Dirty and Unknown: Statistical Editing and Imputation in the SCF

Arthur Kennickell[1]
[1]Assistant Director, Research and Statistics, Federal Reserve Board, Mail Stop 153, Washington, DC 20551; Arthur.Kennickell@frb.gov

Abstract
Prevention of errors in surveys must always be the highest ideal, but in such a complex process as a survey there are limits on what is achievable, because of cost, the absence of strong instruments for control or the emergence of unforeseen outcomes. Thus, effort must be devoted to identifying errors, remediating them, and designing better means of prevention or limitation where that is possible. Editing is typically a key instrument of identification and remediation. However, editing can consume very substantial resources and because the outcome is unlikely to be perfect, the very act itself introduces additional risks to data quality. For these reasons, it has been argued (e.g., de Waal, 2013) that a selective approach to editing, focused as squarely as possible on the core analytical goal of a survey may be more appropriate than detailed review of all survey observations. For surveys supporting multiple uses, particularly ones involving multivariate analysis, there may be a need for a somewhat broader focus, but a more efficient approach may still be possible in such cases. This paper evaluates various approaches to selective editing, using various combinations of fully edited and unedited data from the 2010 Survey of Consumer Finances (SCF), a widely used survey covering household financial behavior and a variety of associated information. The paper also explores the potential importance of contamination of the imputation process under selective editing. While editing has its direct effect on individual data items, it also alters the set of information used in imputing the missing values that result from the unwillingness or inability of respondents to provide answers or from the resetting of values to missing during the editing process. The results of the paper support a selective approach to editing and they indicate that any resulting contamination of imputation is relatively minor in the case of the SCF.

Key Words: Editing, imputation, nonsampling error, quality assurance

1. Introduction

Although we should always strive to eliminate data errors at their source in a survey, there may be classes of problems that have a lower bound in the likelihood of their incidence, and some may be more controllable than others. Errors related to oversights by the survey designer ought to be highly controllable and relatively rare, certainly over repeated waves of a survey. In contrast, the actions and interactions of respondents and interviewers allow for far more varied and complex problems in a place beyond direct control of subject-matter experts. Although there are frequently potential strategies to minimize the incidence of this second type of error and working toward that end should be a primary goal, it may be that the most we can hope for in general is to find clearer means of identifying errors at an early stage and designing the survey process to capture as much information as possible that might be helpful in later remediation. Such remediation has both a long-term and a short-term aspect. For the long run, care should be taken to refine processes to the degree possible in order to avoid perpetuating known problems.

For the short term, the remediation process most often involves editing of the survey data. Editing applies some sort of filter to observed information, whether mechanical or judgmental, to identify values or observations that pass a threshold of seriousness for investigation. Resolution of such instances requires supporting information or a binding structural framework.

Editing, particularly approaches that include examination of each observation individually, is often a laborious and difficult process. Much cleverness may be expended in developing tools to make the activity less mechanically difficult, but the heart of the work remains a process of deciding whether data are sensible relative to the limits of knowledge. Where that activity has a component of human judgment, as is often the case, an additional element of potential error is introduced by the human tendency to find "patterns" even in random data. Some perceived patterns may reflect genuine problems. But in the absence of binding constraints of some sort, human intervention may also lead to "over editing" with some information being inappropriately altered. More generally, because editing can only be driven by what is observable, asymmetries in observability may create biases in editing—for example, by addressing only values that are "too large" and not those that are "too small" in some appropriate dimension.

Overall, editing is a costly and risky activity, if often a necessary one. In that light, it would seem rational to do what is possible to minimize the cost and risk, subject to a required level of data quality and a need to capture any structural knowledge revealed in editing to improve the subsequent data-generation process. Grandquist and Kovar (1997) portray extensive editing as a poor use of resources, if only because of the limits on what can be achieved in most circumstances. Indeed, Grandquist (1998) goes further, arguing that reducing editing might actually improve data quality in the longer run, if the resources were redirected toward other, more direct means of quality improvement. In a recent issue of the *Journal of Official Statistics* focused on editing, de Waal (2013), Arbués *et al.* (2013), Pannekoek *et al.* (2013) and Di Zio and Guarnera (2013) discuss much of the rest of the extensive earlier literature questioning the appropriate scope for statistical editing and propose additional alternative approaches for selective editing.

Another aspect of editing that has been less discussed than the direct effects of identifying and correcting errors is the potential effects of editing on the imputation of missing data in a survey. Any imputation based on data within a given survey (as opposed to imputations determined externally, such as through the use of register data) is a type of extrapolation of the observed structure of information, and thus subject to distortion by any errors present among the non-missing values in the data (Little and Smith, 1987). Any effort to reform editing should also take this secondary need into account.

This paper uses the 2010 Survey of Consumer Finances (SCF) as an experimental platform for investigating the effects of various strategies for selective editing on the estimated distribution of net worth, a key outcome variable. The goal is to identify practical guidelines that might support a more selective approach to editing than is currently the case, without significant loss of data quality. For this survey, the unedited data are available along with edited data resulting from a detailed case-by-case review. By varying the combinations of data in these two states, it is possible to trace out the effects of editing on the estimated wealth distribution. The paper also attempts to isolate the effects of unedited data on the estimates of population moments used to drive the imputation of missing data in the SCF.

The next section of the paper discusses the reasons editing is usually necessary and it underscores the need for an approach more focused on quality assurance over time than quality control in a specific survey. The third section describes the SCF and the history of editing in that survey. The fourth section lays out the experimental approaches and provides evidence of each on the effects on the estimated distribution of wealth. The fifth section examines the indirect effects of the various experimental approaches on imputation. The final section concludes and points to additional research needed.

2. Editing and Quality Assurance

Editing is typically seen as an important tool for identifying and addressing erroneous values in a data set. For U.S. government surveys, guidelines from the U.S. Office of Management and Budget (2006) require appropriate editing to mitigate or correct errors detectable based on available information, and an audit trail of any changes made. In practice, the proximate drivers for editing are generally outliers in some appropriate dimension, logical improbabilities or other indicators of inconsistency or incoherence. Such drivers may derive from high-level inspection of the data for distributional anomalies, post-survey filters of multivariate relationships, comments or mechanical flags set during an interview, or brute-force inspection of individual cases. Based on a review of the available evidence, values may be left untouched, set to missing and imputed, or set to another value. In many situations, this decision process introduces a risk that error is actually added to the data. Consequently, the more intensive is the editing process or the finer is the level of data review, the more risk is accepted of adding errors. Moreover, editing is usually a resource-intensive process. Thus, for reasons of both error avoidance and operational efficiency, it is necessary to find an appropriate balance in editing between intervention and abstention.

The need for editing represents a failure of some sort in the survey process. There are two important sets of failures relevant here—those that were avoidable with sufficient attention and those that were not. The most painful of the former type is a failure through negligence to address fully the information known at the time of survey design and implementation; this type of failure should not be confused with determinations made with the additional benefit of *ex post* information. Sometimes "avoidable" failures are the conscious result of a cost-benefit decision. Other failures stem largely from the complex nature of surveys, particularly through the roles played by respondents and interviewers (where there is one). Such failures may result from behaviors that are not fully controllable or ones that are not yet understood fully.

Engagement with a survey is mediated through the words of a questionnaire and the administration of the questionnaire to the respondent by an interviewer, where present. Even with great care in constructing a survey, the respondent is generally manipulable or controllable only distantly by the survey designers, and interviewers are at best very difficult to monitor or control in detail. Respondents may misunderstand questions through inattention, deficiency of literacy in the subject matter, or a different understanding of the words used in the questions or their surrounding framing. Where there is a need to address very complicated subject matter, it may not be feasible to have a questionnaire that works in every possible situation. The level of interest in the survey or trust in the process may also shape respondents' behavior. Sometimes respondents intentionally provide insincere answers. Interviewers have great potential for clarifying or otherwise controlling the progress of an interview to minimize errors, but they may make errors of their own and their behavior also may induce reactions from respondents that increase the likelihood of error.

A potential benefit of editing is that the surrounding investigation can provide insights into the deeper nature of errors and point to ways of eliminating or minimizing the errors or to more systematic means of detecting such errors for focused attention. For repeated surveys particularly, an editing process that is not structured to capture structural knowledge is not an efficient process.

If one accepts the idea that complete *ex post* elimination of error from data collected in surveys is generally an illusory goal, then there is a need for some guidance about how far editing should aim to go. With no resource constraints, a diffuse view of the relative importance of particular cases or variables for the sake of current error correction or future corrections, and a belief in the high reliability of editing decisions, it would be sensible to review every case and every variable with equal vigor. However, even with only a minimal resource constraint or a competing alternative, it becomes sensible to consider whether key analysis variables, clusters of variables or other relationships might be identified along with a means of assessing the likelihood that problems with those variables might make a detectable difference, particularly when sampling error is taken into account. As the resources become more precious, this argument should hold correspondingly more strongly.

No matter how structured and efficient an editing process may become, there are risks that the inherent decision rules for repeating surveys may become outdated for known classes of error or that new types of error will arise. To hedge against that possibility, it may be reasonable to include a (possibly stratified) random selection of additional cases or variables in an edit review.

3. Background on the SCF

This paper uses data from the 2010 SCF.[1] The SCF is designed primarily to measure the wealth and income of U.S. households, along with related information needed to support or interpret the wealth and income information. The survey employs a dual-frame sample design, with one part derived from a multi-stage area-probability sample and the other from a list sample based on statistical records derived from individual income tax returns. The list sample uses a modeling technique to create a proxy for wealth, which is used to stratify the sample and support oversampling of wealthy households. In 2010, the area-probability sample had a response rate of about 70 percent and averaged across all the strata the list sample had a response rate of about 30 percent; the realized sample included 6,492 observations, of which 1,480 were from the list sample.

The survey questionnaire is long and it can be very complicated from the perspective of some respondents. The typical interview requires about 75 minutes, but some interviews run for several hours, perhaps split over multiple sessions. Given the focus of the survey,

---

[1] See Bricker *et al.* (2012) for summary information on the data, a general overview of the survey methodology and references to more detailed technical material.

it is necessary to cover a variety of financial categories, some of which may be unclear for people with a low level of financial sophistication or inadequately detailed for some respondents with complex financial situations. The questionnaire is carefully designed in terms of its wording, sequencing and other framing to minimize errors. Since 1995, the SCF has used computer-assisted personal interviewing (CAPI) to collect the data for the survey. This technical approach allows for an elaborate protocol to support data quality at the point of data collection—through the use of conditional routing, the inclusion of various real-time edit checks, the availability of an electronic glossary, and a facility for recording comments on any exceptional situations. At the conclusion of every interview, interviewers are obliged to complete a "debriefing" about their experience in administering the interview and any problems that arose there; in addition, they have an opportunity to clarify or amplify any of the comments they have recorded during the interview.

As in most surveys, item nonresponse is a problem in the SCF. Although there is nonresponse across a wide variety of variables in the survey, the problem is most acute for questions with answers given in dollar terms. To mitigate this problem, the survey employs an automated probing technique that aims to obtain a range response when a complete response cannot be obtained.[2] As indicated by the information shown in table 1, range responses tend to be the dominant form of incomplete information. Missing data in the survey are imputed using a multiple imputation system that uses the range information to draw imputations from a truncated version of the conditional distribution of the missing data.[3] The core of the imputation process is a type of randomized regression procedure.

Table 1: Initial status of selected dollar variables, where the item is present or unknown whether present; percent.

Item	*Good value*	*Range*	*Missing*
House value	92.0	7.4	0.6
Main checking account	80.2	16.6	3.2
Certificates of deposit	75.7	17.7	6.7
Stock mutual funds	74.9	14.7	10.4
Wages of household head	85.7	12.4	1.9

The data collected in the 2010 survey were subjected to a very intensive editing process, which is described in more detail below. To understand how this process arose and to provide insight into where it might go in the future, it may be helpful to review the history of editing in this survey. In 1983, when the modern SCF began, the survey editing process was radically different. Interviewers, who completed interviews using a paper questionnaire, were required to perform a review of each case to ensure that all necessary questions had been answered (and in the "right place," since a very similar question might appear in more than one place), that the answers had been recorded clearly and that any complications were duly noted, either in marginal notes at the relevant point in the questionnaire or in a "thumbnail sketch" completed by the

[2] See Kennickell (1997) for a discussion of range data in the context of the SCF.
[3] See Kennickell (1998) for a description of the FRITZ system for multiple imputation developed for the SCF.

interviewer after the interview. The managers of the interviewers would then review at least a selection of interviewers' work. Subsequently, all questionnaires were returned to a central processing point, where a team of editors performed a further review of each case to prepare it for data entry. The task of the editors included coding open-ended responses and ensuring that information was recorded in the correct columns of structured fields. More importantly, they identified problems, which they sorted into those that could be resolved using the information already in the questionnaire or readily available externally (for example, using date-appropriate exchange-rate information to record a value given in non-local currency) and those that needed additional technical knowledge or judgment. Subject-matter experts on the project staff at the Federal Reserve would periodically visit the processing center and address the cases identified by the editors for such additional review. At every stage up to the point of data entry, any changes or additions were carefully noted in colored pencil, where the color signified the role of the person taking the action. Once the data were available in electronic form, they were processed by the Federal Reserve project staff, using computer algorithms intended to identify additional potential problems, which were then addressed in more depth. The combination of an escalating process of case-level review and computer-driven error detection was efficient and effective. User feedback about the data ultimately released to the public was effective in identifying a wide variety of small problems undetected in the earlier review.

Two important factors changed the context in which editing for the SCF took place, and the consequence was that the editing process was ultimately radically changed. First, the SCF moved to CAPI in 1995. While CAPI did make possible many types of real-time control that would be virtually impossible with a "primitive" paper questionnaire, the change was not uniformly positive. Above all, the paper questionnaire provided a concrete representation of the substance of the entire interview that was open an accessible to all parts of the survey team, both during and after the interview; in contrast, information in an electronic instrument is a more abstract object. Electronic data were less amenable to broad interaction, without extensive programming which was not feasible in 1995; by now, such programming should not be a binding constraint, but the previous framework to make use of it has been lost. The second factor was a change in the labor market that provided the talent necessary to support the work of the processing center. Many aspects of society in the 1980s were shaped by a different expectation than now about the role of women and their possibilities for a career. That limited labor market had advantages for areas that depended on worker sophistication at low wages. In a limited sense, this was a "golden age" for surveys, and it did not last. As the market changed, it became progressively harder to attract people who would be willing to perform the primary central editing at a wage that was seen to be possible at the time. With increasingly limited ability to attract reliable editors, the editing task was pulled progressively to the level of the subject-matter experts, who had previously only edited information that had been identified as needing attention by a human editor or a computer algorithm. Pulling the entire edit task to the central subject-matter experts was an expensive step, in terms of money as well as the difficulty of finding people who could and would do such work as well as perform research and policy work at other times. The change was intended to be temporary.

The centralized editing for the SCF currently entails a review of each survey observation, and potentially every variable. The work is driven by the comments and debriefing notes provided by interviewers and by the output of a computer algorithm descended from the one employed in the earlier, decentralized review. In addition to identifying particular items for review, the algorithm assigns a score to each observation. The score is the sum of the subjectively assigned scores assigned to the individual items identified. Once a review is begun, all parts of an interview and related material might be reviewed. An editor would have available a readable representation of the case-level data, the interviewer's comments, verbatim responses by the respondent, the items identified by the algorithm, some case-specific summary data, and a variety of tools to perform calculations based on the data or to look up potentially relevant information in external sources. Over time, the process has become increasingly automated, in an attempt to lessen the mechanical burden on the editors. Nonetheless, it remains a difficult, tedious and time-consuming process.

Aside from the obvious benefit of having the necessary editing work done somewhere, centralization had the positive side effect of greatly clarifying the nature of many types of nonsampling error. This positive result stemmed from matching high expertise in data interpretation with close examination of individual cases and the cumulation of patterns of behavior. Such understanding has led to changes in the questionnaire wording and sequencing, real-time edit checks, interviewer training and retraining, management of field work, and structure of the computer algorithms used to highlight particular variables or clusters of variables for closer examination. Systematic capture of such information is valuable, and it should remain part of the editing process, no matter the shape of any future reform of that work.

However, there has been a perennial problem in the exploitation of this information in the short term to change behavior during the course of a survey, through retraining tailored to a particular interviewer or a more diffuse group. The problem appears to be largely one of communication. Unfortunately, it is all too easy for such interventions to appear overly critical or too ambiguous to someone who has not interacted directly with the data, particularly when the interventions originate from a group of people seen as remote from the process of data collection. Involving people closer to the data collection in the data review might have the benefits of more effectively neutralizing opposition or ill feelings among those targeted for re-training, allowing a broader group to understand the nature and consequences of errors in the questionnaire administration, as well as lessening the editing burden on the central staff. A prototype effort of this sort was undertaken for the 2013 SCF, focusing on a set of problems that were relatively well defined.[4] The participants reported a desire for more extended involvement and preliminary evaluation of this effort suggests that it was successful in addressing problems. Further progress in re-outsourcing editing may be helpful in recapturing some of the earlier benefits of escalating review. However, more detailed investigation of the relative benefits of editing, as described below, may be even more important in increasing the efficiency of the editing process.

4. Experimental Approaches to Editing

As noted above, detailed editing is very burdensome and it may lead to over-editing that creates error where none existed previously. If any cost of editing is taken into account, it is sensible to consider how the process might be made more efficient while maintaining a sufficient level of data quality. Obviously, to be useful in any practical way, any change would need to be based on information that is available before editing takes place. Because the 2010 SCF was edited in its entirety, it is possible to use those data to create

[4] See Bricker and Kennickell [2013] for a summary of this effort.

various mixtures of edited and unedited data and examine the consequences for key outcomes.

For purposes of this work, the distribution of net worth is taken to be sufficiently indicative of the core purpose of the SCF. Calculating net worth in the SCF requires aggregation over a large number of assets and liabilities, each of which might be affected materially by error. The seven approaches listed in table 2, which are described in more detail below, involve the creation of different mixtures of edited and unedited data from the 2010 SCF that are intended to span a sufficient set of alternatives to judge the feasibility of more selective editing for the SCF. In each instance, the data were fully imputed and weighted independently.[5]

Table 2: Experimental variations.

Experiment	*Edited data included for:*
1	All cases
2	No cases
3	Cases with "substantial edits"
4	List sample caes only
5	Top quarter of cases by interviewers' comments
6	Top quarter of cases by algorithmic score
7	Top 15% of cases by interviewers' comment or top 15% of cases by algorithmic score

The baseline for the comparisons, the full set of edited data (#1 in the table), was re-imputed and re-weighted to ensure comparability with other variations. The second alternative uses only the unedited data.[6] The third option includes edited data for the nearly 40 percent of cases that were edited in a substantial way, in the sense that at least two dollar values had an absolute change of at least $1,000 between the unedited and the edited data or two dollar values originally not missing were set to missing. This option is not one that could be designed *ex ante*, and it is included here only to provide perspective on an approach that might otherwise seem the appropriate aspiration.

The design of the remaining four alternatives is based only on information that could be known independently of any knowledge obtained through editing; each was structured to include approximately the same number of edited survey cases. In the fourth version, edited data are included only for the list sample cases (about 23 percent of the total). The list sample contains the overwhelming majority of the wealthiest cases in the SCF and a rapidly decreasing share below that wealth level. The fifth version relies on an indicator of the subjective view of interviewers about their cases. Because the interviewers' comments should be largely driven by problems or ambiguities detected during an interview, one would expect the amount written to be generally reflective of what needed

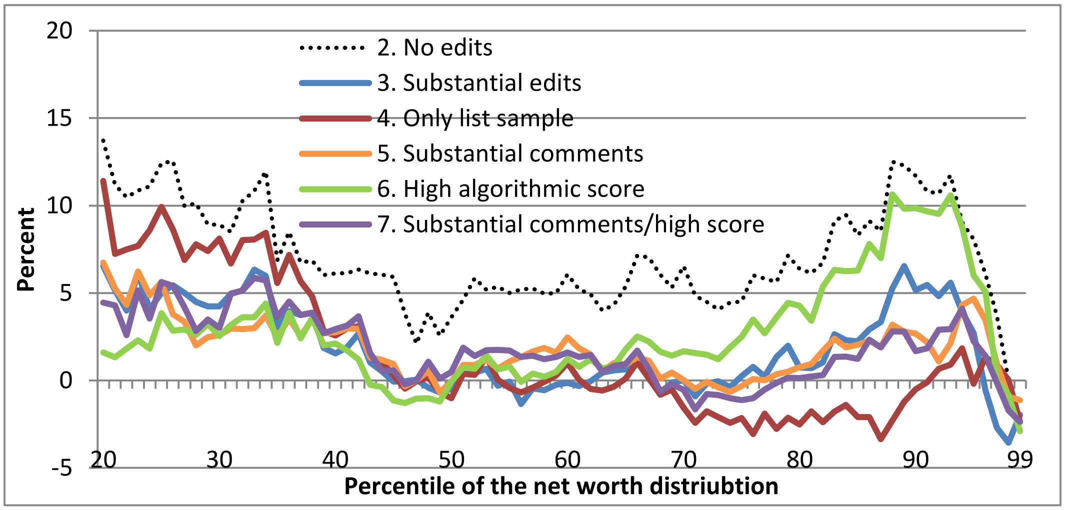
[5] In contrast to the normal practice of multiple imputation for missing data in the SCF, the data in the experimental data sets were only singly imputed. This simplification was taken to make the work necessary for the experimental comparisons manageable. Sensitivity analysis indicates that the results presented here are not meaningfully affected by this choice. Because the SCF weighting design is, in part, dependent on imputed data, it was important for consistency to weight each experimental data set separately.

[6] There is some small degree of editing involved in producing the set of unedited cases. Approximately 65 interviews were deleted for issues related to falsification, excessive missing data, completing the interview with someone other than the correct respondent, or other glaring issues. These interviews are not included in any of the experimental data sets here.

to be said to describe the problems. For this purposes of the experiment, edited data were included for the 25 percent of cases with the largest total length of interviewer comments and debriefing remarks. The sixth alternative uses a more objective *ex ante* indicator of potential problems, the score from the algorithmic review of the data. For this option, edited data were included for cases with a score at or above the 75th percentile of the distribution of the scores. Owing to clumping around the 75th percentile, this approach selects about 27 percent of cases. The final experimental version combines the fifth and sixth approaches. To maintain approximately the same number of observations as in each of these versions for the sake of comparability, edited data were used for cases at or above the 85th percentile of either the distribution of the length of the comments or the distribution of the score; this selection results in including edited data for about 27 percent of all cases. More than half of all cases are in none of these last four groups.

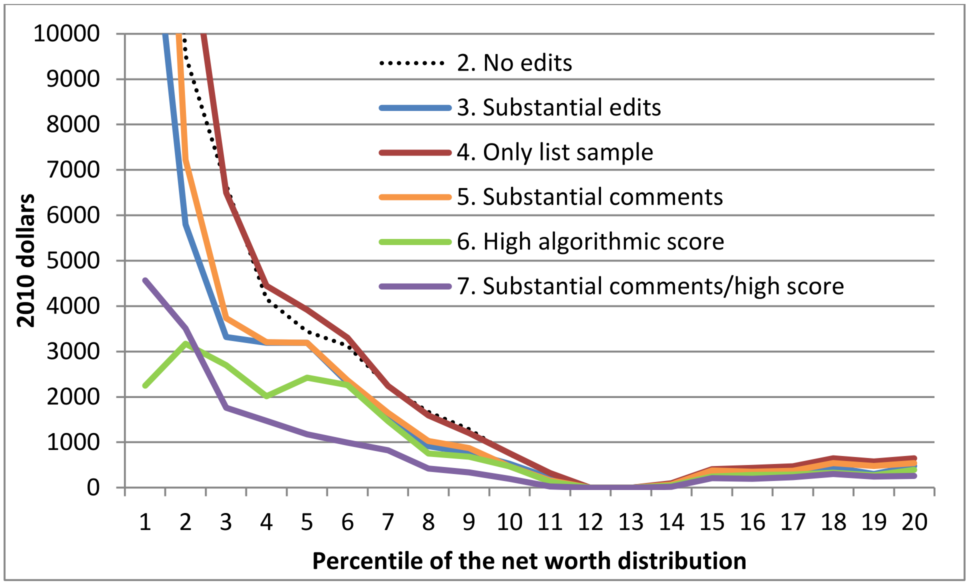
Figure 1 shows a quantile-difference plot for the distribution of net worth under the second through seventh options, taking the fully edited first option as the baseline.[7] The difference is defined in percentage terms as the value of the distribution at a given percentile under the edited data minus the value under the experimental version, divided by the absolute value of the value of the edited data at the same percentile. The absolute value is used in the denominator so that the change shown reflects what would be seen in changes in levels. This approach works well for the region of the wealth distribution above about the 20th percentile.[8] Below that level relatively small difference in levels can imply very large percentage changes. For clarity, the region below the 20th percentile is dealt with separately in terms of the differences in levels in Figure 2.

Figure 1: Percent quantile-difference plot for net worth under various experimental scenarios.



[7] For present purposes, it is assumed that errors and distortions induced by editing are sufficiently outweighed by the positive effects of editing that such problems can be ignored.
[8] The 20th percentile of net worth according to the final version of the 2010 SCF data was $4,400, the 10th percentile was $-950 and the 1st percentile was $-90,000.

Figure 2: Quantile-difference plot for net worth under various experimental scenarios.



The data show clearly that editing matters for the SCF. As indicated by the dotted black line (#2) in the figures, the wealth quantiles of the edited data are about 5 to 10 percent higher than those of the entirely unedited data in the range above the 20th percentile, and they differ by substantial dollar amounts in the lower percentiles. Although not achievable *ex ante*, the inclusion of edited data only from cases that had substantial edits of the sort described above (#3) might be thought to provide a reasonable lower bound on what might be achievable through other schemes of selective editing. In fact, this option does show a substantial improvement (closer to the zero line), particularly in the range between the 45th and the 75th percentiles of net worth. However, substantial differences remain elsewhere in the distribution and as discussed below, the remaining alternatives are implementable *ex ante* and perform about as well in this region and better in other regions.

If only the cases in the list sample are edited (#4),the outcome is much better in the top 10 percent of the wealth distribution than under option 3, about the same in the middle, but worse or about equally badly almost everywhere else. The superior performance at the top of the distribution is not surprising, given the great concentration of list-sample cases at the top of the wealth distribution. The fact that this option "over-shoots" between about the 70th and 90th percentiles must reflect a greater proportion of area-probability cases in this region that had offsetting reductions in the fully edited data.

The option driven by the extent of interviewers' comments (#5), the most directly subjective approach to targeted editing considered here, performs about as well as option 3 across the broad middle of the distribution and better or nearly as well elsewhere except in the top decile. A sizable gap remains for the wealthiest 20 percent. In contrast, the option driven by the algorithmically determined score (#6) performs only somewhat better than the entirely unedited data across the wealthiest 40 percent of the distribution, but it performs better or about as well the other options elsewhere. Option 7, which pools options 5 and 6, captures the relatively superior performance of option 5 at the top of the distribution and option 6 at the bottom of the distribution, while maintaining good performance across the middle.

Other strategies or combinations of strategies, such as including edited data for certain strata of the list along with a mix like option 7, might improve performance at the top of the distribution without adding greatly to the fraction of all cases edited. Closer examination of observations with important changes as a result of editing might also lead to improvements in the basis of the algorithmically determined option. Modeling the differences between the edited and unedited cases more generally might lead to additional insights into the sources of detectable differences, some of which may be sufficiently recognizable *ex ante*.

Any realistic approach to selective editing is necessarily based on observables. Selection on those observables should generally be guided by evaluation of past practice and outcomes, but there is an argument for looking more broadly. Non-stationarities may arise throughout the process of data collection and processing and such changes may have important effects on data quality. Including a random sample—or a sample stratified on some relevant dimension—of additional cases would provide a degree of protection against unknown changes as well as give a means of tuning a selective editing process over time.

5. Indirect Effects of Editing on Imputation

As noted earlier, data errors may affect outcome measures directly or indirectly via contamination of estimates needed to support the imputation of missing data. In the largely regression-based imputation approach in the SCF, there is a clear way to delineate the conceptual issues related to contamination. The box below lays out a basic regression model for imputation and the basic differences in implementation with edited or unedited data.

Box: Effects of editing on imputation.

Let the imputation model be given by Y = $X'\beta + \epsilon$.
Let $\ddot{Y}$ be a vector of values subject to imputation, where all values have been reviewed.
Let $\dot{Y}$ be the vector of values subject to imputation, where none of the values have been reviewed.
Let $\ddot{X}$ and $\dot{X}$ be similarly defined for a matrix of variables used to condition the regression.
In the edited data, $\ddot{\beta} = [\ddot{X}'\ddot{X}]^{-1}[\ddot{X}'\ddot{Y}]$.
In the unedited data, $\dot{\beta} = [\dot{X}'\dot{X}]^{-1}[\dot{X}'\dot{Y}]$.
Thus, in the edit data the imputed value of element i is $\hat{\ddot{Y}}_i = \ddot{X}_i'\ddot{\beta} + \tilde{\ddot{\epsilon}}_i$ and in the unedited data the imputed value is $\hat{\dot{Y}}_i = \dot{X}_i^i\dot{\beta} + \tilde{\dot{\epsilon}}_i$,
Where $\tilde{\ddot{\epsilon}}_i$ and $\tilde{\dot{\epsilon}}_i$ are draws from the estimated distribution of the error term under the edited and unedited data, respectively.

Clearly, the effects of editing on an imputation are felt both through the direct effect of errors in the value of the conditioning variables, X_i, and indirectly through the estimated value of β. It would be helpful to be able to understand the degree of damage done by each. Among other things, if editing does not affect imputations in a serious way, then at least the early stages of imputation, if only to facilitate data inspection, could proceed concurrently with editing and allow for a more timely completion of the survey processing. Fortunately, the SCF imputation software is structured so that it is possible to separate these effects. The software is structured so that the data set used to compute the various moments needed for the imputations can be computed from a different data set than the one being imputed. Thus, one can look separately at is $\widehat{\dot{Y}_\iota} = \dot{X}_i^i \dot{\beta} + \tilde{\dot{\epsilon}}_i$ and $\breve{\dot{Y}}=\dot{X}_\iota' \dddot{\beta} + \tilde{\dot{\epsilon}}_i$ to gauge the effects of contamination in the estimates of β.

Figures 3 through 8 show percent quantile-difference plots for net worth, under each experimental alternative described in the preceding section and for two approaches to imputation. As before, the baseline in all cases is option 1 in table 2. In each of the figures, the blue line reproduces the results for each of the earlier experiments separately—that is, results based on $\widehat{\dot{Y}_\iota}$. That is, both the conditioning variables and the relevant moments are based on the data set appropriate to each experimental alternative. The calculations underlying the red line use imputations of the form $\breve{\dot{Y}}$, with the moments computed from the fully edited data and the conditioning variables taken from the data set appropriate to the experimental alternative.

Generally, using fully edited data to compute the moments for imputations in the completely or partially unedited data results in a worse outcome than using the unedited or partially edited data both for estimating the moments and for supplying the relevant conditioning variables. The two outcomes are most similar for the versions including no edited data (#2, figure 4), the version including edited data for cases that had substantial edits in the sense described above (#3, figure 5), or the version including edited data for cases with substantial comments or a high algorithmic score (#7, figure 8). Differences are pronounced for the other versions, but generally not seriously so.

The result for the entirely unedited data is perhaps most interesting, in that it implies that most of the harm is done by bad data in imputation operates through the misrepresentation of the characteristics of individual cases, rather than through contamination of the moments needed for imputation. For the other alternatives, the results suggest that not using fully edited data does not result in a worse outcome.

What underlies this result? Although there are a variety of possible explanations, one plausible reason is that noise in elements of X may lead to estimates of the corresponding element of β being biased closer to zero. Thus, the effects of noisy conditioning variables would be reduced. Another factor may be the collection of partial information in the SCF, in the form of ranges, when a value is not otherwise available. As noted earlier, the great majority of incomplete responses are accounted for by respondent-provided ranges and this bounding information necessarily places a constraint on how far wrong imputation can go. Other surveys that do not allow for range responses may experience different effects of editing on imputation.

Figure 3: Percent quantile-difference plots for net worth; covariances computed using edited data and covariances computed using unedited data (#2).

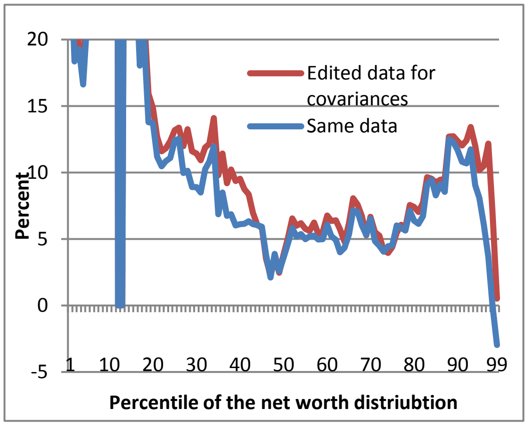


Figure 4: Percent quantile-difference plots for net worth; covariances computed using edited data and covariances computed using data set containing only substantial edits (#3).

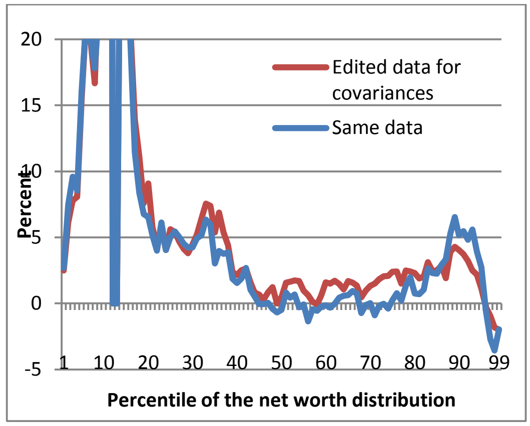


Figure 5: Percent quantile-difference plots for net worth; covariances computed using edited data and covariances computed using data set containing edits only for the list sample (#4).

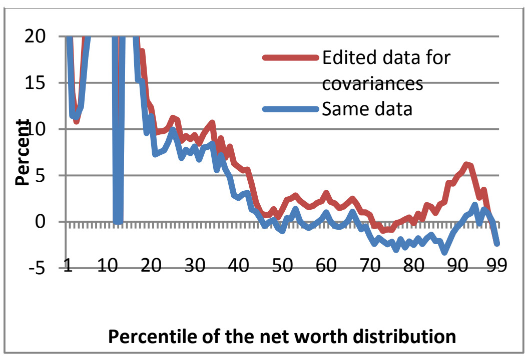


Figure 6: Percent quantile-difference plots for net worth; covariances computed using edited data and covariances computed using data set containing edits only for the cases with substantial comments (#5).

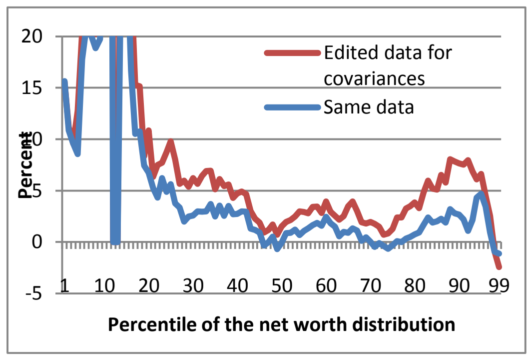


Figure 8: Percent quantile-difference plots for net worth; covariances computed using edited data and covariances computed using data set containing edits only for the cases with a high algorithmic score (#6).

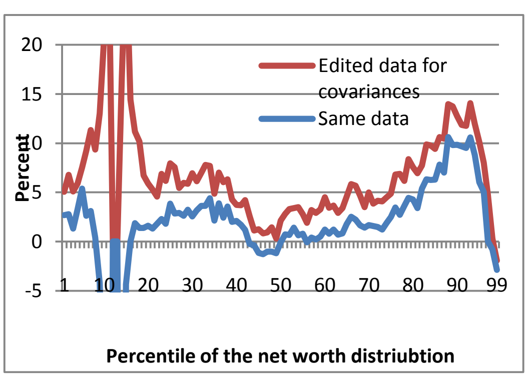
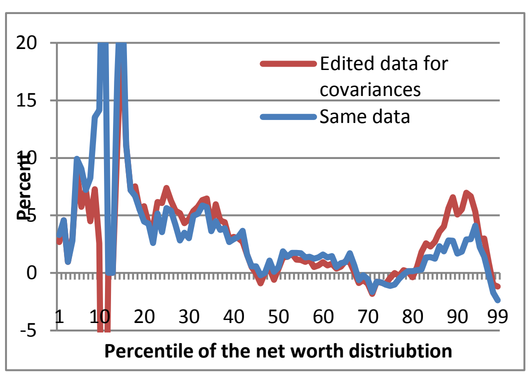


Figure 7: Percent quantile-difference plots for net worth; covariances computed using edited data and covariances computed using data set containing edits only for the cases with substantial comments or a high algorithmic score (#7).



6. Conclusion and Future Research

Historically, the SCF has been sharply focused on data quality and editing has been an important element in supporting data quality, both in terms of mitigating problems with the data in a given survey and in terms of developing strategies to prevent or limit error in the future waves of the survey. In a survey as complex as the SCF, editing will doubtless always have a central role in such quality control and quality assurance. But like many other activities that have positive consequences, editing has also costs. Detailed editing of every individual case, as currently practiced in the SCF, is expensive and cognitively challenging, and absent serious constraints on data changes it increases the likelihood of introducing additional error into the data.

This paper has considered a set of alternative, experimental protocols for editing that might justify a substantial reduction in the level of effort devoted to editing. Because the 2010 SCF, the basis of this paper, was fully edited, it is possible to consider various mixtures of edited and unedited data in order to play out the consequences for the survey outcomes of editing only a part of the cases. Here, the relevant outcome is taken to be the full distribution of net worth, which is the sum of a large number of detailed questions in the survey; a more complicated approach based on multiple outcomes could also be considered. For each mixture of edited and unedited data, the data set was separately imputed and weighted, in order to enhance the realism of comparisons across the protocols. The results show clearly that editing matters. Furthermore, the experiments indicate that it is possible to reduce the level of noise in the data very substantially, even when only about a quarter of the survey cases are examined. Overall, the best performance among the alternatives considered was obtained from a rule requiring review of all cases with substantial amounts of interviewer notes (a subjective indicator) or a high score from an automated review of the data for potential errors (an objective indicator).

Other mixtures based on *ex ante* knowledge should be considered to refine this judgment to the point of an operational rule. For example, including the wealthiest few strata of the list sample, or some part of those strata, along with the cases with high values of the subjective or objective indicators would seem likely to improve the outcome for the top decile of the wealth distribution. The subjective indicator might be further improved by prior review to identify commentary that contains actionable information; such work might be the next step in the return of part of the editing task to project staff closer to the point of data collection. The objective indicator might also be improved by more systematic review of the effectiveness and importance of each test in the review. Further insights might be gained by modeling the difference between edited and unedited data directly, in terms of other characteristics, such as demographic or other variables collected in the survey or the measures of the difficulty of obtaining the interview or other aspects of paradata. To guard against changes in the underlying processes generating errors, it would be advisable to include an additional random (possibly stratified) sample of cases for review.

As a side benefit of the investigation, the paper also provides evidence on the effects of editing on model-based imputation. Using unedited data for imputation poses the risk that the parameters estimated for the model might be seriously contaminated as well as that the conditioning variables against which those parameters are applied may be erroneous. With the information available, it is possible to distinguish the effect of error through the parameter estimates. The data show that at least for the SCF, distortion via the parameter estimates is not a large factor relative to other error. Indeed, the results indicate that when the edited data are used to compute the parameters and those parameters are applied to unedited conditioning variables, the deviation from the fully edited outcome actually tends to be larger. This result may reflect a biasing of coefficients of particularly noisy variables in the unedited data toward zero, or it may be related to the relatively constrained nature of SCF imputation for dollar-denominated variable, for which partial information in the form of ranges is an important type of missing data.

Acknowledgements

The opinions expressed in this paper are those of the author alone and do not necessarily reflect the views of the Federal Reserve Board or its staff. I am grateful to many people for their contributions toward preparing the data used in this study, but I can only mention a small number in this place. The principal editors for the 2010 SCF were Jesse Bricker, Brian Bucks, Gerhard Fries, Traci Mach, Kevin Moore, John Sabelhaus and the author. The central office staff at NORC, particularly Catherine Haggerty, Micah Sjoblom, Katherine Del Ciello, Shannon Nelson and Karen Veldman, provided the essential guidance for field operations and participated in an experiment to reform the overall editing process. The hardest work always takes place in the field by interviewers and their managers, without whose dedication to following the complex SCF protocol the data would likely be meaningless. Finally, I am grateful to the thousands of SCF respondents, without whose generosity and patience there would be nothing at all to say here.

References

Arbués, Ignacio, Pedro Revilla and David Salgado [2013] "An Optimization Approach to Selective Editing," *Journal of Official Statistics*, Vol. 29, No. 4 (December), pp. 489–510.

Bricker, Jesse, Arthur B, Kennickell, Kevin B. Moore, and John Sabelhaus [2012] "Changes in U.S. Family Finances from 2007 to 2010: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, http://www.federalreserve.gov/pubs/bulletin/2012/pdf/scf12.pdf.

Bricker, Jesse and Arthur B. Kennickell [2013] "Shared Understanding and Data Quality in the SCF," *Proceedings of the Government Statistics Section*, 2013 Joint Statistical Meetings.

de Waal, Ton [2013] "A Quest for Efficiency and Data Quality," *Journal of Official Statistics*, Vol. 29, No. 4 (December), pp. 473–488.

Di Zio, Marco and Ugo Guarnera [2013] "A Contamination Model for Selective Editing," *Journal of Official Statistics*, Vol. 29, No. 4 (December), pp. 539–556.

Grandquist, Leopold [1996] "The New View on Editing," in *Data Editing Workshop and Exposition*, Federal Committee on Statistical Methodology Working Paper 25 (www.fcsm.gov/working-papers/wp25a.html), pp. 16–23.

Grandquist, Leopold [1997] "Editing of Survey Data: How Muich Is Enough?" in *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Duppo, N. Schwarz and D. Trewin (eds.), New York: Wiley, pp. 415–435.

Grandquist, Leopold [1998] "Efficient Editing—Improving Data Quality by Nidern Editing," paper presented at the Conference on New Techniques and Technologies for Statistics, Sorrento, Italy.

Hedlin, Dan [2003] "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics," *Journal of Official Statistics*, Vol. 19, No. 2 (June), pp. 177–199.

Kennickell, Arthur B. [1997] "Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances," www.federalreserve.gov/econresdata/scf/files/rangepap0197.pdf.

Kennickell, Arthur B. [1998] "Multiple Imputation in the Survey of Consumer Finances," www.federalreserve.gov/econresdata/scf/files/impute98.pdf

Little, Roderick J.A. and Phillip J. Smith [1987] "Editing and Imputation for Quantitative Survey Data," *Journal of the Americal Statistical Association*, Vol. 82, No. 397 (March), pp. 58–68.

Pannekoek, Jeroen, Sander Scholtus and Mark van der Loo [2013] "Automated and Manual Data Editing: A View on Process Design and Methodology," *Journal of Official Statistics*, Vol. 29, No. 4 (December), pp. 511–538.

U.S. Office of Management and Budget [2006] "Standards and Guidelines for Statistical Surveys" (September), www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.