Measuring Data Quality In the 1998 Survey of Consumer Finances[1]

Arthur B. Kennickell, Federal Reserve Board
Mail Stop 153, Federal Reserve Board, Washington, DC 20551

Key words: Data quality, Editing

Survey data quality is notoriously difficult to measure. The analysis here focuses on the effects of data editing on data quality in the 1998 Survey of Consumer Finances (SCF). Editing in the SCF is driven by a variety of types of text data provided by interviewers, and by a priori logical and institutional consistency requirements. Each interviewer comment and each possible logical or institutional inconsistency was examined to determine whether the data should be changed as a result. Data changes after the interview are often highly inefficient and sometimes such changes entail the generation of substantial amounts of missing data.

I. Data

The SCF is designed to collect detailed information on the assets, liabilities and other financial characteristics of families, and on other variables that are analytically useful in interpreting that information (see Kennickell, Starr-McCluer and Sundén, 1997). The full set of SCF questions supports a total of about 3,000 final variables, but because of the skip patterns of the questionnaire, it would not be possible for any given respondent to provide information on all of these. Indeed, the maximum number of responses given in any 1998 SCF interview was about 780. The average interview in 1998 lasted about 75 minutes.

The participants in the 1998 survey include 2,813 respondents from an area-probability sample, and 1,496 respondents from a list sample designed to over-sample wealthy households.[2] The response rate for the area-probability sample was about 70 percent; the response rate for the list sample varies strongly by stratum with progressively wealthier households having substantially lower response rates.[3]

The 1998 survey was conducted by the National Opinion Research Center at the University of Chicago between the months of June and December. Interviewers collected the data using a CAPI program. During their training, interviewers were taught both how to record respondents' data using the instrument, and how to record auxiliary information that might either further specify or qualify a response. In addition to verbatim and text information interviewers recorded during the interview, they were also required to complete a "debriefing" interview after each main interview. This information is structured around a set of questions that look for specific types of misreporting, and a final question asks for any additional information that might be useful in resolving later questions about the coherence of the interview. The comment and debriefing information was used primarily in the first stage of editing of the main data. Sometimes this information triggered very substantial rearrangements of the original data.

A final source of information used in this paper is a self-administered interview of the interviewers after their training. The questionnaire asked about aspects of their work experience as interviewers, their attitudes, and demographic information. The response rate was only 75 percent, but these cooperating interviewers accounted for 86 percent of the completed main interviews.

II. Indicators of Data Quality

Frequently, the information obtained from various interviewer comments and from more mechanical data review provokes large changes in the interpretation of the original data. There are a total of almost 48 thousand differences between the final version of the dataset and a version of that dataset excluding the two classes of edits. For comparison, there are about 2 million fields in total that do not contain a code signifying that the variable was an inapplicable item in either dataset. In over 15 thousand instances, editing "created" new missing data. In some ways, these figures overstate the true effect of editing. For example, an edit may determine that a particular response was unreliable; as a consequence, it becomes unknown which of several alternative sequences of subsequent questions the questioning should follow, and all of these subsequent questions in each branch are treated independently as missing data. In many other cases, edits are relatively simple rearrangements of the original data. Nonetheless, a very substantial fraction of the edits represent important changes to the data. Given the nature of the survey, a better indicator of the effects of editing may be the changes induced in dollar variables. Out of a total of about 173 thousand non-inapplicable data values for dollar variables, there were about 7,200 changes of any sort, and of those changes about 2,600 resulted in new missing values.

As shown in figures 1a and 1b, there is a broad dispersion in the number of edits across observations.[4] While 58 percent of all observations had at least some data change as a result of editing, only about 27 percent gained a new missing value, and the median number of such missing values was seven. In terms of edits to dollar

Figure 1a: Density of Number of All Edits and All Edits Yielding New Missing Values in Any Variables, Across Observations; Excluding Zeroes

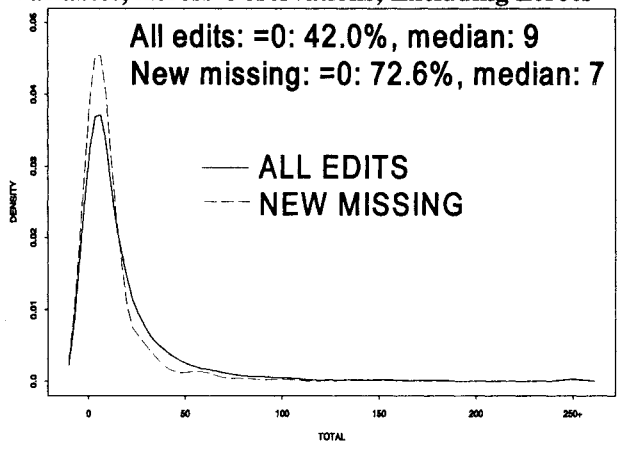


Figure 1b: Density of Number of All Edits and All Edits Yielding New Missing Values in Dollar Variables, Across Observations; Excluding Zeroes

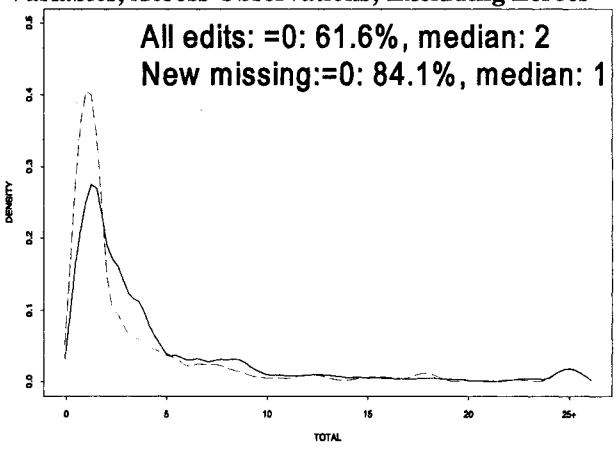


Figure 2a: Density of Number of All Edits and All Edits Yielding New Missing Values, Across Variables; Excluding Zeroes

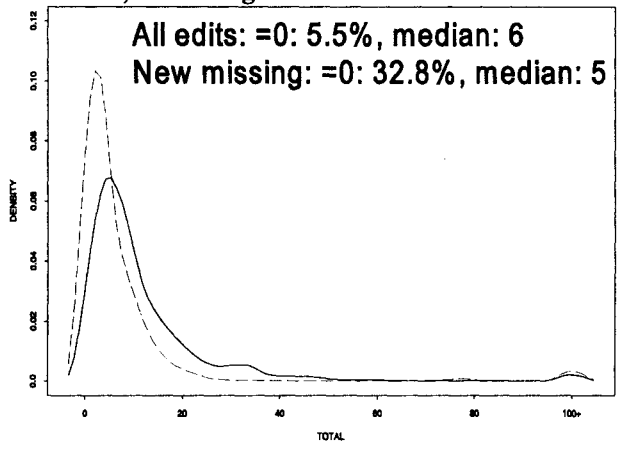


Figure 2b: Density of Number of All Edits and All Edits Yielding New Missing Values, Across Dollar Variables; Excluding Zeroes

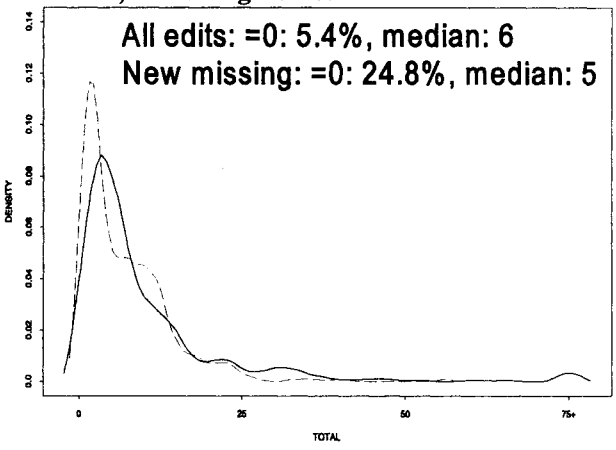


Figure 3a: Density of Mean Number of All Edits and All Edits Yielding New Missing Values in Any Variables, Across Interviewers; Excluding Zeroes

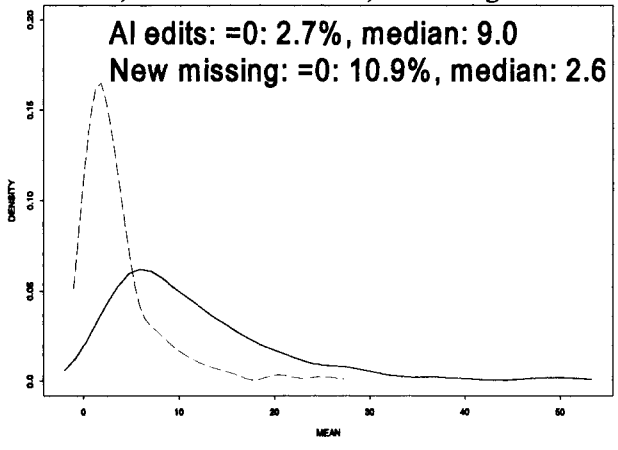
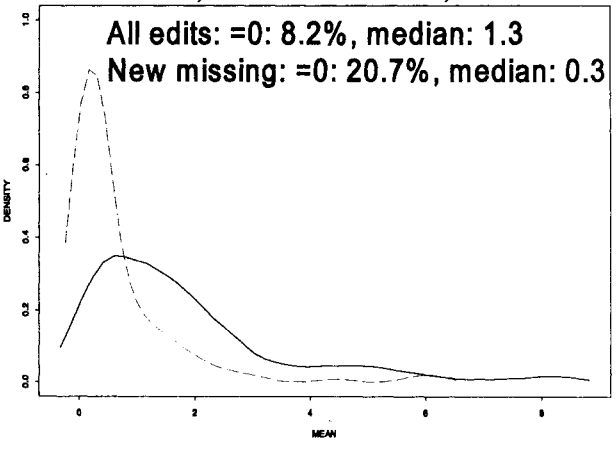


Figure 3b: Density of Mean Number of All Edits and All Edits Yielding New Missing Values in Dollar Variables, Across Interviewers; Excl. Zeroes



variables, the scope of editing was much more limited: almost 62 percent of cases had no edits, only about 14 percent gained a new missing value, and median number of new missing values for the latter group was one.

The great majority of variables—including dollar variables—had at least some changes made, and the proportion with at least one new missing values is also high (figures 2a and 2b). Although incidents of changes were broadly spread across variables, a much smaller number of variables accounted for a disproportionate share of the total edit changes.[5]

Data on the number of different types of edits across interviewers suggests a very wide variation in data quality problems across interviewers. However, this inference may be incorrect for two reasons. First, some interviewers completed far more cases than others, so even if all interviewers had the same level of problems, those with more completed cases would appear to have more problems. However, as shown in figures 3a and 3b, the same conclusions hold when looking at the distribution of the averages of the edits across each interviewer's cases. A second potential problem is that a large fraction of the edits was determined by information provided by the interviewers: about 87 percent of total edits and total new missing values were determined from interviewer comments. Because the responses of some SCF respondents are quite complicated, there may not always be an easy fit with the questionnaire. In such cases, interviewers are trained to record auxiliary information that might be useful in the proper recording of information later. Although such material is an important part of the auxiliary comment data, there is great heterogeneity in what is recorded. Unfortunately, there is no easy way to sort this information by its contents. Some interviewers record information that could have been fully incorporated directly into the questionnaire, while others report subtleties that allow the editors to salvage what would otherwise be missing data. Another approach is to look separately at the patterns in the "indirect" edits identified using logic rules and past experience. This approach removes the possibility of biasing the measures of data quality against interviewers who were simply doing a good job of documenting problems with difficult cases. However, the data show that there is still substantial variation over interviewer in the degree of edit problems even looking at this restricted class of edits.

III. Modeling Data Quality

If the data quality indicators above can be related to other measurable characteristics of interviewers and respondents, it may be possible to design interviewer training and additional respondent aids that may improve the quality of the data collected. This section develops some descriptive models to relate the quality measures developed earlier in the paper to other information about respondents, respondents' neighborhoods, and characteristics of interviewers.

Problems may arise during an interview from the respondent's reaction to the interviewer or the instrument, or from the interviewer's reaction to the respondent or the instrument. There are several dimensions that seem plausible determinants in explaining problems. For example, respondents who have relatively complicated situations might be more likely to have trouble fitting their responses into the desired format. Those who are less trusting of the interview process might tend to limit the amount of information revealed. Respondents who felt more pressed for time might not devote as careful attention to the questions as others. Respondents who are more sophisticated in their understanding of questions and their ability to express themselves might be more likely to give the analytically desired responses. One would hope that experienced interviewers are more comfortable in probing respondents to answer questions and fit responses into the analytically desired framework, and that they are better able to follow the standard protocol for administering the interview. It also seems reasonable to think that interviewers who have personality traits that make them outgoing and persistent should have better success in controlling the interview. Unfortunately, the data available to test these relationships are often are indirect.

There are two particularly important additional

Figure 4: Two-Stage Model of Data Quality Indicators

1. $Pr(E_i = e) = Exp(-\lambda_i)\ (\lambda_i)^e / e!$
2. $Ln\ (\lambda_i / Q_i) = X_i\ \beta + D_{j(i)}\varphi$
3. $Expected\ (E_i) = Q_i\ Exp(\ X_i\ \beta + D_{j(i)}\varphi\ )$
4. $D_{j(i)}\ \phi = F_{j(i)}\gamma + \eta_{j(i)}$

Where:

E_i is the number of errors/edits for observation i

X_i is a vector of characteristics of observation i

β is a vector of parameters compatible with X

$D_{j(i)}$ is a vector of dummy variables representing each interviewer j who interviewed observation i

φ is a vector of parameters compatible with D

Q_i is the number of questions of the sort modeled to which observation i was exposed

$F_{j(i)}$ is a vector of characteristics of interviewer j who interviewed observation i

γ is a vector or parameters compatible with F

ϕ is the estimated value of φ

$\eta_{j(i)}$ is a error term reflecting a combination of estimation error in φ & modeling error in equation 4

obstacles to straightforward modeling. First, the assignment of cases to interviewers was nonrandom. Generally, interviewers are given an initial local assignment, and where the number of interviewers is great enough, there is an attempt to "match" interviewers to the cases where their supervisors think they will be most successful in getting an agreement to do the interview. Interviewers who are unusually good and who manage to complete their local assignments early in the field period are often offered the chance to travel to other areas with large numbers of pending cases. Unfortunately, it would be virtually impossible to control for all dimensions of selection that might be operating in this process.

A statistical problem is that errors in modeling the effects of interviewers on data quality are not independent across observations, because most interviewers completed more than one case. Fortunately, this problem is more amenable to statistical treatment. In the models that follow, it is assumed that interviewer contributions to the model can be captured as fixed effects, which are then modeled in a second stage in terms of interview characteristics.

The selection of appropriate dependent variables for the analysis is not obvious. Any choice other than modeling quality indicators at the level of individual variables allowing for correlation across variables, requires aggregation. Aggregation implies something about the relative importance of the component data items. While recognizing the inherent arbitrary nature of any aggregation, ease of interpretation argues for examining a small number of key dimensions. Because of the importance of dollar amounts in the SCF, the modeling of editing focuses only on such variables. The variables modeled are "EDOLL," which counts the number of changes in dollar variables as the result of editing, and "NMDOLL," which counts the number of new missing values generated by editing.

The model takes the two-stage form described in figure 4. Equation 1 is the mathematical form of the poisson model which is the first stage. The log of the poisson parameter λ is specified as a linear function of a vector of respondent characteristics, and a vector of dummy variables constructed for each interviewer, scaled by the level of exposure to the relevant type of questions (equation 2). As shown in equation 3, the expected value of the number of errors or edits takes the form of a rate multiplied by the number of questions of the relevant type to which the subject was exposed. This stage was estimated by running the maximum likelihood poisson procedure in Stata on a subset of 4,168 observations for which the associated interviewer completed at least four cases (there were 167 such interviewers).[6] The second stage, given by equation 4, regresses the estimated individual interviewer effects from the first stage on measured interviewer characteristics. This stage was estimated using the robust regression routine in Stata.[7] Results from the first- and second-stage estimations are given in table 1.

The data provide strong support for the overall importance of interviewers effects in each model: a likelihood ratio test of H_0: $\{\varphi_i=0, \forall\, i\}$ is rejected at less than the one percent level. The first-stage results also support the claim that respondent characteristics are important in explaining the aspects of data quality modeled, though the interpretation of the results is not always straightforward. The data suggest that payment of incentives to respondents to participate in the survey was negatively associated with the problem rates, as one might reasonably expect as part of the "bargain." There is an indication that the rate of problems "scales up" with household size, probably as a consequence of relatively greater complexity. Older respondents appear more likely to have higher rates of new missing data in editing. Curiously, more educated respondents had higher error rates in both models. As one might expect, there are also indications that respondents who were less interested in the interview, had a less good understanding of the questions, or did not express themselves clearly had more problems. Comfortingly, respondents who used records has a lower rate of problems. Respondents who showed signs to the interviewer of being suspicious after the interview were more likely to have had problems in their data. The rates of dollar edits and new missing values are positively associated with income, assets, and debts show a mixed pattern. Other work with the data suggests that these factors are picking up two dimensions: on the one hand, respondents with higher levels of income and wealth often tend to be less cooperative with the survey process, and on the other hand, such people have greater levels of financial sophistication that would make it more likely that they would understand the intent of the survey questions. Contrary to expectations, there is some suggestion that average commuting time in the census tract is associated with *lower* error rates for NMDOLL. Interview length is positively associated with higher level of EDOLL, possibly an effect of the actual length of time required to accommodate confusing responses through interviewer comments, or of some other aspect of a higher level of complexity among such cases.

In earlier explorations using OLS to estimate the second stage of the model, there was very little consistent sensible variation despite the strong overall significance of interviewer effects. Robust regression reveals somewhat more structure. Interviewers with longer years of experience, and those with experience on the SCF tend to have lower levels of EDOLL and NMDOLL. Such interviewers appear to be better at capturing data correctly

Table 1: First-Stage Poisson Regression of NDOLL and NMDOLL, and Second-Stage Robust Regression

Stage 1	EDOLL		NMDOLL	
	Est.	S.E.	Est.	S.E.
CONST	-8.595#	0.453	-11.910#	0.962
RFEE	-0.007#	0.001	-0.006#	0.002
AGER	0.002	0.001	0.012#	0.002
NWHISP	-0.004	0.150	1.348*	0.584
RHEALTH	0.023	0.017	0.014	0.031
MARRIED	-0.211#	0.034	0.092	0.061
HHSIZE	0.063#	0.010	0.046*	0.020
REDN	0.023#	0.006	0.024*	0.011
INTEREST	0.175#	0.018	0.268#	0.029
EXPRESS	0.095#	0.032	0.046	0.058
UNDERSTD	0.104#	0.031	0.078	0.056
RECORD	0.000	0.027	-0.293#	0.050
SUSPA	0.118#	0.034	0.524#	0.053
RWORK	-0.116#	0.031	-0.155#	0.053
DHOWNER	-0.432#	0.041	-0.256#	0.076
TIMEAREA	0.049#	0.012	-0.024	0.020
NORMY	0.110#	0.014	0.194#	0.025
ASSETS	-0.040#	0.008	0.070#	0.017
DEBT	-0.019#	0.003	-0.036#	0.004
COMMTIME	0.012	0.027	-0.090*	0.043
MHORIZ	-0.041#	0.010	-0.048#	0.018
TIMEIW	0.369#	0.037	0.087	0.062
REG1	0.469#	0.129	0.860#	0.226
REG2	0.510#	0.082	0.828#	0.150
REG3	0.300#	0.095	0.720#	0.170
SRPSU	0.010	0.050	-0.137+	0.082
MSA	-0.247#	0.074	0.147	0.125
Coefficients for interviewer-specific dummy variables and list sample strata are not shown				
LR{H_0 φ_i=0 $\forall$ i}	2874#		2402#	
N	4168		4168	
Stage 2	*EDOLL*		*NMDOLL*	
	Est.	S.E.	Est.	S.E.
CONST	-1.623+	1.158	-5.836#	2.228
LIWEREXP	-0.107#	0.032	-0.127*	0.061
DISCF	-0.132	0.159	-0.627*	0.306
LIAGE	0.576#	0.032	1.342#	0.446
ICOLLEGE	0.222+	0.141	0.038	0.271
ITYPEF	0.165	0.151	0.500*	0.291
ITYPES	-0.239+	0.154	0.076	0.296
IRESEAR	0.206*	0.114	0.300+	0.219
INEIGHB	-0.083	0.092	-0.140	0.177
DFIGURE	-0.148*	0.076	0.066	0.147
DIGET	-0.044	0.054	-0.070	0.104
DIRRESP	-0.013	0.058	0.072	0.112
IOUTGO	0.144*	0.078	0.119	0.151
DCONOUT	-0.146#	0.053	-0.125	0.103
LCOUNT	0.101	0.094	0.091	0.181
N	120	120		

#=p-value<=1%, *=pvalue<=5%, +=p-value<=10%

RFEE: Amount of any fee paid to respondent
RAGE: Age of the respondent.
NWHISP: =1 if the respondent nonwhite/Hispanic.
RHEALTH: Health: 1=excellent,...,4=poor.
MARRIED: =1 if married/living with a partner.
HHSIZE: Number of people in household.
REDN: Years of education.
INTEREST: R's interest in interview: 1=very high,...,5=very low.
EXPRESS: R's self-expression: 1=excellent,...,4=poor.
UNDERSTD: R's understanding of questions: 1=excellent,...,4=poor.
RECORD: =1 if R used records.
SUSPA: =1 if R suspicious at end of interview.
RWORK: =1 if R working.
DHOWNER: =1 if homeowner.
TIMEAREA: Log(years in area)..
NORMY: Log("normal" income).
HIINC: =1 income unusually high.
ASSETS: Log(assets).
DEBT: Log(debts).
MHORIZ: Planning horizon: 1=next few months,..., 5=longer than 10 years.
COMMTIME: Log(avg. commuting time in min. for people living in the Census tract.
TIMEIW: Log(interview length in sec.).
DSTR1-1: =1 if list sample stratum 1-7.
REG1: =1 if case in NE region.
REG2:=1 if case in NC region.
REG3: =1 if case in W region.
SRPSU: =1 if case in self-representing PSU.
MSA: =1 if case in any type of MSA.
LIWEREXP: Log(years as interviewer)..
DISCF: =1 if worked on the SCF before 1998.
LIAGE: Log(age of interviewer).
ICOLLEGE: =1 if interviewer had some college education.
ITYPEF: =1 if interviewer was fast typist.
ITYPES: =1 if interviewer was slow touch-typist.
IRESEAR: "I like being a part of a research project.": 1=strongly agree,...,5=strongly disagree.
INEIGHB: "I enjoy the challenge of visiting unfamiliar neighborhoods.": 1=strongly agree,...,5=strongly disagree.
DFIGURE: "Most of the time I can figure out what a respondent's real objections are.": 1=strongly agree,...,5=strongly disagree.
DIGET: "It's better to persuade a reluctant presondent to participate than to accept a refusal, even when you feel they won't give very accurate answers.": 1=strongly agree,...,5=strongly disagree.
DIRRESP: "We should respect respondents' rights to refuse by not pushing when they say 'no'.": 1=strongly agree,...,5=strongly disagree.
IOUTGO: "I am generally very outgoing.": 1=strongly agree,...,5=strongly disagree.
ICONOUT: Iwer confident in confidentiality protections: 1=strongly disagree, ..., 5=strongly agree
LCOUNT: Log(number of cases completed by the interviewer).

within the instrument and recording information that is useful to support unusual situations. Older interviewers are mode likely to have dollar edits and new missing values. There is some indication that interviewers who are proficient typists are more likely to have problems, perhaps because it is relatively easier for them to record responses verbatim than to probe the respondent. The effects of the self-reported attitudes are generally weak and occasionally inconsistent. One interesting such variable is the confidence that interviewers have that NORC would protect the identifying information in the survey. Interviewers who expressed higher levels of confidence had significantly lower rates of edits.

IV. Conclusions and Future Research

This paper has examined data quality in terms of editing in the 1998 SCF. As a byproduct, the analysis has led to the identification of question sequences that need particular attention in the revisions of the instrument and in the design of future interviewer training. In some cases, the results suggest that it may also be important to develop other means to orient respondents to the desired data reporting framework.

The results of the modeling exercise provide at least some evidence of systematic variations in coding and editing problems over different types of respondents and interviewers. However, the data do not provide as much structural insight as one would like. For the future, two additional sources of information would be very helpful in identifying the sources of data quality problems. First, it would be helpful to have a set of objective measures of interviewers' abilities in addition to the self-reported information available now. Second, it would be useful to have information directly from the survey respondents on their impressions about the interview. Perhaps to encourage frankness, a short series of questions about the interviewer and the interview could be left with the respondent after the interview for the respondent to complete and mail in independently.

Another important facet of data quality that is not explored in this paper is the effect of data changes on the ultimate estimates using the data. The presumption has always been that the SCF data should be as correct as possible, without going to the point that highly unusual, but legitimate, variations in the data are suppressed. However, so far there is no broad quantitative measure of the benefits of such review. One possibility, to be developed in a subsequent paper, is to compare a set of analyses using the fully reviewed and imputed data to the same analyses using an imputed version of the edited data. Given the available software for the SCF, it should be straightforward to construct the additional dataset. It is highly likely that the data review greatly damps spurious swings for some variables in tail-sensitive measures, such as the mean and concentration ratios, but it is not clear how much other measures typically viewed as being more robust, such as the median, are affected.

Bibliography

Kennickell, Arthur B. [1998] "List Sample Design for the 1998 Survey of Consumer Finances," working paper, Federal Reserve Board.

_____ [1999] "Revisions to the SCF Weighting Methodology: Accounting for Race/Ethnicity and Homeownership," working paper, Board of Governors of the Federal Reserve System.

_____, Martha Starr-McCluer, and Annika E. Sundén [1997] "Changes in Family Finances in the U.S." Federal Reserve Bulletin, January, pp. 1-24.

Endnotes

1. The author is grateful to Annelise Li and Amber Lynn Lytle for research assistance and to Gerhard Fries, Martha Starr-McCluer, Annika Sundén, and Brian Surette for their work on the data underlying this analysis, and to the staff and interviewers at NORC who collected the data. Particular thanks go to the SCF respondents who so generously gave of their time for the interviews. This paper is the responsibility of the author alone and the views expressed herein do not necessarily reflect the opinions of the Board of Governors of the Federal Reserve System.

2. See Kennickell [1998] for a more extended discussion of the sample.

3. See Kennickell [1999] for additional information on response rates.

4. In these density estimates figures and the ones that follow, the estimates exclude zero values. In each case, the percent of zero values is shown on the figure.

5. An appendix to the longer version of this paper enumerates the variables with particularly large numbers of problems.

6. The data used for estimation were the singly-imputed first imputation iteration of the 1998 SCF.

7. The second-stage model is heteroskedastic because different interviewers completed different numbers of cases. To weight implicitly, the robust regression was run on the 3,416 observations corresponding to the set of interviewers who completed the interviewer questionnaire and who completed at least four cases, and the standard errors were adjusted to reflect the artificial inflation of the sample size.