

How Do We Know if We Aren't Looking? An Investigation of Data Quality in the 2004 SCF

Arthur B. Kennickell
Federal Reserve Board

Paper prepared for the 2006 Annual Meetings of the American Statistical Association
Seattle, Washington

This paper is dedicated in memory of Pat Doyle.

September 1, 2006

Abstract

“Data quality” is a term often used fairly vaguely to indicate some cluster of desirable traits of a set of data. This paper argues that the most useful notion of data quality is one that turns on the utility of data for the analytical purposes for which they were originally intended. Although the standard attributes such as missing data rates are important, other more subtle matters can be critical. Unfortunately, many such factors are quite difficult, or even impossible in the absence of advanced AI technology, to identify mechanically. This paper focuses on an exercise undertaken with the 2004 Survey of Consumer Finances (SCF) with the goal of improving data quality. The SCF data have been routinely reviewed in detail by subject-matter experts since the inception of the survey. However, prior to the 2004 round of the survey, no systematic attempt had been made to integrate the findings from the data review with survey operations. In 2004, a structure was created to allow routine feedback from the data reviewers to the field staff. Following preliminary analysis by Athey and Kennickell [2005], this paper lays out the procedures used and demonstrates the importance of error correction on the ultimate uses of the data.

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, interviewers and respondents 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.

Usually survey designers have at least a core set of analytical purposes in mind for the data to be collected. Such analysis requires, of necessity, that these data be fit for those purposes. The extent to which the information recorded in a survey interview differs meaningfully in this sense from the ideal interview is one compelling definition of interview data quality. Although there are other important aspects of data quality, particularly ones concerned with the representativeness of the set of respondents on whom data are available, this paper focuses on the events that shape the interview data available for analysis.¹

The information retained from a survey interview usually represents all we know about how respondents answered the questions asked, and it may differ in two major ways from what analysts may have wanted to know. First, a respondent may not provide the answer ideally desired by the analysts, because the questions are badly designed, or because the respondent does not hear the questions clearly, is inattentive, lacks knowledge on the subjects of the questions, or gives answers that are not clear. Such problems may also arise because the respondent and question designers have different understandings of words and arrangements of words; even with extensive question testing, it is often not possible to determine a question wording that will be both transparent and effective with all respondents. Second, interviewers may make errors of commission and omission. They may read questions unclearly or with inappropriate emphasis, record answers incorrectly or make other procedural errors during the interview. In some surveys, interviewers are expected to clarify questions to respondents, probe answers that are unclear or not germane to the question, or define items in the questions when the interviewer is

¹See Groves [1989] for a review of a broad range of errors that affect the quality of a survey as a measurement instrument. In an economic framework of interviewers' incentives, Kennickell [2006] considers the data quality issues discussed in this paper as well as the quality of the application of effort toward case completion.

aware of a problem; in such cases, interviewers may also err when they fail to intervene when needed or when they intervene inappropriately.

The Survey of Consumer Finances (SCF), the data source for this paper, expects interviewers to be actively engaged with the respondents to ensure, to the extent possible, that the questions are understood, the respondents are well-motivated, answers are probed as necessary for content and clarity, and that the answers are recorded correctly. It is not generally possible for survey designers to have more than indirect influence over the decisions and actions of respondents, but interviewers are effectively agents of the designers and, thus, assumed to be more controllable.

Interviewers face a variety of incentives that shape their behavior in important ways. Most commonly, they face a strong pressure to complete interviews at the lowest possible cost. Over time, there is selection along this dimension both by workers and data collection organizations. Economic theory would argue that workers who continue as interviewers should have the highest level of persuasive skills available at the wage offered them. Ideally, persuasive interviewers should also collect good data, but without some system of monitoring and feedback, there is nothing more concrete than the hope of a positive correlation between these two desirable traits. In the abstract, a high correlation seems questionable, and Kennickell [2002] finds little correlation in such traits across interviewers for the 2001 SCF. Persuasion depends on interpersonal skills and intuition, whereas successful administration of an interview is more dependent on analytical abilities; often these groups of traits do not come together in equal strength. However, there may well be interviewers who do have a balance of such native abilities, but who have not been given the detailed practical guidance to perform at their best.

Thorough training of interviewers on the objectives and methods of a study is critical, of course. But training cannot possibly cover every situation explicitly. At best, it can provide a clear structure, familiarize interviewers with the resources they have after training, and motivate them to review details periodically and note problems that arise while they work. Alone, it can do nothing to control interviewers who choose not to continue their education or who set a lower standard of behavior with respondents than would be desired by the ultimate data users.

To counter a decline in data quality detected in earlier SCFs, the 2004 survey introduced a new system of monitoring and feedback to interviewers. Reporting regularly to the field about the quality of the information collected makes it possible for interviewers to learn to collect better data and for field managers to make more efficient use of the talents at their disposal.

The first section of this paper provides background on the SCF. As an indication of the importance of data quality management in the survey, the next section examines differences between the edited and unedited versions of the 2004 SCF data. The third section presents the initiative undertaken for the 2004 survey intended to improve data quality. The next section evaluates the changes in 2004 and a final section concludes and points toward further work.

I. The Survey of Consumer Finances

The SCF has been conducted every three years since 1983 by the Federal Reserve Board (FRB), in cooperation with the Statistics of Income Division of the Internal Revenue Service.² Since 1992, data for the survey have been collected by NORC, an national organization for

²See 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.

research and computing at the University of Chicago. This paper mainly uses data from the 2004 survey, along with some results from the 2001 survey to provide context.

The survey is intended to collect data for the analysis of the financial behavior of households and the structure of financial markets relevant for households. To this end, the survey asks detailed questions on individual portfolio items, the financial institutions used by the survey households, current and past employment, pensions, income, demographic characteristics and attitudes. Interviewers use computer-assisted personal interviewing (CAPI) to collect the data from respondents. Although an attempt is made to engage with every survey households in person at least initially, 55.3 percent of the interviews completed in the 2004 survey were conducted by telephone, often at the request of the respondent; the remaining interviews were conducted in person.

The survey sample is a dual-frame design. One part is a national multi-stage area-probability sample of households. The other part is a list sample selected from a set of statistical records derived from tax returns, using a model-based stratification to oversample wealthy families. The two parts of the sample serve somewhat different purposes. The area-probability sample provides robust representation of general financial behavior, whereas the list sample allows meaningful analysis of portfolio items and behaviors that are dominated in monetary terms or prevalence by wealthy families. In addition, the structure of the list sample provides a means of adjusting for unit nonresponse that is correlated with wealth and income. The full sample consisted of approximately 10,000 cases, which were about equally divided between the two sample types.

Unit nonresponse rates in the survey are high relative to those in most other U.S. government surveys. In 2004, 31.3 percent of the eligible respondents in 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

sample either declined to participate or could not ever be reached (table 1).³ Although overall only 30.2 percent of the eligible selected list sample cases participated, the participation rate varied strongly over the wealth-index strata. 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. An attempt was made to reach every respondent in both samples by mail before an interviewer approached them. Additionally, list sample cases were offered an opportunity to refuse the interview definitively by returning a postcard, and 12.9 percent of them did so.

Research indicates that, among other things, nonresponse is positively correlated with wealth. In the end, 4,522 completed cases were available for analysis, with about one third of those cases coming from the list sample.

The CAPI program for the main interview has three features noteworthy for this paper. First, throughout the questionnaire, clarifying instructions and the most relevant definitions for each question are included directly on the question screens, as the research of Dillman [2005] suggests. Second, there is a facility that allows interviewers to make comments at any point during the interview. In their training, interviewers are strongly encouraged to use this feature whenever there is any ambiguity about what is recorded in the main data fields. If the situation

³For the area-probability sample, ineligible units are ones that are not housing units or that are uninhabited; units with temporarily absent residents are treated as inhabited. For the list sample, ineligible units are ones where the respondent is abroad for a 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 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.

Table 2: Reporting status for selected variables, all completed interviews, unweighted, 2004 SCF.

	Inapplicable	As a percent of all responses not known to be inapplicable			
		Complete	Range	Unknown*	Other missing
Have any credit cards	0.0	100.0	NA	NA	0.0
Balance on bank-type cards	23.0	96.1	2.8	0.1	1.0
How often pay off credit card bill	20.5	99.7	NA	0.1	0.2
Value of house (ex. farm/mobile home)	31.2	90.2	8.7	0.0	1.2
Have a mortgage	27.4	99.8	NA	0.0	0.2
Amount borrowed on mortgage	56.1	90.5	5.3	0.3	3.9
Amount of mortgage payment	56.6	93.0	4.4	0.4	2.2
Have a checking account	0.0	99.9	NA	NA	0.1
Amount in 1 st checking account	8.9	85.3	9.4	0.1	5.2
Have certificates of deposit	0.0	99.6	NA	NA	0.4
Amount of certificates of deposit	84.7	75.4	13.0	2.6	9.1
Value of first business	72.1	73.0	14.8	0.8	12.1
Have wage income	0.0	99.7	NA	NA	0.4
Total wage income	23.2	79.9	16.9	0.5	2.8

* "Unknown" designates the situation where a higher-order variable was missing, so it was not known whether or not the variable referenced was actually missing or simply inapplicable.

during the interview makes a comment inadvisable at that time, there is an alternative place to record comments after the interview, which is discussed below. Third, for questions with a monetary response, there is a set of options that allows the interviewers, where necessary, to record a range instead of a complete response. Respondents may report such a range directly, select one from a range card, or provide a range implicitly by answering a sequence of questions in a decision tree. In situations where the respondent answers a dollar-denominated question either "refuse" or "don't know," the program automatically generates text to guide the interviewer in obtaining a range. This procedure was intended to automate and make routine the probing activities that interviewers had been instructed to undertake in the surveys before the introduction of CAPI.⁴

⁴As shown in 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.

Many of the topics covered in the survey, particularly the dollar-denominated variables, are commonly viewed as sensitive, and in some instances the information requested involves technical details that may be or seem complicated. As shown by the examples in table 2, item nonresponse rates in the survey vary. For questions about ownership and attitudes, nonresponse rates tend to be quite low. For monetary questions, the rate of complete responses is lower, but this difference is usually offset in part by a substantial fraction of range responses. All missing data are imputed using an iterative multiple imputation technique that is largely based on a regression-like modeling approach (see Kennickell [1998] for an overview).

Along with the main interview data, each case in the SCF includes several files filled out wholly or in part by the interviewer independently of the respondent. The most important such file for the purposes of this paper is an electronic debriefing instrument that interviewers are required to fill out for every completed main interview after leaving the respondent. The purpose of the debriefing is to provide an opportunity for interviewers to reflect on the quality of information obtained, to clarify situations that might seem confusing without additional information, and to add relevant information obtained from the respondent that may have been omitted from the interview for some reason. Where an interview appeared to the interviewer to proceed without problems or unusual events, the interviewer was instructed to say so in the debriefing. Most of the questions in this instrument are structured by topic, but there is also an open-ended field to use in describing any situation the interviewer thinks is relevant.

In field surveys, given an acceptable instrument and sample design, good interviewers are the best hope for successful data collection. As the agent of the survey designers, the interviewer represents the goals of the project to the respondent and, if the interviewer is successful in gaining cooperation, navigates an interview with the respondent. A skillful, knowledgeable and

highly motivated interviewer will be more persuasive with respondents, and will do a better job of navigating complex interviews. For the SCF, interviewer recruiting focuses on finding people who would not be intimidated by asking people to participate in a financial interview. In the 2004 survey, additional emphasis was given to selecting interviewers who had qualities that would tend to make them better able to collect high-quality data. Training for most interviewers consisted of more than three days of in-person instruction covering general background, questionnaire content, questionnaire administration, persuasion skills and administrative matters.

Over the course of the approximately six-month field period, the 186 interviewers trained fell, reaching 128 by the twentieth week. This attrition reflects a wide variety of situations—from the case of interviewers who completed their assignments and did not wish to travel to those who were terminated because they were unable to complete interviews at a sufficiently high rate or within an acceptable range of costs per case (or, in very rare instances, because they violated a critical part of the survey protocol). The productivity of the interviewers varied greatly. No interviews were completed by five of the interviewers, and only one was completed by each of 16 interviewers. At the other end of the spectrum, the most productive interviewer completed 116 interviews, and the most productive 30 interviewers accounted for nearly half of all completed cases.

The assignment of interviewers to a set of cases was not purely random. Rather it was driven primarily by geographic considerations, 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.

II. Editing and Data Quality

Despite years of evolution of the SCF questionnaire to avoid confusion and accommodate changes in the financial marketplace, and despite care in hiring and training interviewers, data errors of various sorts remain a pressing problem in the SCF. A key element in maintaining the quality of the final SCF data over time has been the allocation of resources to building structures to detect and limit the effects of errors. As noted above, the survey relies on interviewers to document problems they observe during the interview—either in the main instrument as they detect them or in the debriefing they are required to complete for every interview. But interviewers vary in their abilities to make and record observations as well as in their commitment to doing so. The CAPI program for the 2004 survey contained a relatively small set of tests to ensure that some key elements of logical consistency, beyond the basic questionnaire routing, were maintained. The SCF also employs a series of automated checks applied to the data at the FRB to detect more complicated patterns in the data that have been associated with problems in the past.

For each case, the interviewer's comments are formatted together with the results of the automated checks and a variety of other case-specific information; in addition, the main interview data are formatted as a pseudo-questionnaire. A subject-matter expert examines the interview data using the interviewer-provided and computer-generated comments as a guide. Changes to the data specified by the editor are intended to be a minimally intrusive response to the available information. Simple inconsistency is not necessarily a sufficient justification for a change. The better the comments provided by the interviewer in cases with serious problems, the less likely that originally reported information must be set to missing and imputed.

Although all edits are flagged using a system of shadow variables, the complexity of the information used in making edits is extremely difficult to reflect meaningfully in a usable system of codes for those shadow variables. Thus, for example, a variable set to missing because the interviewer said explicitly that the value was a nonsensical response cannot be distinguished from the case where the editor's examination of the interview data indicated that the response was logically impossible. Similarly, when data are moved from one part of the instrument to another based on interviewer comments, an outside user could only see from the shadow variables that some originally reported information had been moved into a set of variables and some other information had been eliminated elsewhere; use of comment data to lower the value of a variable would be indistinguishable from any other edit that did not result in a missing value. Incorporating the judgments of the subject-matter experts and the additional feedback of a final reviewer, the edited data set serves as the basis of all further data processing and analysis.

In the ideal, interview data quality in the survey would be judged against the true answers to the ideal understanding of the questions asked. Such information is not available for the SCF— or for almost any other survey. The alternative used in this paper is to compare the edited and unedited SCF data, with the former serving as a proxy for the true data. However, the comparison is not straightforward, in large part because of the complications introduced by missing data. Thus, imputation of the missing data in both versions of the data set is a practical necessity in such a comparison.

The multiple imputation technique used in the SCF to address missing data normally takes as input a data set that has been fully edited. The process begins by using all non-missing values in the data to estimate various moment matrices, which are then used to compute randomized imputations of starting values for an iterative process. In subsequent iterations, the

most recently imputed iteration is used to compute the moment matrices, which are then applied to “available” conditioning variables (variables originally non-missing or already imputed in the sequential variable-by-variable path within the iteration) to make multiple imputations.

Use of unedited data in this imputation system would have two effects: an indirect effect through distortions in the estimation of the various moment matrices that are inputs to the imputation models, and a direct effect through the conditioning variables used as the basis for imputation of individual missing items. In order to examine the sensitivity of these two effects, two imputations of the full unedited SCF data set were made. One used the final iteration of the edited and imputed data set as input for calculating the required moment matrices, which were applied to the unedited conditioning values to calculate the imputations. The other approach used only the unedited data and went through two iterations of the imputation model.

An enormous set of comparisons could be made. Because data problems often cluster in different areas of the interview, it is surely the case that any detailed analysis of such areas would be affected by editing. However, probing of such small areas would not give a good sense of how editing affected the broader results of the survey. Net worth is an important outcome variable, and because it is constructed from a large number of variables in the survey, it gives a sense of the overall nature of the changes, and by implication of the quality of the original data.

Figure 1 is a quantile-difference plot showing the value at each quantile across the distribution of net worth under the iterated unedited data minus the corresponding value under the edited and imputed data.⁵ At the 25th percentile of the distribution, the value for the unedited

⁵It should be emphasized that the difference shown is not the difference in values of the same observations in the two versions of the data; the relative positions of households in the distribution also may shift as data values change. The dollar values are scaled using the inverse hyperbolic sine function with a scale parameter of 0.0001; this transformation has the property of being approximately linear near the origin and approximately logarithmic at larger values.

data is about \$1,200 below that for the edited data; at the median it is \$1,098 below; and at the 75th percentile it is about \$3,700 above the level of the edited data. Differences are much larger at the extremes of the distributions; at the 1st percentile the unedited value is about \$14,200 below (more negative, in this case) the edited value and at the 99th percentile the unedited value is about \$369,000 above the edited value. As a proportion of the edited values (figure 2), the differences appear somewhat less striking. At the top of the distributions aside from the very top, the unedited values exceed the edited ones by about 10 percent. The profile of relative differences around the median is small and flat. Lower in the distributions, the pattern is more mixed. In the region below the 20th percentile in both distributions the underlying values fall toward zero and continue below; in the region around zero, quite small dollar differences between the distributions become magnified. Still lower in the distribution, many changes were more than 20 percent, the range of the figure.

Figure 1: Quantile-difference plot of net worth: value of distribution from unedited data with iterated imputations minus value of distribution from final edited and imputed data, by percentiles of the distributions, 2004.

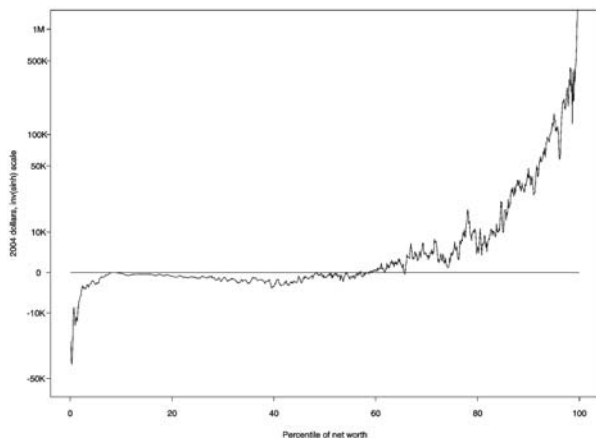


Figure 2: Relative quantile-difference plot of net worth: value of distribution from unedited data with iterated imputations minus value of distribution from edited data, as a percent of value for final edited and imputed data, by percentiles of the distributions, 2004 SCF.

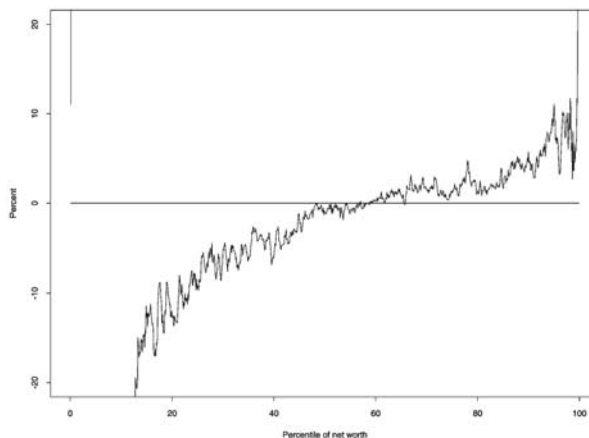


Figure 3: Quantile-difference plot of net worth: value of distribution from unedited data with non-iterated imputations minus value of distribution from unedited data with iterated imputations, by percentiles of the distributions, 2004.

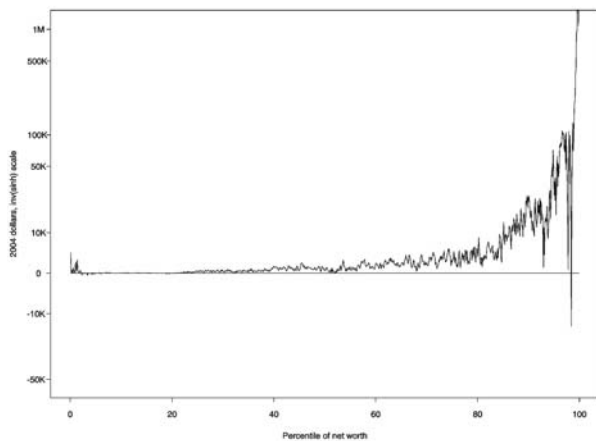


Figure 4: Relative quantile-difference plot of net worth: value of distribution from unedited data with non-iterated imputations minus distribution from unedited data with iterated imputations, as a percent of value for the distribution with non-iterated imputations, by percentiles of the distributions, 2004 SCF.

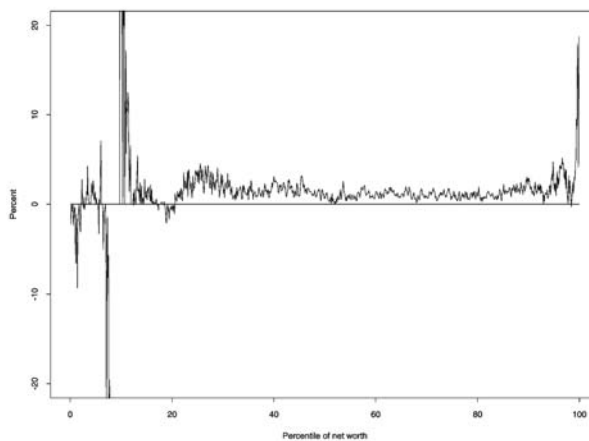


Table 3: Regression of logarithm of total household income using two experimentally imputed unedited data sets and the final imputed edited data set, 2004 SCF.

	Unedited data		Edited data
	Non-iter.	Iterated	
Intercept	5.34*	5.38*	6.97*
	<i>0.60</i>	<i>0.60</i>	<i>0.40</i>
AGE	0.17#	0.17#	0.12#
	<i>0.04</i>	<i>0.04</i>	<i>0.02</i>
AGE2	-0.33	-0.33	-0.22
	<i>0.40</i>	<i>0.39</i>	<i>0.27</i>
AGE3	0.19	0.19	0.12
	<i>0.24</i>	<i>0.23</i>	<i>0.15</i>
D_MARRIED	0.36#	0.36#	0.40*
	<i>0.06</i>	<i>0.06</i>	<i>0.04</i>
D_WORKING	-1.08#	-1.14#	-2.58#
	<i>0.34</i>	<i>0.33</i>	<i>0.43</i>
L_CPSWAGE	0.16#	0.17#	0.28#
	<i>0.03</i>	<i>0.03</i>	<i>0.04</i>
D_FIN	-0.66#	-0.69#	-0.58#
	<i>0.15</i>	<i>0.15</i>	<i>0.10</i>
L_FIN	0.20*	0.20*	0.17*
	<i>0.01</i>	<i>0.01</i>	<i>0.01</i>
D_HOUSE	-2.93#	-2.83#	-2.93*
	<i>0.37</i>	<i>0.37</i>	<i>0.26</i>
L_HOUSE	0.26#	0.25#	0.27*
	<i>0.03</i>	<i>0.03</i>	<i>0.02</i>
D_VEHIC	-0.54*	-0.57*	-0.62#
	<i>0.27</i>	<i>0.27</i>	<i>0.18</i>
L_VEHIC	0.08#	0.09#	0.07#
	<i>0.03</i>	<i>0.03</i>	<i>0.02</i>
D_BUS	-2.25*	-2.29*	-2.55*
	<i>0.26</i>	<i>0.26</i>	<i>0.18</i>
L_BUS	0.18*	0.18*	0.22*
	<i>0.02</i>	<i>0.02</i>	<i>0.01</i>
D_NNRESRE	-0.48	-0.58+	-0.55*
	<i>0.38</i>	<i>0.37</i>	<i>0.25</i>
L_NNRESRE	0.05+	0.06*	0.05#
	<i>0.03</i>	<i>0.03</i>	<i>0.02</i>
D_ORESRE	-2.15#	-1.92#	-1.82#
	<i>0.37</i>	<i>0.37</i>	<i>0.25</i>
L_ORESRE	0.20#	0.18#	0.17*
	<i>0.03</i>	<i>0.03</i>	<i>0.02</i>
D_DEBT	0.15	0.15	0.42#
	<i>0.18</i>	<i>0.19</i>	<i>0.13</i>
L_DEBT	-0.01	-0.01	-0.05#
	<i>0.02</i>	<i>0.02</i>	<i>0.01</i>
R ²	0.51	0.51	0.69

Definitions for table 3.

Dependent variable: $\ln(\max(1, \text{total household income}))$

Conditioning variables:

AGE: age of household head

AGE2: age squared divided by 100

AGE3: age cubed divided by 1000

D_MARRIED: head married or living with a partner

D_WORKING: head currently working

L_CPSWAGE: $\ln(\max(1, \text{sum of total annual wages predicted for head and spouse/partner using CPS data, conditional on detailed occupation, age, education, race, self-employment status, part-time status}))$

D_FIN: =1 if have financial assets, =0 otherwise

L_FIN: $\ln(\max(1, \text{financial assets}))$

D_HOUSE: =1 if own home, =0 otherwise

L_HOUSE: $\ln(\max(1, \text{value of owned home}))$

D_VEHIC: =1 if own any vehicles, =0 otherwise

L_VEHIC: $\ln(\max(1, \text{value of owned vehicles}))$

D_BUS: =1 if own any businesses, =0 otherwise

L_BUS: $\ln(\max(1, \text{value of owned businesses}))$

D_NNRESRE: =1 if own nonresidential real estate, =0 otherwise

L_NNRESRE: $\ln(\max(1, \text{value nonresidential RE}))$

D_ORESRE: =1 if own residential real estate other than home, =0 otherwise

L_ORESRE: $\ln(\max(1, \text{other residential RE}))$

D_DEBT: =1 if have any debt, =0 otherwise

L_DEBT: $\ln(\max(1, \text{value of outstanding debt}))$

Significantly different from zero at: #=1% level, *=5% level, +=10% level.

If the imputed version of the unedited data created using the edited data to estimate the distributional moments used in imputation had been compared with the data resulting from the usual process of editing and imputation, the differences would have been larger. This is seen most clearly by comparing the two imputation versions of the unedited data using quantile-difference and relative quantile-difference plots (figures 3 and 4). Across nearly the entire distribution, the levels corresponding to the unedited data imputed using the edited data as input for the moment estimations lie above the values for the unedited data imputed iteratively. Most likely, this difference results from noisier data used in the iterated version implicitly attenuating the relationships between variables and, thus, lessening the effects of outliers used as conditioning variables.

The gini coefficient for net worth shows substantial differences between the edited and unedited data. The value for the edited data is 0.805 with a standard error reflecting both sampling and imputation of 0.005. The values for the unedited data are much larger than this, but similar to each other—the value for the iterated version is 0.861 and that for the non-iterated version is 0.866. Thus, the unedited values are at least 11 standard errors higher than the edited value.

Univariate statistics on net worth may be particularly sensitive to the changes induced by editing. Perhaps variables taken in a multivariate relationship would be more immune to removal of erroneous data values. Classical econometrics can be taken as encouraging the assumption that measurement error may be taken as sufficiently uncorrelated with the error term in a well-specified regression model to be ignored. Table 3 presents the results of a slightly augmented “accounting” regression of the logarithm of total household income. The dependent variables are largely wealth factors that should be direct generators of income; other wealth

factors that may serve as indirect indicators of income; a measure of labor income predicted from a model estimated using the Current Population Survey conditional on the characteristics of the survey respondent and that person's spouse or partner, including the relevant detailed occupation data; and a set of basic demographic characteristics. Income is a variable that has a moderate amount of editing, often triggered by a discrepancy in the amounts of income components respondents report and the total amount of their income, which is computed by the CAPI program and offered to the respondent for confirmation. The wealth variables used for conditioning include all the major components of net worth.

For most variables in the model, the estimated sign, relative magnitude and multiple-imputation-corrected standard errors are sufficiently similar that an analyst interested in this model would not draw wildly incorrect conclusions from the different versions of the data. However, there are some notable differences across the models. First, the R^2 measure of fit is more than a quarter lower for the two models based on the unedited data. Second, the effects of debt, work status and predicted wage income are larger under the edited data than under the unedited data; in the case of the debt variables, they do not even register as significantly different from zero in the models run on the unedited data.

The point of this section of this paper is to suggest that attention to data quality is a critical issue in the SCF and that editing is a critical step in maintaining data quality. The main way that editing may fail is in the situation where the information available for making decisions is inadequate. In such cases, erroneous values may be left in the data or values may be set to missing because there is no other means of dealing with complicated inconsistencies or other obvious misreporting. The next section deals with other more case-specific and interviewer-

specific measures of quality that address some of the problems underlying the variability seen as a result of editing.

III. Monitoring and Managing Interview Data Quality

In the SCF, as in most surveys, interviewers are encouraged during their training to read questions to respondents and record their answers as carefully as possible. However, because the survey is largely focused on the collection of factual information on technical topics that some respondents may find unclear or confusing, interviewers have routinely been asked to go beyond the basic tasks of reading and recording. A key role of an SCF interviewer is in ensuring that the intended meaning of the questions reaches the respondents and that the answers recorded are truly responsive. To that end, interviewers are trained to use active listening, probing and explanations.⁶ The interviewer is expected to answer explicit requests from the respondent for clarification and to offer clarification without prompting when the respondent appears confused.

Nonetheless, over time it has been difficult to get all interviewers to practice such behavior uniformly and to the desired degree. Because the actual interviewing of field interviewers is, at best, very difficult to observe, developing indirect means of supporting compliant behavior is important. Recognizing the place of interviewers' incentives in field operations, the 2004 SCF introduced a several changes, the most important of which for purposes of this paper is a set of procedures designed to monitor interview data quality and to

⁶Based 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.

provide feedback to interviewers about their performance.⁷ These procedures were in two parts. One involved a computer-generated set of interviewer-specific reports that were available soon after interview data for each case were transmitted to the NORC central office. The other was the product of an intensive data review undertaken by subject-matter experts at the Federal Reserve Board as a part of the data editing; owing to the complexity of this process, the results of this process were generally available at a somewhat longer lag.

For the computer-generated reports, the data for every ostensibly completed case were screened mechanically to determine the proportion of missing values for monetary variables and the number of keystrokes stored in a key set of comment fields in the main questionnaire and the in the open-ended response fields in the debriefing interview.⁸ Missing data have an obvious and basic connection with data quality. Although interviewers cannot force respondents to give answers when they are unable or unwilling to do so, it is clear that interviewers can have a strong influence on respondents' willingness to cooperate. For example, by making respondents more comfortable about the protection of their data, interviewers may need to record fewer missing values.

The comment data have a more indirect connection to data quality. Successful editing of the SCF cases turns critically on the quality of comments that interviewers provide. Interviewers were told in their training and repeatedly thereafter that good documentation of their interviews

⁷See Athey and Kennickell [2005] for a discussion of the monitoring system and preliminary analysis of the resulting data. See Wang and Pedlow (2005) for a discussion of a part of the monitoring system designed for rapid turnaround.

⁸In addition to direct tallying of "don't know" and "refuse" responses to monetary questions, the measure of missing values also includes missing responses to a selection of higher-order questions that might lead to key monetary questions. If the respondent provided range information, the value was not treated as missing.

was essential. As noted earlier, even in situations where a case appeared to the interviewer to be perfectly free of problems, they were instructed to give a brief description of the interview and to say explicitly that there had been no problems; such comments are tremendously useful in editing when data anomalies arise. Although the total length of the comments says nothing directly about the usefulness of what is written, historically, cases with few such comments were very often ones where such information was most needed to achieve satisfactory edits. To reinforce the need to make comments, guidance was given to interviewers on the quality of their comments as a part of the feedback based on data editing.

The mechanically generated measures of interviewers' performance were aggregated on a weekly basis and formatted into a simple form for use by the interviewers' supervisors. An interviewer whose performance either differed greatly from that of other interviewers or fell below a critical level, was examined by the supervisor during weekly calls to review the interviewer's performance.

The second part of the data quality monitoring was driven by the editing of the SCF data by subject-matter experts as described earlier. In addition to specifying the necessary edits to each case, the subject-matter experts assigned a score to the case, which was intended to reflect how well the interviewer coped with the interview, and wrote a brief evaluation of both the strengths and weaknesses of the interview. The scores and accompanying evaluations were transmitted to the survey field managers for use in the weekly calls.

The case-specific scores assigned by the subject-matter experts indicate the seriousness with which the field manager was expected to review each of an interviewer's cases with the interviewer (table 4). In the most serious cases (score=1), the interviewer could be asked to

Table 4: Definition of case-level data quality score.

- | |
|---|
| <ol style="list-style-type: none"> 1: High priority problem with case 2. Medium priority problem with case 3. Minor problem with case 4. No important problem with case |
|---|

recontact the respondent to obtain clarifying or missing information; in some cases the interviewer could be required to repeat the entire interview with a different (correct)

respondent. In such instances where the respondent could 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. 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 for which the editor thought the interviewer bore no meaningful responsibility.

The review process was time consuming and the number of such editors was small. Thus, it was not possible always to keep pace with the interviewers in the field, particularly early in the field period when the rate of case completion was relatively 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 temporarily skipped a week of cases in order to be able to return comments on cases more likely to be 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.

Provision of the two types of feedback to the interviewers was expected to have three effects. First, feedback should have provided continuing education on how to administer a questionnaire successfully. In past surveys, some interviewers had complained that they were never given an indication of how successful they were in collecting useful data and what they could do to improve. Second, the feedback should have signaled to the interviewers and their managers that coherent information is important, that the project staff were watching for

deviations, and that there could be serious implications for collecting data of poor quality. Third, the overall effect should be to increase data quality.

It was learned in the project debriefing held after the end of data collection that interviewers' managers differed in the stringency with which they reviewed the feedback with their interviewers. Although such differences should have led to differences in quality across management groups, the expected overall effect is still positive.⁹ Because the number of productive interviewers was already fairly thin relative to the number of sample areas and the because the number of cases was high relative to the number of interviewers in some areas (such as New 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 that they could be called upon to explain their work to their manager. Moreover, the decision to delete a nominally completed interview was made at a higher level than the field manager; because interviewers' retention was conditional on meeting their production goals, the threat of losing such credit should be have been an important motivating factor.

IV. Evaluation of Quality Control Procedures

Although the intention was that the scores assigned by the subject-matter experts be free of respondent-level effects, this is unlikely to have been purely so. Some interviews were done with families with more complex circumstance than others, and may, thus, have had more

⁹It would be interesting, in principle, to control for manager-level 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 5: Priority score by sample type and list stratum, percent distribution, 2004 SCF.

Score	All	AP samp.	<i>List sample stratum</i>						
			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
All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 6: Distribution across interviews of the percent of missing dollar values, by sample type, 2004 SCF.

Percentile of distribution of number of missing values	All cases	AP cases	LS cases
25	0	0	0
50	6	4	9
75	18	15	24
90	37	33	41

chances to experience problems. One sign that reinforces this issue is the higher frequency of low (bad) quality scores for cases in the wealthier (higher) strata of the list sample (table 5).

The percent of missing dollar values in the rapidly-available feedback also shows considerable variation across cases (table 6).¹⁰ List sample cases, on average, also have more “opportunities” to have missing data because on average they tend to have more complex financial 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.

Comparison of the rate of missing values with the priority score for each case suggests that the two have fairly different patterns (table 7). Although these measures of quality are

¹⁰The denominator for the percentage calculation is the number of variables to which it was known that a dollar response should have been made plus the number to which it was not known because a response to a higher-order question was missing.

Table 7: 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				All
	1	2	3	4	
<i>All cases</i>					
<5%	2.5	5.5	23.3	16.0	47.4
5%–9.9%	0.5	1.8	6.8	3.7	12.8
10%–24.9%	1.4	3.5	10.3	6.2	21.4
≥25%	1.6	4.7	9.0	3.2	18.5
All	6.0	15.5	49.4	29.1	100.0
<i>All area-probability cases</i>					
<5%	2.5	5.3	25.0	18.9	51.7
5%–9.9%	0.4	1.3	7.1	4.1	12.8
10%–24.9%	1.0	2.9	9.2	6.9	20.0
≥25%	1.1	2.8	7.8	3.8	15.6
All	5.0	12.2	49.1	33.7	100.0

negatively correlated (that is, quality by the two measures is positively correlated), the connection is weak—only -0.18 for all cases and -0.12 for the area-probability cases alone. This finding indicates that providing separate feedback based on each measure may be productive.

The levels of all of the simple quality statistics are determined, 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 by the changes in procedures; 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.

¹¹Monitoring should have reduced the levels of the two outcome measures relative to the unmonitored state as well as increased the correlation between them.

Table 8: Percent of all dollar values missing, mean priority score, mean length of interviewer debriefing comments (number of characters), and number of observations; by biweekly interval of the field period; 2001 and 2004 SCF.

Biweekly period	% missing \$ values		Mean priority score		Mean length debriefing comments		Number of observations	
	2001	2004	2001	2004	2001	2004	2001	2004
<i>Full sample</i>								
1	15.6	13.3	NA	3.01	162	310	199	355
2	15.9	11.4	NA	2.88	163	296	389	574
3	12.4	13.1	NA	2.99	179	324	356	504
4	15.9	13.1	NA	3.04	191	385	630	427
5	16.1	12.9	NA	2.94	173	429	411	372
6	17.4	12.6	NA	3.04	191	416	233	359
7	14.9	13.4	NA	2.95	198	391	547	246
8	13.7	13.2	NA	3.07	201	391	332	250
9	13.8	13.3	NA	3.09	196	416	425	263
10	16.6	13.9	NA	3.09	213	405	438	227
11	13.5	11.0	NA	3.10	227	377	252	196
12	16.2	11.3	NA	3.17	193	392	140	240
13	21.0	9.8	NA	2.83	192	436	97	124
14	.	11.2	NA	3.14	.	396	.	167
15	.	11.3	NA	3.14	.	322	.	100
16	.	14.0	NA	3.15	.	441	.	110
<i>AP Sample</i>								
1	15.6	12.6	NA	3.01	162	292	199	295
2	15.9	10.3	NA	2.97	163	278	389	446
3	12.4	11.4	NA	3.05	179	300	356	369
4	15.4	11.9	NA	3.11	163	350	442	262
5	15.7	11.4	NA	3.03	165	368	265	241
6	14.2	10.2	NA	3.10	179	398	141	209
7	14.5	11.2	NA	3.05	162	336	307	140
8	10.2	10.1	NA	3.17	166	415	168	154
9	11.2	12.4	NA	3.24	154	385	213	138
10	14.2	12.2	NA	3.26	158	311	214	139
11	11.0	10.9	NA	3.28	142	321	102	124
12	15.1	10.1	NA	3.32	179	311	76	163
13	25.2	9.0	NA	2.99	156	302	45	72
14		8.3	NA	3.33	.	357	.	119
15		8.9	NA	3.33	.	262	.	70
16		11.3	NA	3.54	.	332	.	65

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 periods to support meaningful separate analysis here.

Table 8 provides a summary of the available data for the 2001 and 2004 surveys across the set of biweekly intervals of the field period. The overall 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 considerably higher in every period in 2004 than in 2001. The fact that the amount of comments also rose more sharply over the first intervals in 2004 suggests that monitoring and feedback had an immediate effect by reinforcing the message interviewers received in their training.

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

Table 9: 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

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 9).

One might expect that cases would tend to become *more* difficult over the field period, and thus increase the likelihood of *lower* scores. Adjusted for difficulty, the pattern might be stronger. To make such an adjustment, the quality scores were filtered using a regression technique. The full set of scores was regressed on 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 was 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.

Table 10: Mean unadjusted and adjusted quality scores, by biweekly interval of the field period, as a percent of overall mean score of each type, 2004 SCF.

Biweekly period	Unadjusted	Adjusted
1	99.87	96.86
2	95.56	93.55
3	99.16	97.52
4	100.74	100.17
5	97.35	96.53
6	100.69	100.83
7	97.74	98.18
8	101.88	101.82
9	102.39	103.14
10	102.42	102.15
11	102.71	100.83
12	105.02	104.13
13	93.88	94.22
14	104.06	103.14
15	104.14	105.13
16	104.32	103.47

Very similarly to the case with the unadjusted scores, the relative pattern of means of the adjusted scores shows a mild rise in data quality over the field period (table 10). The fact that the adjustment makes so little difference could indicate that the setting of the score was largely successful in allowing for the difficulty of each case; alternatively, it might reflect omission of other important variables in the model or selection effects generated as the pool of interviewers shrank 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

Table 11: Distribution of cases over rounded mean standardized quality scores and case completion groups, 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	3.3	0.6	0.6	11.1
3	24.3	21.0	21.0	15.5	81.8
4	4.4	0.0	0.0	0.0	4.4
All	38.1	24.3	21.6	16.2	100.0
<i>Memo items:</i>					
Std. dev. of standardized quality score over interviewers in group					
	0.693	0.332	0.216	0.217	0.484
Mean of standardized quality score over interviewers in group					
	2.82	2.91	2.93	3.10	2.91
Number of interviewers in group					
	69	44	39	29	181

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 they should be related. In the case of the 2004 survey, across interviewers at all levels of case completion, there is a distinct concentration of the standardized quality score around 3, a level reflecting small and non-critical concerns with the data (table 11). In seeming contrast to the conclusion of the earlier study, variability around this point was least for interviewers with a relatively large number of 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.

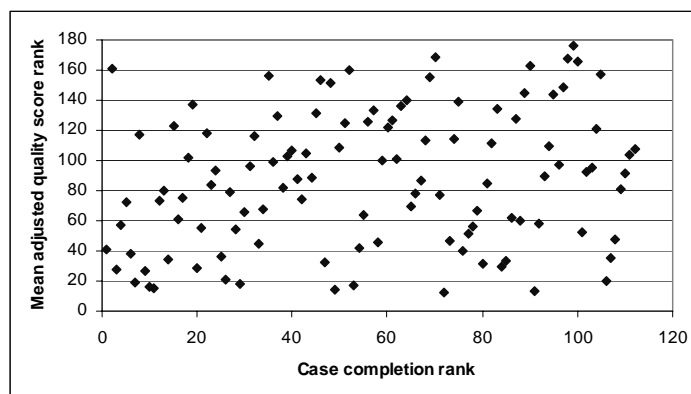
Table 12: Distribution across interviewers of parameter on biweekly reporting period in a regression of standardized quality score on reporting period, for interviewers completing various numbers of cases.

%ile	<i>Number of completed cases</i>		
	26-50	>50	>25
99	0.267	0.067	0.267
95	0.123	0.054	0.102
90	0.102	0.055	0.090
75	0.076	0.033	0.047
50	0.034	0.008	0.017
25	-0.005	-0.003	-0.005
10	-0.024	-0.022	-0.022
5	-0.058	-0.027	-0.039
1	-0.092	-0.039	-0.093
Mean	0.036	0.013	0.026
Std. dev.	0.061	0.027	0.051

If the standardized quality score for each case is regressed by interviewer on the biweekly wave number, the result is an estimate for each interviewer of the trend level of quality change over the field period. Table 12 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. However, 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.

However, a more detailed examination of the data reveals that although interviewers who were highly productive in terms of completing interviews do show somewhat higher data quality,

Figure 5: Interview rank in terms of number of completed interviews vs. interviewer rank in terms of mean adjusted data quality score; interviewers completing more than 10 cases, 2004 SCF.



as a group (confirming the finding noted above), this relationship clearly does not hold for the entire group of such interviewers (figure 5).¹² Indeed, the interviewer who completed the second-largest number of interviews had one of the worst average adjusted data quality scores. For the group away from the top producers, the impression is one of little or no correlation. Among interviewers who completed at least 11 cases, the rank correlation between an interviewer's number of completed interviews and mean adjusted quality score was only 0.23; among interviewers who completed between 11 and 25 interviews, the correlation was only 0.10.

V. Conclusion

This paper focuses on data quality in field surveys, using data from the SCF. As indicated by an analysis of the editing process for the 2004 SCF, data quality is a pressing concern in the survey. The paper argues that interviewers have a critical role in the collecting

¹²In the figure, a completion rank closer to 1 indicates a larger number of completed cases and a data quality rank closer to 1 indicates a higher level of data quality. Interviewers who completed only 10 cases or fewer are excluded because the mean quality score cannot be reliably calculated for this group.

high quality data and in facilitating successful editing where data collection is less successful. In a field survey, interviewers are normally the only project staff who have direct contact with respondents. Interview protocols developed to promote data quality depend on the compliance of interviewers. Because interviewers' performance in the field is usually only indirectly observable, their abilities and motivations are critical factors. Desirable behavior can be shaped through hiring decisions, training, and the management of the incentives interviewers face. The strand followed here is the creation of incentives for interviewers through the use of monitoring and feedback during the field period.

The 2004 survey employed two types of feedback to interviewers on their performance in collecting data. First, a pair of indicators that could be computed directly from the raw data were available quickly for managers to use in their regular performance reviews with their interviewers. The indicators were the percent of dollar-denominated missing values in an interview and the number of characters typed into a set of key comment field available to interviewers during and after the actual interview. The latter type of information is an important input into the data editing process for the SCF. Second, the data collected, along with any comments and a variety of auxiliary information, were examined by subject matter experts while editing the data. For each case, the editor assigned a data quality score and wrote a short report on the interviewer's performance. The report provided praise and constructive criticism as appropriate for each case. Because the editing process is time consuming, the score and the report were available to the interviewers' managers with a lag. Despite some problems in the design and implementation of the system of monitoring and feedback, these process appears to have had the desired effect of bolstering the quality of the data ultimately produced for analysis. Nonetheless, it is clear that problems remain.

A key finding for future work is that even with this process in place, the relationship between interviewers' success in completing a large number of interviews (the classic measure of an interviewer's performance) is only loosely correlated with their performance in terms of collecting good data. The incentives interviewers face relative to each of their tasks are important. If after a period of mentoring an interviewer fails to secure enough interviews, the consequence is termination, a very salient outcome; rewards come to interviewers who are successful in this dimension. Management of data quality requires a similar management framework. Although data quality is not so easily and objectively measured as case completion, as this paper shows, it is possible to develop useful instruments for data quality management, some of which may need to be project specific, as is the case for the SCF. Without some metric that can be shared with interviewers and used by managers, data quality improvement is an exercise in wishful thinking.

One strategy to be followed in the 2007 SCF is to make the idea of data inconsistencies more salient to interviewers by highlighting a set of actual or potential problems for them during the interview. Unlike the traditional "hard checks" and "soft checks" executed in CAPI programs, these tests will offer three options. First, the interviewer will be able to correct erroneous information. Second, the interviewer will be able to write an explanation of a seeming inconsistency during the interview. Third, the interviewer will be able to defer a response until the debriefing interview the interviewer is required to complete alone after each interview. In addition, during the debriefing interview, all comments made during the interview, including the responses to such tests, will be available to allow the interviewer to evaluate the adequacy of the documentation and to add any additional clarifying information that might be required. In addition, the incidence of such screens can be used as another quality monitoring tool.

Finally, a possibility that should be considered in the future is splitting the tasks interviewers are currently required to perform into those associated with persuading respondents to participate and those associated with conducting an interview. Informally, and most often to a limited degree, this practice is followed in many surveys. Persuasion skills for interviews have a great similarity to those appropriate in selling any other type of idea or product; interpersonal skills are highly important in such work. Although success in conducting an interview also requires persuasive skills, the main skills are technical ones—reading, listening, comparing answers with the intent of the question, responding to or even anticipating questions appropriately, recording information within the framework of the questionnaire, and navigating the surrounding computerized or paper apparatus required to complete an interview. By splitting the persuasion and interview administration tasks, it would be possible to hire interviewers with more specific abilities and to provide more intensive training on the issues more pertinent to each task. If feasible, this process could offer quality improvements in both phases of interviewing.

A simple experiment conducted in the planning stage for the 2004 SCF explored the possibility of using field interviewers to gain respondents' cooperation and then using another interviewer to administer the interview by telephone. This test indicated that there might be serious logistical problems in arranging for a sufficient pool of interviewers who would be ready to work at any time, and that there were equally serious questions about how incentives could be structured to give the persuasion-focused interviewers a financial and personal sense of accomplishment that is normally provided by completing an the entire interview process. Because of the possibility of strong quality gains, work should continue here.

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.
- Dillman, Don A., Arina Gertseva and Taj Mahon-Haft [2005] “Achieving Usability in Establishment Surveys Through the Application of Visual Design Principles,” *Journal of Official Statistics*, Vol.21, No.2, 2005. pp. 183-214.
- 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>.

_____ [2006] “Who’s asking? Interviewers, Their Incentives, and Data Quality in Field Surveys,” paper presented at the 2006 meetings of the International Association for Research in Income and Wealth, Joensuu, Finland.

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.