed by interviewer on the biweekly wave number, the result is an estimate of the trend level of quality change over the field period for each interviewer.

Table 11 provides the distribution of these estimates across interviewers who completed 26–50, >50 and >26 cases. Although the mean change for all groups is positive (that is, quality improved over the field period), the distribution shows a high level of relative variability. Moreover few of the estimates for the

underlying interviewer-specific models are statistically significant. Thus, the data suggest that these groups may have begun work at a higher standard of data quality than less productive interviewers and then did not much alter their behavior as a group. Monitoring and feedback may still have had the effect of maintaining the credibility of the data quality program laid out in the interviewer training sessions.

IV. Interviewers and the “Quality” of Nonresponse

As well as influencing the quality of the information collected in an interview, interviewers may also affect the set of people who become participants. Groves and Couper [1996] develop the argument that an interviewer’s ability to understand and respond to respondents’ comments and reservations during the negotiation of agreement to do an interview. But even beyond the point of negotiation, in field surveys interviewers have a power to alter the

set of cases that are ultimately completed by varying the intensity of their efforts in ways that are difficult or impossible to monitor directly.¹²

Sample selection is a complex task that seems to be misunderstood sometimes even by people who have worked closely with surveys for years. Field staff appear to be most highly motivated by the idea of completing interviews, regardless of how the set of interviews that remain incomplete might differ in any systematic way from those that are completed. It appears to be a common belief that statisticians will find a way to make any set of completed interviews represent the desired population, no matter what pattern of selection takes place in the field, either as a results of interviewers' actions or respondents' actions.

If all cases have an unambiguous threshold of decision about survey participation, and effort is expended on every case until that point is reached (with a definitive positive or negative outcome), then at least any bias in the set of participants directly reflects only the attitudes and motivations of those participants.¹³ But if respondents have some degree of persuadability, interviewers have expectations about whether cases are more or less likely to be completed, those expectations are rational, and interviewers act on those expectation, then interviewers will tend to amplify patterns of nonresponse that would tend to arise from a neutral application of effort. As suggested by the simple behavioral model presented earlier in this paper, an incentive structure based only on completed interviews will tend to drive interviewers in this direction. Indeed, evidence from earlier waves of the SCF (see Kennickell [2004]) suggests that there were systematic variations in effort that are correlated with observable respondent characteristics.

¹²See Kennickell [2005] for a specific model of such behavior.

¹³There may also be effects generated by respondents' reactions to unchangeable characteristics of individual interviewers, but such effects are ignored here.

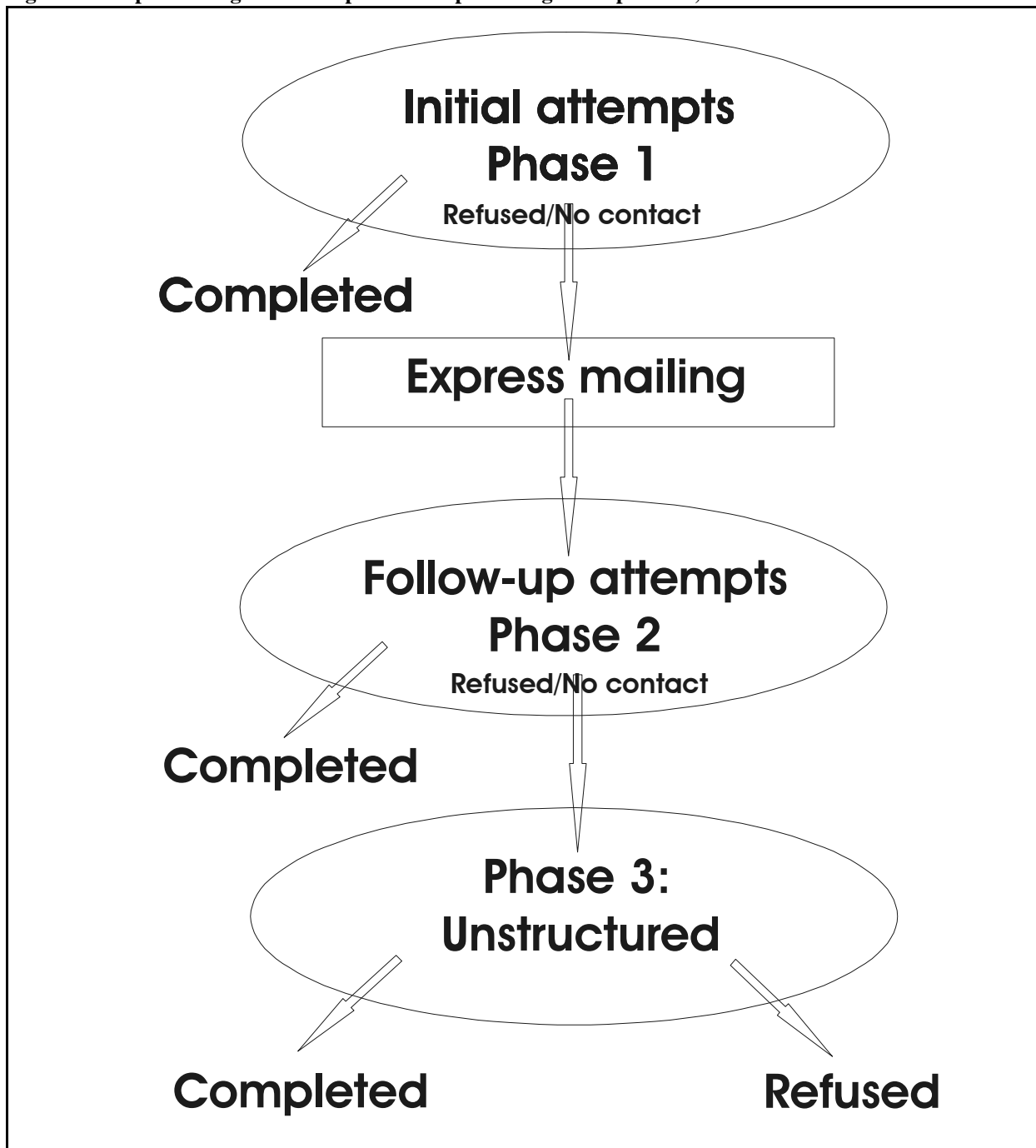
In an attempt to reduce the behavioral component in nonresponse that does not originate purely with the behavior of the respondents, the 2004 SCF introduced a method of case management specifically designed to provide a more measured “dose” of effort to respondents. To be useful, a protocol for respondent contacts must have enough flexibility to be workable in the complex set of situations interviewers face, it must not distort interviewers’ incentives in a way that might lead to deterioration of data quality in another dimension, and it must provide a measurable framework for characterizing the amount of effort applied to each sample case. Figure 1 provides a schematic overview of the protocol employed.

In the first phase of the approach, interviewers were instructed to attempt to make contact with each respondent to explain the survey and solicit cooperation to do the interview.¹⁴ Generally, up to four attempts are allowed in this phase, but interviewers are required to justify their attempts as sufficiently meaningful attempts in the eyes of their managers. For example, passing by a house twice to see if a car is in the driveway over the course of a day while working in a neighborhood is less compelling as a measure of two separate attempts than two visits on the same day where first the interviewer contacts an informant provides information on when the respondent will be home and then contacts the respondent later in the day; the computerized case management system available could not readily distinguish between these two. Human evaluation was seen as necessary to ensure regularity.

In the 2004 survey, 42.0 percent of eligible respondents in the area-probability sample and 15.2 percent in the list sample cases that had not returned the refusal postcard agreed to participate within the first phase. However, sometimes it is not possible for an interviewer to reach the respondent—because there is a physical obstacle, such as a gate, because there is a

¹⁴In this discussion, the problem of ineligible sample units is ignored.

Figure 1: Simplified diagram of the phased sample management protocol, 2004 SCF.



person who serves as a controller of access to the respondent and who refuses to allow access, because no one is ever present at the sample address when the interviewer visits (or when there

is a telephone number available, no one ever answers) and there is no information from neighbors or others to suggest that the respondent is temporarily away, and because of a number of other more minor reasons; 51.8 percent of the eligible area-probability sample that did not reach a final resolution in the first phase without a respondent contact and 61.4 percent of the list sample met this condition (56.7 percent of both samples).¹⁵ In the remaining instances in the first phase, the respondent is contacted and declines to participate, either by actively refusing or by delaying a decision in some way; a refusal automatically terminates the first phase, regardless of the number of attempts. If an interview has not been conducted or scheduled to be conducted at the completion of the first phase, as agreed with the interviewer's manager, then the respondent is sent a specially designed brochure on the SCF by express mail.

The brochure for 2004 contained information about the survey, the Federal Reserve Board, NORC, and uses of the data, as well as a letter from the Chairman of the Federal Reserve Board, a discussion of confidentiality, and a set of contact options for respondents who might want additional information. It was designed to present all of this information in an easy-to-read, uncreenergetic in its approach than any of the routine material available for interviewers to use while negotiating cooperation with respondents. The use of express mail was intended to make the brochure clearly visible to respondents as something out of the ordinary. Interviewers, who were aware of this brochure from their training, had a strong desire to use this tool. Thus, to ensure that attempts made in the first phase were sufficiently credible, approval of an interviewer's manager was required before the marker for the first phase was set and the brochure could be sent to a respondent.

¹⁵In 2.3 percent of area-probability cases and 4.7 percent of list sample cases, the respondent refused participation so vehemently in the first phase that there was no follow-up in the second phase.

In 2004, the first phase of the protocol appears to have worked much as it was intended. After the brochure is transmitted, the protocol specifies that interviewers should apply at most another four attempts (again, with review of quality of the attempts by the interviewers' managers). During this phase, interviewers are expected to exploit the nearly certain knowledge that the respondent has received a substantial dose of information about the project.¹⁶ If the respondent cannot be persuaded to participate within the specified number of attempts or if the respondent refuses to participate before that number of attempts, the case is to be designated as having reached the end of the second phase. At this point, the remaining cases are to be evaluated on an individual basis as to whether they are likely to be amenable to additional persuasion.

Unfortunately, in the 2004 survey many interviewers and even some managers misunderstood the implications of marking the end of the second phase. It seems that there was a belief that somehow their ability to exert additional effort on cases after this point would be restricted in a meaningful way and that they might lose the "investments" already made in trying to interview cases in their assignments. Consequently, there were notable irregularities in the marking of this phase among the 6,383 cases active in that phase. Of the area-probability cases that remained in play from the first phase, 35.4 percent failed to be marked as having completed the second phase after 10 attempts beyond the express mailing; for the comparable list sample cases, the fraction was 39.8 percent (38.0 percent of both samples).

For contemporaneous monitoring or even descriptive purposes *ex post*, it is relatively more difficult to produce a quality score for compliance with the sample-management protocol

¹⁶In a small number of cases, the express mail package was refused. Interviewers could see in their electronic case management system when the package was sent by the Central Office and whether it was returned.

than was the case with data quality. In this discussion, the frequency of noncompliance with the second phase protocol defined as more than 10 attempts without marking the phase termination is one such indicator of violation of the protocol.

The data used to make this determination come from a system used to manage interviewers' work with their case assignments. Like most administrative systems, this one is more attuned to the immediate needs of the users of the system—interviewers, their managers, and certain data processors—than to subsequent analytical needs. For completed cases, there is an unambiguous “interviewer of record,” but other cases are often worked by a variety of interviewers over the course of the field period, and in some cases, the final entries are made by managers, rather than interviewers. By scanning the history of each case to determine the latest interviewer identification number associated with the case; all except 65 cases could be assigned in this way to an interviewer who completed at least one case.¹⁷ The contact information was then linked to the indicators of data quality discussed earlier in this paper.

For various groups, table 12 shows the distribution of the interviewer-specific percentages of cases that showed more than 10 attempts in Phases 2. Of the 106 interviewers who had any cases active after Phase 1, all except 6 of them had a rounded mean standardized data quality score over all their cases of 3; but within this group, there was considerable variability in the frequency of seeming protocol violations. The median interviewer in this data quality group exceeded 10 attempts 36.6 percent of the time, while at the 90th percentile the figure was 57.6 percent and at the 5th percentile only 7.5 percent.

¹⁷A review of a sample of the unassigned suggests that these are ones that were part of the assignments of interviewers who never completed a case.

Table 12: Distribution across interviewers of percent of their cases remaining after Phase 1 that were still classified as being in Phase 2 after more than 10 attempts; by rounded mean standardized data quality score, percent missing dollar values, and number of completed cases; interviewers who completed more than 10 cases; 2004 SCF.

%ile	Rounded mean standardized data quality score*		Percent missing dollar values				Number of completed cases		All cases	
	2	3	<5%	5-9.9%	10-24.9%	≥25%	11–25	26–50	≥51	
90	66.7	57.6	56.0	100.0	59.4	67.2	66.7	57.8	57.1	57.8
75	42.7	49.7	43.2	43.2	49.7	50.0	50.0	44.6	46.0	49.4
50	23.6	36.6	34.7	42.0	39.1	33.3	25.0	36.2	41.2	36.1
25	15.3	20.8	11.8	37.3	25.9	22.2	9.2	20.0	31.3	20.0
10	0.0	7.5	6.7	10.9	15.4	0.0	0.0	9.5	25.0	6.7
Mean	28.6	35.0	31.2	46.7	40.4	33.9	30.7	34.9	39.8	34.6
Std. dev.	23.2	21.4	18.7	32.6	23.0	22.3	27.0	20.5	11.2	21.5
N	6	100	52	5	24	25	40	37	29	106

* With the restriction on the number of completed cases, there are no interviewers with an mean standardized data quality score of either 1 or 4.

Among the set of interviewers with active cases in the second phase, only 5 had an average percent of missing dollar values over their completed interviews in the range from 5 to 9.9 percent; about half had a figure of less than 5 percent and the remainder had a rate of 10 percent or more. The distribution of the percent of times an interviewer exceeded 10 attempts in the second phase shows little consistent relationship across the distribution with missing data rates, but at least the mean percent with more than 10 attempts is lowest for the lowest missing data rate group. In contrast, for interviewers with higher numbers of completed cases, there is a more notable upward shift in the proportion of such protocol violations in the lower half of the distribution as well as in the mean. Overall, the errors in marking the end of the phase appear only loosely correlated with the other quality indicators, as might be expected if there were no monitoring or if incentives were otherwise blurred.

The key difference in the decision to mark the end of the first phase and the second phase is that there was a reward to interviewers for marking the end of the first phase—in the form of

the attractive brochure sent to the respondent—but there was no reward or recognition for marking the end of the second phase. As noted earlier, the misperceptions of the function of the phase 2 completion marker gave interviewers a positive disincentive to set this marker.

Interviewers' managers were sent reports listing cases that had accumulated a substantial number of second phase attempts, but there was little incentive for them to examine these voluminous lists in detail, particularly during the second phase of field work when managers and interviewers are already busy developing strategies for locating or convincing difficult respondents.

Future work using this protocol design must focus on how to align the incentives of interviewers and managers more closely with the analytical interests of the survey as expressed in the protocol structure. One possibility is to attach the setting of the second phase marker to giving access to part of the more intensive assistance offered to interviewers for dealing with relatively difficult cases late in the field period. Monitoring of cases with unusually large numbers of attempts after the express mailing would also have an important role in maintaining the credibility of the approach.

Despite the shortcomings of the compliance of field staff with the case management protocol, it did bring more uniform attention to the sample cases and the information from the first phase proved useful in understanding patterns of respondents' reluctance to participate in the survey (see Kennickell [2005]). Given that one accepts the threshold model of response, a more successful execution of the second phase protocol would provide a marker that could be used to understand the *dynamics* of nonresponse within the field period, uncontaminated by variations in the application of effort, and point toward possibly previously unrecognized sources of bias in the set of final respondents.

IV. Conclusion

Data collection systems have many points of potential non-neutrality with respect to measurement. In this respect, the role of field interviewers is a critical one. Such interviewers are generally the only project staff who interact routinely with respondents; the interview process depends on their motivation of the respondent to participate fully; and the data collected are mediated through the performance of the interviewers in reading and explaining survey questions and in recording respondents' answers.

Two factors make it very difficult to evaluate the quality of the work of field interviewers. First, it is very difficult to observe what they do; normally all that is available is the traces of effort that appear either in questionnaire data or in supporting administrative systems, both of which are substantially shaped by interviewers' actions. Thus, any system established to monitor quality must be, at least in part, compatible with the incentives interviewers face in doing their jobs, whether the system be so directly or through reshaping the structure of interviewers' incentives, such as by expanding the range of secondary items that can be monitored. Second, it is difficult, if not impossible, to define an objective set of evaluation criteria that would be universally recognized as relevant. This paper has examined two aspects of data quality and the interventions constructed for the 2004 SCF. For this survey, these quality aspects are ones that bear on the fitness of the data for the analytical uses for which the survey is designed.

The 2004 survey employed two types of feedback to interviewers on their interviewing performance in an attempt to raise data quality. A set of calculations of rates of missing dollar amounts and the length of some key verbatim fields were used to provide quick feedback to interviewers. With a longer lag, material from a more intensive review of each case by subject

matter expert was used for more detailed discussions of data quality with the interviewers.

Together these efforts were seen as a successful first attempt at creating meaningful incentives to produce high-quality data.

To signal more clearly both the interviewers and respondents the importance of accurate information, the 2007 SCF is planning to employ a flexible framework for error detection and correction. During the interview, the computer program used for interviewing will detect a set of critical inconsistencies as well as a selection of other responses that appear at least superficially likely to be incorrect. When such problems are encountered, the program would produce a pop-up screen for the interviewer to review. The interviewer would have the option of working out the response during the interview where appropriate, with the help of the respondent where necessary. However, if the respondent is uncooperative or the time pressure is too great, the interviewer would have the option of deferring the screen until the mandatory debriefing interview that interviewers are instructed to complete as soon as possible after leaving the respondent. In any case, all such screens would be offered for review and further comment in the debriefing. By signaling the importance of data quality so directly to interviewers as well as continuing the intensive review of individual cases, it is hoped that accidents during the interview will be corrected, interviewers will bring greater attention to the collection of data, and the quality of information collected will improve. Additional training will be given to field managers to ensure that enforcement of quality standards is more nearly uniform across all groups of interviewers.

The 2004 SCF also attempted to employ a formal case management protocol to ensure that all sample cases were worked to a measurable degree and without bias induced by interviewers steering effort toward cases that they perceived to be more likely to resolve as

completed interviews. This system was intended to operate in three phases, the first of which ended either with a completed interview or the sending of a strongly designed informational brochure to respondents. This phase was largely successful, probably because interviewers had an incentive to ask for the material and because the point in the work was highly salient. In contrast, the distinction between the remaining phases turned out to be unclear to the field staff and there was not a sufficiently strong incentive to follow the protocol; moreover, many of the field staff misunderstood the implications of allowing acknowledging transitions between the two remaining phases, thinking incorrectly that such acknowledgment would inhibit further work on any cases marked as having made the transition. As a consequence, an important part of the analytical benefit of the phased approach was lost. Nonetheless, it did serve to organize a significant fraction of the effort and made it possible to examine participation rates by groups uncontaminated by variations in the application of effort in the first phase.

Although it will be relatively easy to correct the misperceptions of interviewers and their managers about the consequences of setting the marker for the end of the second phase, further reform of the second phase of the field management protocol appears somewhat more difficult because of the need to use human evaluation of the weight of work done in the phase. set after four “meaningful” attempts. Because the very particular information necessary to determine whether the bar has been reached rests most clearly with the interviewers, it is desirable that the interviewers should be in the position of pushing their managers to agree to setting the marker. Perhaps attaching a “reward” such as additional assistance (letters or respondent payments) to the setting of the marker would have this effect. Of course, monitoring of cases with unusually numbers of attempts in the second phase would remain important.

The point of this paper is that behavioral incentives for field staff are important in data collection. Like anyone else, interviewers will tend to act in their own interest as they see it. Although most interviewers attempt to do what they think is a good job, in the absence of feedback they will tend to develop an idiosyncratic definition of what that means. Data collection efforts need to recognize the broad incentives interviewers face and structure their jobs to support the outcome desired for the data in a cost-efficient manner. One such important area not discussed in this paper is the structuring of remuneration for interviewers' work that would both reinforce efficient work practices and increase the attractiveness of the work to the highest quality interviewers and encourage their retention. Traditional field management that focuses on rewarding interviewers who complete numerous cases, but this may well be a weaker instrument for supporting data quality goals. Further research in this area, perhaps applying results of work in labor economics, contract theory and agent-principle theory, is needed to design more nearly optimal compensation and other incentive schemes for interviewers.

Bibliography

- Conrad, Frederick G and Michael F. Schober [2005] “Promoting Uniform Question Understanding in Today’s and Tomorrow’s Surveys,” *Journal of Official Statistics*, vol. 21, no. 2, pp. 215–231.
- Athey, Leslie and Arthur B. Kennickell [2005] “Managing Data Quality on the 2004 Survey of Consumer Finances,” paper presented at the Annual Meetings of the American Association for Public Opinion Research, Miami Beach, Florida, May 12–15.
- Bucks, Brian K., Arthur B. Kennickell and Kevin B. Moore [2006] “Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances,” *Federal Reserve Bulletin*, pp. A1–A38.
- Groves, Robert M. (1989): *Survey Errors and Survey Costs*. Wiley, New York
- _____ and Mick P. Couper [1996] “Contact-Level Influences in Face-to-Face Surveys,” *Journal of Official Statistics*, Vol. 12, No. 1, pp. 63–83.
- Kennickell, Arthur B. [1997] “Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances,” working paper, <http://www.federalreserve.gov/pubs/oss/oss2/method.html>.
- _____ [1998] “Multiple Imputation in the Survey of Consumer Finances,” working paper, <http://www.federalreserve.gov/pubs/oss/oss2/method.html>.
- _____ [2000] “Wealth Measurement in the Survey of Consumer Finances: Methodology and Directions for Future Research,” working paper, <http://www.federalreserve.gov/pubs/oss/oss2/method.html>.
- _____ [2002] “Interviewers and Data Quality: Evidence from the 2001 Survey of Consumer Finances,” working paper, <http://www.federalreserve.gov/pubs/oss/oss2/method.html>.
- _____ [2004] “Action at a Distance: Interviewer Effort and Nonresponse in the SCF,” working paper, <http://www.federalreserve.gov/pubs/oss/oss2/method.html>.
- _____ [2005] “Darkness Made Visible: Field Management and Nonresponse in the 2004 SCF,” working paper, <http://www.federalreserve.gov/pubs/oss/oss2/method.html>.
- Yongyi Wang and Steven Pedlow [2005] “Interviewer Intervention for Data Quality in the 2004 Survey of Consumer Finances,” presented at the 2005 Annual Meeting of the American Association for Public Opinion Research, Miami Beach, FL, May 12-15.