Look and Listen, But Don't Stop: Interviewers and Data Quality in the 2007 SCF

Arthur B. Kennickell
Federal Reserve Board

Paper prepared for the 2007 Joint Statistical Meetings
Salt Lake City, Utah

May 11, 2007

Abstract

In most field surveys, the data collection process is observed only by the respondents and the interviewers. Others can observe only the traces reflected in the data and paradata. Careful selection and training of interviewers and thoughtful construction of the survey instrument are, of course, very important in maintaining data quality. But creating a continuing mechanism for clarifying and reinforcing the survey protocols is also important. The 2004 Survey of Consumer Finances (SCF) introduced a new two-part system of data review coupled with regular feedback to the interviewers throughout the field period. Based on the experience with that system, a more refined version was developed for the 2007 SCF. This paper presents a discussion of the process as seen from the perspective of the on-going survey.

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, formerly of NORC, Suzanne Bard, Kyle Fennell, Eric Jodts, Catherine Haggerty, Julia Lane, Steven Pedow, Micah Sjoblom, John Thompson and other Central Office staff at NORC and the field managers, interviewers and respondents for the 2007 Survey of Consumer Finances. The author is also grateful to his SCF colleagues at the Federal Reserve Board, particularly Brian Bucks, Gerhard Fries, Daniel Grodziki, Traci Mach, and Kevin Moore, and colleagues at the Statistics of Income Division of the IRS, particularly Barry Johnson, Michael Parisi and Tom Petska. Thanks to Nancy Gordon for organizing the session in which this paper was presented and to Cheryl Landman for thoughtful comments.

This paper focuses on the control of measurement errors that may arise as the result of interviewer behavior during the administration of an interview for the Survey of Consumer Finances (SCF). Such behavior may be either active—failure to follow an immediate instruction—or passive—insufficient probing of unresponsive or inconsistent reports from survey respondents.[1] Careful selection and training of interviewers are, of course, very important in maintaining data quality. Hiring decisions determine the distribution of skills among interviewers and training provides information to interviewers on the general and survey-specific protocols to follow in administering an interview. However, what is more important is how skills and that information are brought to bear during real interviews. In most field surveys, the final data collection process is observed only by the respondents and the interviewers.[2] Others can observe only the traces reflected in the data, including the main interview data and associated para-data, and most often such data are difficult penetrate in a way that reveals timely information about behavioral patterns among interviewers during data collection..

A continuing and credible mechanism for clarifying and reinforcing the survey protocols during the period of data collection would offer important benefits. Such a mechanism would make the information held by interviewers and survey managers less asymmetric, and thereby highlight problems in individual cases and structural problems in the larger survey process. In addition, if interview data quality is recognized as an important dimension of interviewers' work, this change in the structure of information should alter the incentives for interviewers to be concerned about data quality.

Earlier work summarized in Kennickell (2002) focused on the 2001 SCF indicated a pattern of declining data quality in some key dimensions over waves of the survey. If that trend had been continued, ultimately it would have been pointless to continue the survey. In an attempt

---

[1]Obviously, the two are related if the survey protocol calls for probing of unresponsive, inconsistent or otherwise unclear responses, as is the case in the SCF.

[2]One might record interviews, but generally such actions can only take place with the permission of respondents. Such choice-based sampling could provide useful information, but used alone it risks providing a biased view of behavior if the act of choice alters the behavior of respondents of interviewers, even if the actual use of this technique were randomized in a way unobservable to either the respondent or the interviewer.

to forestall this outcome, the survey instituted a set of new procedures for the 2004 wave of this triennial survey. Athey and Kennickell (2005) and Kennickell (2006) summarize the effects of these protocols on that wave. That analysis of the 2004 process motivated the revisions for the 2007 survey discussed here.

In brief, the approach taken in the 2007 SCF is as follows. Recruiting was guided in part by evaluations of interviewers who had participated in the 2004 survey; those receiving relatively low scores for data quality were not allowed to work on the 2007 survey. Data quality, which has always been a critical message in SCF training sessions, became an organizing principle for nearly all of the instructional material. A complete reprogramming of the CAPI instrument for the survey made it possible to incorporate sophisticated edit routines that require specific interviewer reactions and comments; these checks focused on areas where the most troublesome problems have been found in the past. Finally, a two-pronged approach was used to send case- and interviewer-specific data quality evaluations to the field. One part of this feedback was based on an automated tally and analysis of key interview statistics in the central office of NORC, the contractor for data collection on the survey. The second part was based on more time-consuming in-depth examination of the data in each case by subject-matter experts at the Federal Reserve Board, resulting in a score for the performance of the interviewer and a set of detailed comments on questionnaire administration. Both types of feedback were transmitted to interviewers weekly, after an initial delay which led to an "accidental experiment" described later in this paper. The key benefits of the evaluative steps are that the process took place while the survey was in progress, making it possible provide continuing education and monitoring, an altering, where necessary, the behavior of interviewers in subsequent interviews.

The first section of this paper provides background on the SCF needed for understanding the issues of data collection. The next section reviews the quality control procedures instituted for the 2007 survey. The third section provides an empirical evaluation of the effects of the new procedures, using the data available from the survey, which was on-going at the time this paper was written. A final section offers conclusions and suggests further research and implementation of procedures to increase data quality..

I. The Survey of Consumer Finances

The primary purpose of the SCF is to provide data to support the analysis of the financial behavior of U.S. households and their use of financial services. Since 1983, the SCF has been conducted every three years by the Federal Reserve Board (FRB), in cooperation with the Statistics of Income Division of the Internal Revenue Service.[3] Data for the survey have been collected by NORC at the University of Chicago beginning with the 1992 survey. This paper draws data primarily from the 2004 survey and the on-going 2007 survey. The content and design of the two surveys differ in only minor ways.

The SCF questionnaire collects detailed information on a wide variety of assets and liabilities as well as data on current and past employment, pensions, income, demographic characteristics and attitudes. Although there is an attempt in the questionnaire to modulate the difficulty and sensitivity of the questions asked, the core questions are factual. Many questions require serious thought or access to records. When pruned of purely administrative variables, the final version of the raw data contains over 12,000 variables, though only about 3,500 of these are primary variables potentially answered directly by respondents, and some of these variables correspond to questions in parallel sequences of which only one can be answered. Most sections of the questionnaire are asked only if the respondent answered a question indicating that a more detailed line of enquiry was appropriate. For this reason and because some respondents require more probing and information from the interviewer, the length of time required for an interview varies considerably (table 1). Some of the longer interviews are completed over multiple sessions.

Table 1: Distribution of interview length in minutes, 2004 SCF.

Mean	91
5th percentile	41
10th percentile	48
25th percentile	62
Median	83
75th percentile	111
90th percentile	139
95th percentile	165

At least an initial attempt is made to contact every sample member in person to request participation in the survey; overall, this approach is believed to be important for establishing the credibility of the study with respondents. Where a phone number can be obtained at that point or by other

[3]See Kennickell [2000] for discussion of the survey methodology and references to supporting research. See Bucks, Kennickell and Moore [2006] for a summary of key results from the 2004 survey.

means, the interviewers are instructed to maximize their use of the telephone in subsequent follow-up, in order to control costs. Informal feedback from interviewers suggests that many respondents prefer to be interviewed by telephone because they do not want to let the interviewer into their home or office. Over 55 percent of all completed interviews in the 2004 survey were at least begun using the telephone.[4] Analysis of earlier SCF data reported in Kennickell (2002) could not find significant differences in the quality of data collected in person and that collected by telephone.

Item nonresponse in the interview varies a good deal across variables. Typically, variables that ask about ownership have close to 100 percent response, and variables that request dollar amounts have lower rates of response. As discussed in more detail later in this paper, the SCF CAPI program has the ability to accept range responses for dollar questions; if range responses are not included as missing data, the rate of item nonresponse is usually well below 10 percent.[5] For example, the nonresponse rate by this definition is 1.2 percent for home value, 3.3 percent for wage income, 5.3 percent for the amount in the main checking account, and 12.9 percent for the value of the largest business actively managed by the survey family. All missing data are multiply imputed.

The survey sample is based on a dual-frame design, including both an area-probability sample and a list sample. The area-probability (AP) sample is selected from a geographically based national frame developed by NORC at the University of Chicago (O'Muircheartaugh *et al.* (2002)). The list sample is designed to provides an over-sample of families likely to be relatively wealth. This sample is selected from a set of statistical records derived from individual income

---

[4]Based on the mode of administration identified in the final case status code for each case, it appears that only about 47 percent of cases was *completed* by telephone. There is the potential for a change of mode in a later session for a case. However, a more likely explanation of the difference is the fact that discrimination among final status codes representing case completion is not rigorously monitored and error has minimal immediate consequences, whereas the choice of mode within the instrument must be confirmed and the choice is known to drive the display of information to the interviewer.

[5]As indicated by Kennickell [1997], the very positive outcome in terms of collecting partial (range) information and the absence of an offsetting decline in the frequency of complete responses suggests that previously interviewers, overall, were not sufficiently vigorous in following the protocol for probing.

tax returns by the Statistics of Income Division of the IRS. This set of records is stratified using a model to predict a measure of wealth and records are sampled at progressively higher rates in wealthier strata (Kennickell (2001)). The overall initial sample of approximately 10,000 cases is about evenly divided between the two sub-samples. The SCF employs a number of tactics to engage the positive interest of the sample members in participating in the survey. Although the survey routinely offers most respondents $20 as thanks for participating in the survey, this approach is only a part of a larger focused effort that involves careful management of the level of effort devoted to each sample member.[6] About two-thirds of the 4,522 cases completed in the 2004 SCF derived from the area-probability sample; this represents a response rate of 69 percent for the area-probability sample and 30 percent for the list sample, with substantial variation in rates across the list sample strata. Research indicates that nonresponse in the survey is positively correlated with wealth.[7]

The final internal version of the survey comprises a variety of data and para-data. While collecting the main interview data, interviewers are expected to record comments detailing any issues that arise during the interview about any of the answers; soon after leaving the respondent at the end of the interview, they are required to complete a debriefing interview about the main interview. As discussed in more detail in the next section of this paper, these means of providing such information expanded notably in the 2007 SCF. Further information is available for each case from the detailed call records maintained for each attempted contact or action and from the original sample frame data.

---

[6]All members of the area-probability sample and members of the list sample from the two least wealthy strata of the list sample are initially offered $20. The wealthier members of the list sample are not initially offered anything, but if they ask they are also eligible to receive this amount; the motivation for this approach was that the amount might seem so small as to trivialize the study in their eyes and it might raise suspicions. Respondents have the option of receiving the money themselves or donating it to a charity. Later in the field period, respondents might be offered a larger sum.

[7]See Kennickell (2005) for a description and analysis of the sample contacting strategy. List sample respondents have one more opportunity to decline participation than the area-probability sample cases. The list cases are sent an initial letter along with a postcard to be returned if they do not wish to participate. If the postcard is returned, no additional effort is made to change the respondent's mind.

II. Interview Data Quality Control Cycle in the SCF

For the SCF, control of interview data quality spans recruiting and training of interviewers, questionnaire design and implementation as CAPI, active and guided interventions by interviewers during an interview, post-interview commentary by interviewers, automated and intensive reviews of the data, feedback to the field during the data collection period, systematic evaluation of these processes, and redesign of the next survey. Although each of these aspects of quality control is discussed here, attention focuses most on the questionnaire design and implementation and the data review and feedback.

All field surveys depend critically on interviewers, and recruiting decisions determine the basic pool of talent available. About 20 percent of the interviewers for the 2007 SCF had experience on an earlier round of the survey and about two-thirds of the interviewers had other experience on another NORC survey. The SCF-experienced group was a very select group. Interviewer-specific average data quality scores, adjusted for observed respondent characteristics, were computed, and only those interviewers who had been able to manage a high completion rate and who had a sufficiently high data quality score were eligible to work again on the 2007 survey. Exceptions were made in only a few cases when field managers made special arguments in favor of particular interviewers and the managers pledged to perform a additional monitoring to ensure high data quality. In addition to fitting the usual profile of a successful interviewer, new interviewers for the SCF were required to pass a test showing aptitude in several areas, including the ability to write numbers, the ability to recognize and probe answers that were designed to be nonresponsive in terms of the substance to a question, and the ability to follow instructions.

Interviewer training provides the basic information that interviewers need in order to follow survey protocols and to respond to questions from respondents. Training for the 2007 SCF was organized around the idea of data quality. Particular attention was given to explaining what interview data quality means in an operational sense. Each interviewer was required to study material sent in advance of an in-person training and to complete a quiz to be submitted upon arrival at training. Similarly, interviewers were required to take a “final exam” at the end of training to demonstrate mastery of key technical concepts and to complete a scripted interview to the satisfaction of an observer.

The main questionnaire, implemented as CAPI, is the most important tool available to interviewers. It expresses the desired conceptual framework of the survey as a series of questions which have been tuned by testing and experience to maximize clarity and minimize error. CAPI enforces the logical structure of the questions, given the data typed into the computer by the interviewer; of course, interviewers may enter incorrect information for a variety of reasons.

Administration of the instrument requires the interviewer to interact with both the respondent and the computer—maintaining the respondent's motivation, reading each question, possibly providing information to the respondent about the content of the question, sometimes probing for an answer, listening to the answer to understand it and be certain it is responsive to the question, explaining or probing where necessary or as requested by the respondent, determining how to express the answer in the terms available on the computer, and navigating the computer to encode the response. Clearly, this is a demanding set of tasks.

The questionnaire for the 2007 survey was completely reprogrammed for CAPI using MRinterview, a sophisticated language that employs a flexible browser interface for question display and data entry. This effort allowed refinement and extension of the computer-based tools and real-time editing systems. To aid the interviewer in the interaction with the respondent, the instrument incorporates explanatory material and alternative texts to be read. Instructions and any particularly important definitions are provided directly on the computer screen, as appropriate for each question, sometimes conditioning on information reported earlier in the interview. More extended definitions are available through an on-line glossary that can be accessed from any point in the interview. Pre-scripted probes are given for common problem situations.[8]

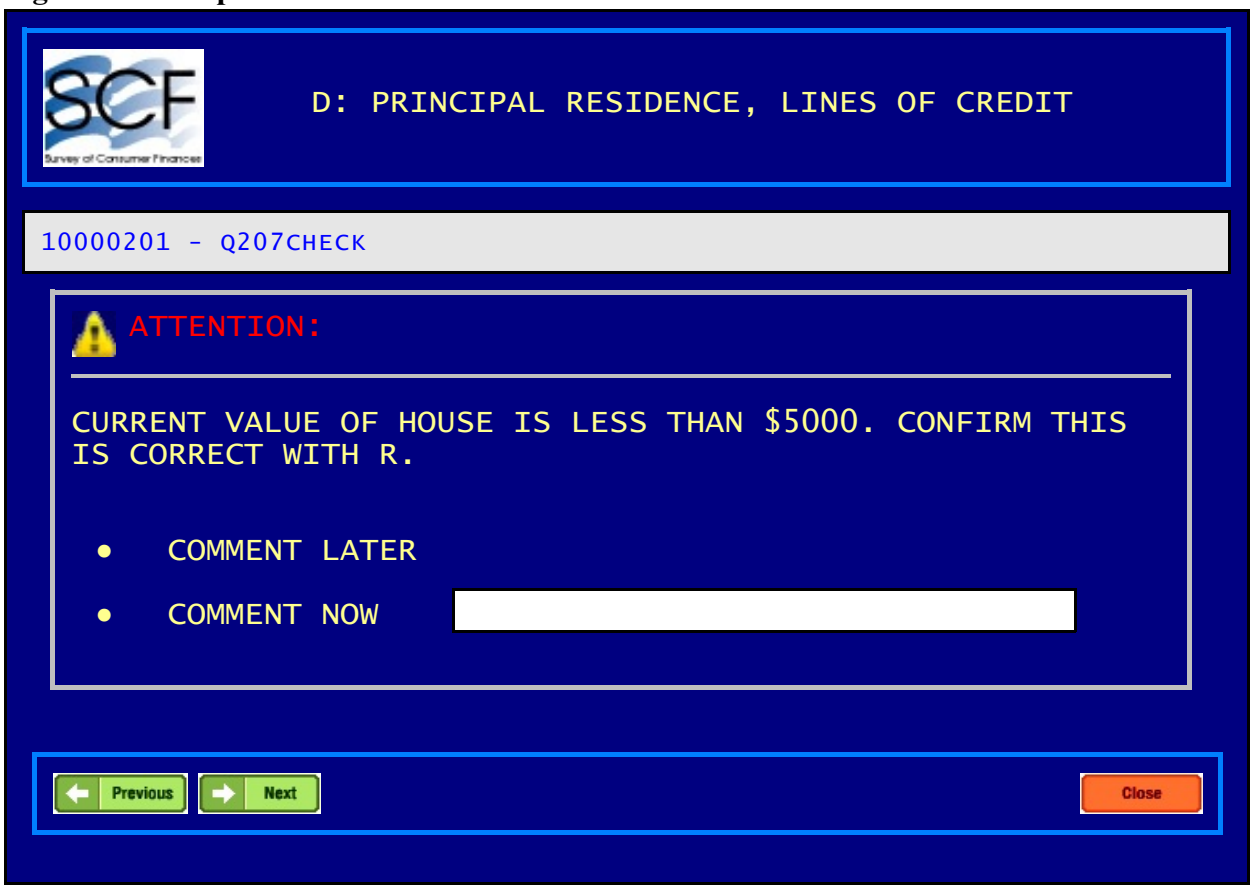
To reduce further the complexity of the interviewer's task, the computer is used to guide some key follow-up interactions with the respondent and to detect some types of response error. Of particular importance in this survey focused on financial data, there is a tool used for all questions with a dollar-denominated response. In the event that the respondent is unable or

[8]Pre-specified probes are also used in places where the question text can be effectively split into two parts—one a part expressing a general concept and the other expressing a more detailed refinement of the concept that would only need to be read if the first part applied. Such questions could be split formally into two separate questions, but they are kept together to keep the framework clear to both interviewers and respondents.

unwilling to give an answer, this tool guides the interviewer in probing for a range, which may take several forms, including an open interval (e.g., "less than $5,000). In all cases, this routine produces a "confirmation screen" that displays the single or range response in words for the interviewer to read back to respondents to ensure that the amount has been captured correctly. This approach has been highly effective in reducing entry errors, particularly for large values.

Like many other surveys, the SCF includes "hard checks"—instructions to the interviewer to correct an impossible data value in order to proceed—and "soft checks"—instructions to the interviewer to confirm or correct an unlikely response in order to proceed. Such checks are most useful with there is a variable whose reasonableness depends either on no other variables or on a variable very close by in the interview. With the reprogramming of the SCF instrument, a more sophisticated generalized edit facility was introduced. As implemented, a screen appears when a logical condition is met, and the interviewer has the option of resolving or explaining the situation at that point or deferring an explanation to the interviewer debriefing instrument

Figure 1: Example of an edit screen in the 2007 SCF.



associated with the case. The motivation for allowing the explanation to be deferred is that there may be times when there is a difficult or very busy respondent who will not tolerate the delay required for the interviewer to record an explanation, or even to probe for an explanation. Such screens were used throughout the interview. Owing to larger difficulties in the initial programming of the questionnaire, it was only feasible to introduce a limited number of such tests in the 2007 survey as a test of the concept. The decision of which screens to include was based on the results of the editing of earlier rounds of the SCF to identify areas that appear to be most robustly identified through logical comparisons of at most intermediate difficulty. Figure 1 gives an example of the content of a screen that would appear if the respondent reports owning a principal residence (other than a mobile home) of very low value.

In addition to being given these structured required comments to complete, interviewers are trained to record comments whenever the respondent provides information that clarifies a reported value, where there are questions about what should be done in the interview, or wherever the interviewer feels additional information would be useful. Such comments are entered in a pop-up box that appears when a computer function key is pressed; the information entered there is tagged with the case ID and the question number. There is also a set of terminal text data fields in the main interview where the respondent can make comments about areas that seemed difficult, comments about anything that was omitted or misclassified, and any other remarks the respondent would like to share.

As noted earlier, the interviewer is required to complete an electronic debriefing questionnaire for each completed interview as soon as possible after leaving the respondent, so that the information requested there is fresh in the interviewer's memory. The items in the debriefing include any edit questions deferred from the main interview, a indicators of the respondent's level of engagement with the interview, a description of any records used by the respondent in answering questions, and a set of open-ended fields for reporting other general information about the interview. In general, interviewers are asked to provide a brief discussion of the survey family and the progress of the interview, to summarize specific problems or questions that arose during the interview, and to provide any other information that in their view affects the reliability of the information reported in the interview.

As in most surveys, all SCF interviewers receive regular feedback on administrative matters and the level of effort they devote to their cases. Beginning with the 2004 SCF, a system was introduced to provide case-specific feedback to the field on interview data quality as well. Although this system did show positive effects, for a variety of reasons—not least that the system was new and more difficult to use than was necessary—its application was uneven. The implementation for the 2007 was intended to make this information an integral part of field operations.

Two sorts of feedback are given on interview data quality. One sort ("quality metric") is generated automatically when cases are transmitted by interviewers to the central office.[9] The other sort ("data utility review") is produced as a byproduct of a manual review of all of the survey data by subject-matter experts in the project staff at the Federal Reserve Board. As discussed later in this paper, in the 2007 survey the return of the two-part feedback to the field staff was delayed for about the first month of the field period.

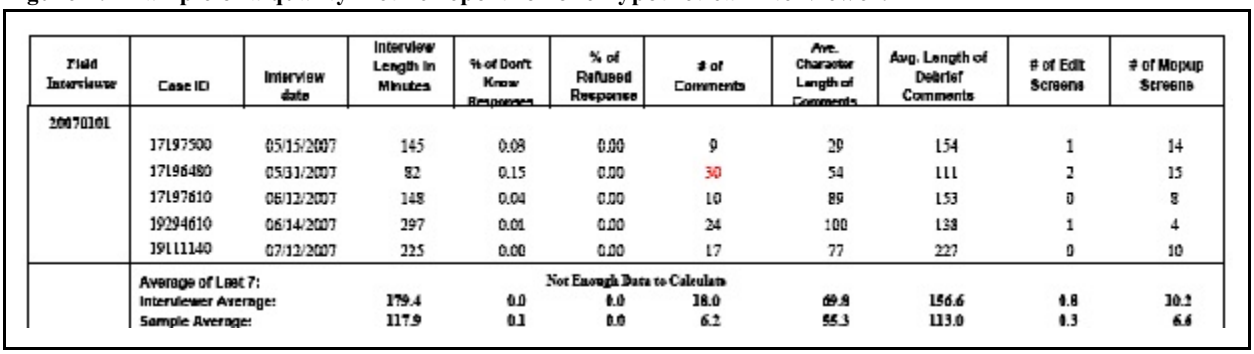
As implemented for the SCF, the quality metric system calculates the interview length, the percent of answers recorded as "don't know" or "refuse," the number of bytes of descriptions entered as comments during the interview or as responses in the debriefing, the number of times an edit screen is triggered in the main interview, and the number of times an interviewer breaks out of a loop of questions before completing the loop.[10] Clearly, each of these characteristics could be affected by the behavior of the respondent, but experience with the 2004 prototype

---

[9]See Wang and Pedlow (2005) for a discussion of the prototype of this system developed for the 2004 SCF and Jodts, Lane and Thompson (2007) for a discussion of the system used for the 2007 SCF. In the future, this system in its generic form will be a corporate standard at NORC available to all studies.

[10]Most looped question sequences in the SCF begin with a filter question about whether a particular item applies to the family, followed by a question on the number of such items. Normally, the interviewer would ask detailed questions on a limited number of items, and then collect summarized information ("mop-up") on all other items. Sometimes the initial counter of the number of iterations turns out to be in error and sometimes the respondent may exert strong pressure on the interviewer. The program allows the interviewer to break out of the loop to the mop-up in such situations. This break is considered a serious deviation from the survey protocol, and as such is intended to be justified in the debriefing interview. Past experience indicates that there is a tendency for some interviewers to over-use this feature.

Figure 2: Example of a quality metric report for one hypothetical interviewer.

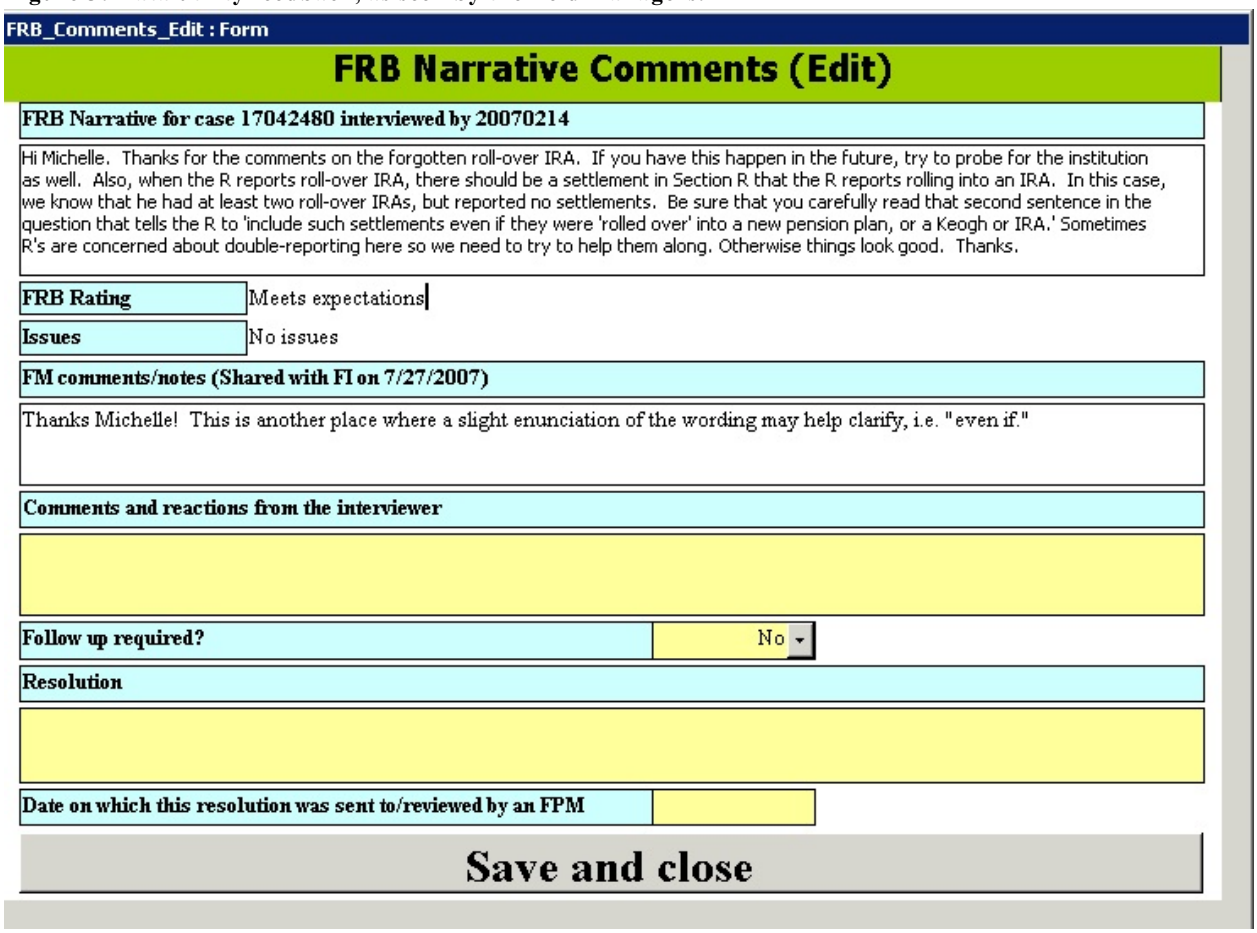


Field Interviewer	Case ID	Interview date	Interview Length in Minutes	% of Don't Know Responses	% of Refused Response	# of Comments	Ave. Character Length of Comments	Avg. Length of Debrief Comments	# of Edit Screens	# of Mopup Screens
20070101										
	17197500	05/15/2007	145	0.08	0.00	9	29	154	1	14
	17196480	05/31/2007	82	0.15	0.00	30	54	111	2	15
	17197610	06/12/2007	148	0.04	0.00	10	89	153	0	8
	19294610	06/14/2007	297	0.01	0.00	24	100	138	1	4
	19111140	07/12/2007	225	0.00	0.00	17	77	227	0	10
	Average of Last 7:				Not Enough Data to Calculate					
	Interviewer Average:		179.4	0.0	0.0	18.0	69.8	156.6	0.8	10.2
	Sample Average:		117.9	0.1	0.0	6.2	55.3	113.0	0.3	6.6

showed that when placed in a distribution of the performance of other interviewers, outlying values of these indicators tend to reflect issues in deeper dimensions of data quality. In addition, they serve the very useful purposes of signaling to interviewers on a regular basis that the quality of their data is important and that it is being monitored. The mechanically generated performance measures are aggregated on a weekly basis and formatted into a simple form for use by the interviewers' supervisors (see figure 2 for an example for one hypothetical interviewer). An interviewer whose performance either differs greatly from that of other interviewers or falls below a critical level, is examined by the supervisor during regularly scheduled weekly calls to review the interviewer's performance.

Although the quality metric system is intended to provide early indicators of problems in interviewers' work, it cannot (yet) address deeper issues in the administration of interviews or interpret and act on interviewers' comments. The data utility review attempts to evaluate the effects of the interviewer's performance on the usability of the data. The editor for an interview draws on several types of information for a case, including specially formatted versions of the interview data, interviewer's comments, the interviewer's record of all verbatim answers provided by the respondent, and the debriefing information, along with a list of potential problems identified by an intensive computer review of the case. In addition to specifying any necessary edits to the case, the editor assigns a score for the interviewer's performance on the case and writes a brief evaluation of both the strengths and weaknesses of the interviewer's work on the case. The scores and accompanying evaluations for every case reviewed are transmitted weekly to both the field managers and the interviewers, and that information is discussed during the weekly performance evaluation calls. Figure 3 provides an example of the feedback given on

Figure 3: Data utility feedback, as seen by the field managers.



a particular case, as seen by the field manager; the version seen by the interviewer contains the same information, but lacks the options that specify a required action.

The case-specific scores assigned by the subject-matter experts indicate the seriousness of unsuccessfully treated problems (table 2). The intent is that the scores reflect the success of the interviewer in addressing problems, but it is inevitable that more difficult cases often offer more ways in which an interviewer might make mistakes. In the most serious instances (score=1), the interviewer could be asked to recontact the respondent to obtain clarifying or missing information; in some instances the interviewer could be required to repeat the entire interview with a different (correct) respondent. In such instances where the respondent cannot not be recontacted, a particularly problematic case might be dropped from the analysis data set and the interviewer would lose the "credit" for the nominally completed case. Many times an interviewer will remember the relevant details of a case (that were incorrectly not recorded in the interviewer debriefing) and help to resolve the critical problems.. A score at the other end of the spectrum (score=4) indicates either that a case has at most minor problems or that it has problems for which the editor thought the interviewer bore no meaningful responsibility.

Table 2: Definition of case-level data quality score.

1: High priority problem in interviewer's handling of case
2. Medium priority problem
3. Minor problem
4. No important problem

The data utility review is time consuming and the number of reviewers is small. Thus, it is not possible always to keep pace with the interviewers in the field, particularly early in the field period when the rate of case completion is relatively high. Nonetheless, is possible to keep up with a selection of cases most likely to be problematic and to ensure that least a selection of the work of all interviewers is regularly reviewed. All cases are ultimately reviewed.

Provision of the two types of feedback to the interviewers is expected to have several effects. First, feedback provides continuing education on how to administer a questionnaire successfully. Second, points of confusion identified among multiple interviewers can be used to provide general clarification to all field staff. Third, the process changes the incentives that the interviewers face from those related only to completing interviews and doing so efficiently, to a

broader set that encompasses multiple dimensions of data quality. The overall effect should be to increase the interview data quality.

III. Evaluation of Data Quality Monitoring

As noted above, problems in the initial phase of the 2007 SCF interfered with the implementation of the pln to give feedback to the field. Data export problems, resulting in part from the full reprogramming of the CAPI instrument, caused a delay of a about a month after the start of data collection (table 3). Entirely fortuitously, this delay combined with the schedule of interviewer training to provide potential opportunity for an experiment to test the short-term effects of feedback to interviewers.

The project interviewers were trained in two groups. Approximately two-thirds of the 194 interviewers were trained in the first session and the remainder ended their training a month later. The first group had about a five-week lag from the end of their training until they began to receive comments from the data utility review. Because the first delivery for this group comprised so many weeks of work, with the available resources it was not possible to review every case within the week available until the first delivery of comments to the field. Thus, a sample of cases was drawn from the first delivery, and at least one interview was reviewed for all interviewers who completed at least one case.[11] From that point forward, at least one case was edited for each interviewer each week.

Table 3 x: Significant dates in the implementation of the feedback system in the 2007 SCF.

End of training #1	May 5
End of training #2	June 4
List sample released to field	June 11
First data delivery	June 6
First utility review feedback: based on data up to 6/7, only training #1 interviewers	June 15
Second utility review feedback: based on data up to 6/14, training #1 & #2 interviewers	June 22
Subsequent utility review feedback	Weekly
First quality metric report	June 29

[11]The remaining cases from this group will be edited later when the flow of new cases diminishes sufficiently.

Table 4: Characteristics of interviewers from trainings 1 and 2: experience and type of home area.

	Training	
	1	2
Experience as an interviewer		
New to NORC	20.6	38.3
Interviewing experience	13.5	25.5
New to interviewing	7.1	12.8
NORC experienced	79.4	61.7
SCF experienced	24.8	10.6
Not SCF experienced	54.6	51.1
Type of home area		
Large metropolitan area	53.7	64.5
Other MSA	27.2	22.6
Non-MSA	19.1	12.9
Number of interviewers	145	50

Because of the closing date for construction of data files for delivery, no cases from interviewers for interviewers from the second training until about 10 days after the close of their training. However, every case for every interviewer in that delivery was reviewed for return to the field on June 22nd. Subsequently, all or most of those interviewers' cases were reviewed weekly for at least the next month.

The quality metric report was first returned to the field almost two months after the end of the first training in a limited form. Thus, there is the possibility of using the first approximately five weeks for the two groups to gauge the returns to the utility feedback, using the interviewers from the first training as a control and the those from the second training as an experimental group. To do so directly requires that other factors be constant. In addition to the trainings and basic operational procedures, the distributions of interviewer "types" and "types" of cases and the patterns of case assignments should be the same. However, this was not a randomized experiment, so in order to draw a conclusion, differences must be explored and if necessary controlled for.

Two trainings were virtually identical except for two things. First, a few minor adjustments were made in the second training to deal with presentational issues that became apparent in the first session. Second, simply because the second training was smaller, some interviewers were more acutely observed by the management team and the observers from the Federal Reserve Board. Overall, it seems likely that any differences from these sources should be negligible.

There are somewhat larger differences between the interviewers who were trained at the two sessions (table 4). Those trained at the first session were more likely to be experienced on the SCF, less likely to be new to NORC but only slightly more likely to be experienced with

NORC on studies other than the SCF. It seems sensible that most of the SCF-experienced interviewers would have been trained first, since *a priori* their productivity in the field should be highest. Interviewers from the second training were more likely to be from one of the largest metropolitan areas, where interviewer turnover is typically greatest.

A straightforward action that would have simplified the analysis of this accidental experiment would have been the creation of random replicate groups of cases reserved for each of the two groups of interviewers. Unfortunately, case assignments were largely completed for the first group by the time it was clear how long the data delivery delay would be. By the time the interviewers from the second training began work, many of the cases in their assignments had already been worked by other interviewers. In addition, the list sample was released a week after the second training ended; cases from this sample are excluded from the remainder of this discussion (except where explicitly noted). Of the area-probability sample cases completed in the first five weeks after the first training, 6.6 percent had previous been worked by another interviewer. The corresponding figure for those from the second training was 47.3 percent. During their first five weeks of work, the group from the first training completed 1,174 area-probability cases (an average of about 8 cases per interviewer), whereas those from the second training completed only 165 cases over a comparable period (an average of about 3 cases). Even including the list sample cases completed by the second group only raises the total to 233 cases. Both the additional effort expended on the later cases and the lower completion rate by the second group suggest that these were more likely to be difficult. Thus, some control for the characteristics of the cases is likely to be necessary in order to evaluate the short-term effects of feedback.

Assignment of area-probability sample cases to interviewers is normally based on geographical proximity, with some lesser consideration of matching 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 other field staff who specialized in securing the cooperation of respondents who had initially refused to participate. There is little data to bring to bear directly on assignment decisions, so analysis of the effect of feedback depends on sufficienc of observed case-specific characteristics.

Table 5: Case-level scores from data utility review, by interviewer training; percent of completed area-probability interviews.

	Case-level score			
	1	2	3	4
All	6.8	29.1	49.9	14.3
Training 1	5.6	29.2	50.1	15.2
Training 2	10.3	28.6	49.2	11.9

Assignment of area-probability sample cases to interviewers is normally based on geographical proximity, with some lesser consideration of matching 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 other field staff who specialized in securing the cooperation of respondents who had initially refused to participate. There is little data to bring to bear directly on assignment decisions, so analysis of the effect of feedback depends on sufficient of observed case-specific characteristics.

At the crudest overall level, the group in the first training produced interviews with higher data quality (table 5). Based on the cases edited as of the time this paper was written, more interviewers in the first training had scores in the best category (4) and fewer in the worst (1) than the interviewers in the second training. Their scores in the other two categories were approximately the sane. To adjust for nonrandom factors, a simple ordered probit model was estimated. The model includes a dummy variable for training group and controls for whether a case had been worked previously by another interviewer, the type of area (large urban area, MSA, or non-MSA), area of the country and interviewer experience. This model suggests that there was no significant difference in data quality between the two training groups. The only statistically significant effect identified by the model is a tendency for non-NORC experienced interviewers to produce interviews of lower quality. It is worth bearing in mind that these results are based on only a selection of cases for interviewers from the first training and the model is fairly primitive.

The final version of the paper will present estimates of a more extended model based on the final edited data. Additional tests will be done extending the evaluation period to include several additional weeks, to allow for the possibility that feedback takes time and repetition to penetrate; this seems likely given the lag in returning cases to interviewers during the busiest part of the field period.

Some factors have emerged from the detailed editing underlying the utility feedback. Perhaps the most important feature that distinguishes interviewers who routinely achieve relatively high score from those who do not is their ability to follow instructions during a complex interview; this fact will be important in designing future tests for interviewer recruiting. In addition, the interviewer-level analysis of cases allowed the discovery of misunderstandings, both by the individual interviewer and larger groups; feedback both through case-level review and general feedback to all field staff through a project newsletter was clearly effective in substantially reducing some specific problems.

Table 5: Case-level scores from data utility review, by interviewer training; percent of completed area-probability interviews.

	Case-level score			
	1	2	3	4
All	6.8	29.1	49.9	14.3
Training 1	5.6	29.2	50.1	15.2
Training 2	10.3	28.6	49.2	11.9

Even if feedback were entirely ineffective in directly changing interviewers' behavior, the process of review has been beneficial in two main ways. First, field managers and interviewers have been presented with an additional standard that is routinely examined. Although the formal evidence at this point is weak, there is a general belief among both the field staff and the project management staff that the revelation of information previously hidden to the field staff has changed the understanding of what it means to be a productive. Managers and interviewers have striven to increase the quality scores. Second, the quality scores have been used as a key factor in decisions about which field staff to retain after the initial phase of field work; interviewers with low scores were released from the project.

Athey, Leslie A. and Arthur B. Kennickell [2005] "Managing Data Quality on the 2004 Survey of Consumer Finances," working paper
http://www.federalreserve.gov/pubs/oss/oss2/papers/Athey_Kennickell_101805.pdf

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.

Jodts, Eric, Julia Land and John Thompson [2007] "Implementing Quality Metrics in Survey Research," paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, Utah.

Kennickell, Arthur B. [2002] "Interviewers and Data Quality: Evidence from the 2001 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 2002 Annual Meetings of the American Statistical Association, New York, NY.

Kennickell, Arthur B. [2003] "Interviewers and Data Quality: Evidence from the 2001 Survey of Consumer Finances," working paper,
http://www.federalreserve.gov/pubs/oss/oss2/papers/asa2002.full.pdf.

Kennickell, Arthur B. [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.

Kennickell, Arthur B. [2006] "How Do We Know If We Aren't Looking? An Investigation Of Data Quality in the 2004 SCF," working paper,
http://www.federalreserve.gov/pubs/oss/oss2/papers/asa2006.3.pdf.

Kennickell, Arthur B. [2005] "Darkness Made Visible: Field Management and Nonresponse in the 2004 SCF," working paper
http://www.federalreserve.gov/pubs/oss/oss2/papers/asa2005.5.pdf.

Kennickell, Arthur B. [2001] "Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances," working paper
http://www.federalreserve.gov/pubs/oss/oss2/papers/scf2001.list.sample.redesign.9.pdf.

O'Muircheartaugh, Colm, Stepahie Eckman, Charlene Weiss [2002] "Traditional and Enhanced Field Listing for Probability Sampling," *Proceedings of the American Statistical Association Social Statistics Section*, pp. 2563-7.

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.