EDITING THE 2001 SURVEY OF CONSUMER FINANCES[1]

Ryan Bledsoe, Federal Reserve Board; Gerhard Fries, Federal Reserve Board
Gerhard Fries, FRB, Mail Stop 153, Washington, DC 20551; gfries@frb.gov

Key Words: Data quality, editing, coding

The primary goal of the Survey of Consumer Finances (SCF) is to provide household-level data that accurately reflect the balance sheets of households in the United States. For a variety of reasons, the information recorded during a survey interview may deviate from what is desired: respondents may misunderstand questions, interviewers may record answers incorrectly, complex situations sometimes fit awkwardly into the structure of the survey interview, etc. In an effort to improve the quality of the data, every SCF interview is subjected to intensive review using comments and other data recorded during the interview, information provided by interviewers in a debriefing report they are required to make for every case, and reports generated mechanically from the raw data. Where potential problems are detected, the data are reviewed more closely to build a case for altering the original data, if necessary. Changes are made only where there is a clear preponderance of supporting information. A strong effort is made to develop simple rules, but when simple rules fail, decisions are guided by a review of "case law" developed over the history of the survey in an effort to maintain conceptual consistency. This paper provides an overview of the editing process that occurred for the 2001 SCF.

The first section provides a brief summary of the SCF, covering the sample design, data collected, and issues involving nonresponse and variance estimates. The second section discusses the editing procedures followed by SCF staff for the 2001 survey. The third section summarizes the editing process.

I. Background on the SCF

The SCF is a triennial household survey sponsored by the Federal Reserve Board with cooperation from the Statistics of Income Division (SOI) of the Internal Revenue Service. Data are collected on household finances, income, assets, debts, employment, demographics, and attitudes. Interviews for the 2001 SCF were conducted via Computer-Assisted Personal Interviewing (CAPI) by the National Opinion Research Center at the University of Chicago (NORC) between June and December of 2001. The median length interview required approximately 79 minutes, although complicated cases took substantially longer. Most interviews were obtained in-person, but 35 percent were conducted by telephone, generally as an accommodation to respondents' preferences. Data are collected on items that are not always widely distributed (e.g. non-corporate businesses or tax-exempt bonds). In order to provide adequate coverage of such variables and to provide good coverage of broadly distributed characteristics (e.g. home ownership) in the population, the SCF combines two techniques for random sampling. The sample is selected from a dual frame that is composed of a standard, multistage Area-Probability (AP) sample and a list frame (see Kennickell and McManus [1993] for details on the strengths and limitations of the sample design). The list frame is based on statistical records derived from tax returns. The list sample is stratified on an estimated "wealth index", with higher values selected at a higher sampling rate. These records are made available for this purpose under strict confidentiality rules. The list sample is designed to oversample relatively wealthy families but excludes people mentioned by *Forbes* magazine as the 400 wealthiest in the U.S.

Of the 4,449 completed interviews in the 2001 survey, 2,917 families came from the AP sample and 1,532 from the list sample. The response rate for the AP sample was about 68 percent. The overall response rate for the list sample was about 30 percent, where the rate was only 10 percent for the part of the list sample containing the wealthiest families.

Both unit and item nonresponse are important issues for the SCF. Weighting adjustments compensate for nonrespondent households. The adjustments include post-stratification to known, external control totals for age, location, and home ownership. For the list sample, frame data on financial income and the wealth index are also used (see Kennickell and Woodburn [1996]). Multiple imputation deals with missing data (see Kennickell [1998]).

Both imputation error and sampling error are measurable for the SCF. Estimates of the variance due to imputation are computed using five imputation replicates ("implicates"). Estimates of the variance due to sampling are computed using replication methods where samples are drawn from actual respondent records in such a way that the important dimensions of the original sample design are incorporated. These estimates can then be combined to yield standard errors for analysis (see Kennickell [1999]).

II. Editing the SCF

In an effort to improve data quality, two techniques were employed to thoroughly review the data. First, comments and other data recorded during the interview, as well as information provided by interviewers in a debriefing report, were examined by SCF staff. The

combination of these types of data is generically referred to as auxiliary information. Second, SCF staff reviewed outcomes from a set of logical and institutional consistency checks. These checks are based on *a priori* logical and institutional requirements. Many times, these checks were developed to ensure that reoccurring patterns of errors were handled consistently across cases. Additionally, a small percentage of edits resulted from a review of outlier and influence plots.

Each case was reviewed by SCF staff in order to detect possible inconsistencies in the data. When inconsistencies warranted a rearrangement of the data, an edit was created to adjust the errant data. Edits account for 52,201 alterations to the data (see Kennickell [2002]). This figure includes 14,647 new missing values. Nearly 32 percent of cases had at least one missing value created due to editing. In 1998, edits accounted for 47,787 alterations to the data (see Kennickell [1999]). This figure includes 15,428 new missing values, where 27 percent of the cases had at least one new missing value as a result.

A. Auxiliary Information Edits

Interviewers were encouraged to enter comments at almost any point during the interview, and specifically where they felt ambiguity existed in the data. They were not constrained in the content or the amount of information they included in these comments. Some auxiliary information was recorded to qualify responses to categorical variables when a pre-specified code frame was inadequate (e.g. an unusual frequency of mortgage payment). When sufficient information was recorded, these verbatim responses were coded after the completion of the survey interview by either NORC or SCF staff. Sometimes interviewers used this facility in the instrument to record broader information of the sort that would be expected in an interviewer comment field.

At times, the CAPI instrument asked the interviewer to enter a verbatim response before they could proceed with the survey interview. For example, the final question for each survey interview was a verbatim question that asked, "Is there anything you would like to add to any subject we've discussed?" The interviewer was required to enter some text for this question in order to complete the interview. Required verbatim questions also appear in the survey to qualify other types of assets, other forms of income, mutual fund holdings, other bond holdings, and life insurance holdings, as well as to identify double-counted assets, income, and liabilities. Finally, interviewers were required to complete a debriefing report before a survey interview was considered complete. The questions asked of the interviewers probed for information that would be useful to SCF staff during editing. This set of auxiliary information was reviewed during the editing process.

Each case's set of auxiliary information was combined to create a working text file. A working text file's header includes: the case's ID number, the case's wave number, the interviewer's ID number, the respondent's age, and the length of the interview in minutes. Information recorded in the interviewer's debriefing report followed the header in its own section, while comments and other data recorded during the interview were displayed in the order they were entered following the debriefing report output with a label identifying the associated question where the comment or other verbatim was recorded. The end of each working text file includes the shell of a SAS program where edits were entered. These files were included as a part of a layer program to install the edits and their logical consequences. All the text material was treated as a series of SAS comments.

Data were received from NORC in waves, with each wave of data including approximately two weeks of interviews. Thirteen waves of data were processed during the 2001 cycle. Editors reviewed one wave of working text files at a time. For each wave of working text files, a corresponding batch of "caseview" files was generated. Each caseview file consisted of the recorded answer to every survey question along with the variable name and an identifying label. The caseview and working text files were the primary tools used to edit the data.

It is useful, conceptually, to separate the discussion of auxiliary information edits into the following three parts discussed below: systematic edits, categorical edits, and irregular edits.

"Systematic" Edits

Some verbatim questions were only asked when certain conditions were met. Edits generated from this set of verbatim responses are referred to as "systematic". Special care was given when reviewing "systematic" auxiliary information since resolution was required every time a verbatim response of this type appeared in a working text file. A subset of these verbatim questions prompted editors to alter the data every time they were answered. These verbatims correspond to the double-counting of future pension benefits or the double-counting of assets and liabilities held by members of the NPEU. (people who "usually live" at the respondent's primary residence but are "financially independent" of the household).

The question "Which account or pension?" was only asked when the respondent stated the future pension they reported had also been recorded earlier in the survey interview. This verbatim question asked the respondent to identify where the future pension was double-counted. Given the response to this verbatim and a caseview file, the editor resolved "Is this pension part of an IRA, Keogh, or other pension plan you already told me about?" from 1 (Yes, IRA/Keogh) or 2 (Yes, pension) to 3 (Pension(s)

remain(s) after removing plans reported earlier), 4 (Contrary to respondent's answer, unable to identify any such plans reported earlier), or 0 (Inappropriate). For example, when the response to "Which account or pension?" stated "IRA", the editor reviewed the caseview file for an Individual Retirement Account (IRA) that matched the future pension. If the editor determined the IRA was a double-counted future pension, the IRA was removed and the answer to "Is this pension part of an IRA, KEOGH, or other pension plan you already told me about?" was set to "pension remains after removing plans reported earlier". If the future pension was actually an IRA, the future pension was removed, implicitly setting this question to inappropriate. Finally, the answer to this question was set to "Contrary to respondent answers, unable to identify any such plans reported earlier" if no reported data matched the future pension. In such instances it is assumed that the respondent may have mentioned the pension earlier in passing, but the interviewer correctly did not record it.

NPEU assets, liabilities, and income received in 2000 are recorded separately in a section toward the end of the survey interview. Any time the respondent reported NPEU assets, liabilities, or income, a question asked if the item had also been recorded earlier in the survey interview. If this question was answered "yes" it was followed up with a verbatim question that asked where the item was recorded earlier. The corresponding variable was always resolved from 1 (yes) to either 3 (yes, amount edited out earlier) or 4 (yes, but no apparent match in the data: nothing changed).

The above set of "systematic" verbatim responses called for altering the data every time they appeared in a working text file. The rest did not necessitate such action. These verbatim questions were asked in order to clarify types of mutual fund, other bonds, and life insurance holdings. They were asked when certain conditions were met and resolved each time they appeared in a working text file. Sometimes the resolution was simply to impute asset type (i.e. no edit was required).

The verbatim question "Please explain type of mutual funds" was asked when the respondent reported mutual fund holdings but was unable or unwilling to report a mutual fund type. If the respondent answered "yes" to "Do you have any mutual funds?"a series of questions was asked to determine the composition of the household's mutual fund holdings (e.g. the household owns stock funds and tax free bond funds). If the respondent answered "no", "don't know", or some combination of these two answers to each of these questions, the respondent was asked to report the total value of all mutual funds holdings, while a verbatim question asked the respondent to clarify the type of mutual funds held by the household. Ideally, the response to the verbatim question provided the editor with enough information to move the recorded value of the household's mutual fund holdings to one of the types in the mutual fund grid (e.g. stock funds). For example, when this verbatim response stated, "Mid-Cap Growth Fund" the editor set the variable corresponding to "Do you have stock funds?" to "yes" and moved the dollar value of mutual fund holdings to the variable corresponding to "Total market value for stock funds". When the verbatim response read "don't know" the editor simply moved on, allowing mutual fund type to be imputed. The review of other verbatim responses in this group was similar. Unlike the first set of "systematic" verbatim responses, it was appropriate at times for no action to be taken when one of these verbatim responses appeared in a working text file.

"Categorical" Edits

Frequently, verbatim responses were recorded as answers to categorical questions (e.g. unusual frequency of mortgage payment). For these questions, only the most common responses along with the response "other" were displayed on the computer screen during the interview. When interviewers fielded a response that was not a common response, they simply chose the response "other" to enter a verbatim response. Most of the time these verbatim responses were resolved to existing codes by either NORC or SCF staff. In some instances, however, these verbatim responses brought to light errors in the data. These errors were often the result of a misreported asset, liability, or income type.

For example, a question about the specific type of savings account the respondent held revealed that a certificate of deposit had been misclassified as a savings account. This type of data error occurred most frequently within the "other" assets income sequences. These questions were only asked after all other questions concerning assets and income had been reported, respectively. Consequently, specific types of assets (e.g. mutual funds) or income (e.g wage and salary) were sometimes erroneously reported as an other asset or other type of income. Any time a respondent reported having other assets the interviewer entered a verbatim response for the type of asset. Similarly, types of other income were recorded with a verbatim response. Therefore, data errors were always documented. These mistakes were remedied by moving the errant data to the appropriate asset or income section. When the verbatim response for other asset type read, "tax free bond fund," for example, an edit moved the reported dollar value for this asset to the "tax free bond fund" section of the mutual fund grid. Similarly, when the verbatim response for other income type read "part-time job", the dollar value reported as other income was moved to wage and salary income.

"Irregular" Edits

The remaining auxiliary information is referred to

as "irregular". There was nothing systematic about this set of comments. These comments ranged in context from stating the interview was successful for one case to identifying ambiguities throughout the entire interview for another. "Irregular edits" were the most difficult of the edits generated. No steadfast set of rules existed for resolving these comments. Many times, the editor was required to consider several comments in conjunction to justify altering the data.

At times, irregular edits were straightforward. Due to the complexity of the interaction between the respondent and interviewer, assets and/or liabilities were sometimes missed or double-counted. Comments occasionally appeared in the debriefing report (or toward the end of the survey interview), stating something analogous to, "the respondent has a $20,000 money market account that was not recorded". For this comment, the editor checked to see if a money market account was recorded. If not, a $20,000 money market account was created. Specific assets and liabilities were also at times double-counted. For example, respondents sometimes recorded the same asset as both an IRA and a 401k. Usually, a comment identified whether the asset was actually an IRA or a 401k. If the editor determined the 401k was indeed a double-counted IRA, an edit removed the IRA. Similarly, if the editor determined that the asset was an IRA, the 401k was removed.

Unfortunately, most irregular edits were not this straightforward. Frequently, irregular auxiliary information conveyed data inconsistencies that could not be resolved without supporting information. "Case law" was developed during the 2001 editing process to ensure editors handled sets of comments as consistently as possible across cases. The project director served as the chief "judge" in the creation of these rules by drawing on broad generic concepts, earlier unclassified decisions, and knowledge of the ultimate uses of the data. SCF staff members continue to document and update "case law" to ensure sets of comments are treated uniformly in subsequent rounds of the survey.

"Case law", for a specific comment, outlines the supporting information required for an edit to be generated. For example, "pension" was a common verbatim response to "types of direct deposits". Upon reviewing this verbatim response, an editor checked the data for a recorded pension payment and for reported pension income. If the respondent reported receiving both a pension payment and pension income, the editor simply left the case alone since the verbatim response was consistent with the data. Under certain conditions, however, the editor created a pension payment. Consider a case where: the verbatim response for types of direct deposits read, "pension", no pension payment was recorded, social security was recorded as a type of direct deposit, social security payments were recorded, but reported pension income exceeded the amount of reported social security income. The verbatim response and the data suggested that a pension payment was missing. This set of conditions constituted the "case law" necessary to lead editors to create a pension payment.

B. "Indirect" Edits

Most of the remaining edits were the result of a review of logical and institutional inconsistencies. SCF staff developed a set of a priori logical and institutional software checks. Primarily, these checks attempt to capture instances when the interviewer mistyped a response or the respondent did not understand the question. The development of these checks was dynamic. As patterns of inconsistencies appeared in the 2001 data, checks were created to ensure that similar patterns of errors were treated consistently across cases. SCF staff reviewed the outcomes of each of these checks. When a preponderance of evidence suggested that data were erroneous, an edit corrected the data. Since numerous inefficiencies are introduced by editing the data after the completion of the survey interview, mere inconsistency was rarely sufficient to warrant altering the data. SCF staff also reviewed a set of outlier and influence data plots. At times, the plots highlighted errant data missed in earlier stages of editing. For these instances, edits were generated to correct the data.

Logical Checks (Examples)

A set of survey variables collects data regarding time frame information (e.g. the year the household moved into their primary residence or the year that the respondent expects to receive a future pension benefit). Responses to these questions via programmable logical checks were compared to the survey year (2001) in order to highlight logical inconsistencies. For example, a warning message from a logical check was produced when the data indicated that the household moved into their primary residence in some year after the survey. Similarly, a warning from a logical check was generated each time a respondent reported that they were expecting to start receiving a future pension payment in some year prior to the survey year. The CAPI program prevented interviewers from entering obviously inconsistent data, in general, but the complexity of the instrument made complete enforcement of logical consistency infeasible. SCF staff reviewed the output from each of these checks to determine if the inconsistency warranted altering the data. Editors took every measure possible to code errant data to specific years. When other applicable information provided no guidance for resolving the inconsistency, the errant data were set to missing.

The SCF also collects household liability data. The liabilities covered include mortgages, lines of credit, education loans, other types of installment loans, margin accounts, pension loans, credit cards, and other types of

debts. A set of logical checks compared the amount borrowed to the amount owed for each reported loan. A warning message was produced each time the amount owed on a loan exceeded the amount borrowed. Most of these inconsistencies involved the mortgage on the household's primary residence. The most common error occurred when the respondent reported the amount of home equity rolled over as the total amount borrowed. If possible, the editor resolved the amount borrowed to a specific dollar value. If potentially accurate, the data were left alone. When the data did not provide direction concerning the actual loan amount and the error was too egregious to maintain, the amount borrowed was set to missing.

These are a few examples of logical checks used by the SCF staff during the 2001 editing process. Some of the other checks reviewed include: cases when the respondent has two jobs and reports his/her work status as part-time, anytime a respondent who is less than thirty years old reports expecting to receive a future pension benefit, and anytime a reported car loan was taken out prior to the purchase of the automobile.

Institutional Checks (Examples)

Guidelines set forth by the Social Security Administration, the Internal Revenue Service, state law, credit card companies, automobile dealers, and other institutions were reviewed to develop *a priori* institutional checks. Warning messages from these checks were reviewed in a similar manner. For example, a warning message from an institutional check was generated if a $6,000 monthly payment was reported as the respondent's Social Security benefit (a check was generated for any payment that exceeded $25,000 per year). Generally, the editor assumed the value was misreported and should be $600 per month or $6,000 per year. The household's pension income was reviewed to determine which payment the data supported. If payment amount remained ambiguous after reviewing all applicable information, the editor set this payment to missing.

The software generated another warning message if reported property taxes exceeded five percent of the value of the house. This inconsistency usually occurred when a yearly tax payment was recorded as a monthly payment. This error was remedied by setting the payment frequency to "yearly". When the data provided no clear resolution the tax payment was set to missing.

These are a few examples of institutional checks used by the SCF staff during the 2001 editing process. Other checks include: any instance where income exceeded some minimum threshold and the respondent reported not filing taxes, anytime the credit limit or interest rate on a credit card exceeded some expected maximum level, and any instance where the number of years a car is leased exceeds some expected maximum level, to name a few.

Influence and Outlier Plot Edits

Much of the data editing and initial imputation processing is done in parallel and specific types of plots are used to check for possible outlier values. If extreme outliers are found, they may indicate a missed edit or an incorrect edit. One type of such plot is a scatter plot of the survey variable versus an indication of sampling stratum. List cases can be discerned from AP cases and imputed data from reported or edited data. Almost every survey variable is plotted in this way. Another type of plot shows the influence of each case for a particular variable. These plots use a constructed analysis weight and can reveal whether a variable for a given observation is contributing too much to the overall weighted total for that variable. Graphical analysis is quite effective in the editing process to help ensure high-quality data, and its use can not be stressed enough.

III. Summary

The primary goal of the Survey of Consumer Finances is to provide household-level data that accurately reflect the balance sheets of households in the United States. In an effort to ensure accurate data, each case is subjected to an intensive review. This includes a review of auxiliary information recorded during the interview as well as *a priori* logical and institutional checks developed during the 2001 cycle and earlier survey years. These reviews are necessary since, for a variety of reasons, the information recorded during a survey interview may deviate from what is desired: respondents may misunderstand questions, interviewers may record answers incorrectly, complex situations sometimes fit awkwardly into the structure of the survey, etc. When a preponderance of supporting information suggested that the data were erroneous, changes were made to correct the data. Simple rules guided most editing tasks. However, for complex sets of reoccurring comments, "case law" was developed in an effort to maintain conceptual consistency across cases. This "case law" has been documented to help guide editing in the 2004 SCF.

References

Fries, G., and R.L. Woodburn [1995] "Using Graphical Analyses to Improve all Aspects of the Survey of Consumer Finances," *Proceedings on the Section of Survey Research Methods*, 1995 Annual Meeting of the American Statistical Association, Orlando, FL.

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

Kennickell, A.B. [2001] "http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html"

Kennickell, A.B. [1999] "Measuring Data Quality in the 1998 Survey of Consumer Finances ," *Proceedings of the Section on Survey Research Methods*, 1999 Annual Meetings of the American Statistical Association, Baltimore, MD.

Kennickell, A.B. [1998] "Multiple Imputation in the Survey of Consumer Finances," *Proceedings of the Section of Survey Research Methods*, 1998 Annual Meetings of the American Statistical Association, Dallas, TX.

Kennickell, A.B., and D.A. McManus [1993] "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section of Survey Research Methods*, 1993 Annual Meeting of the American Statistical Association, San Francisco, CA.

Kennickell A.B., D.A. McManus, and R.L. Woodburn [1996] "Weighting Design for the 1992 Survey of Consumer Finances," Federal Reserve Board Working Paper.