Who Provides the Best Data: Respondent Characteristics, Financial Literacy, and Data Quality in the Survey of Consumer Finances

Joanne W. Hsu[1], Richard A. Windle[1]
[1]Federal Reserve Board, Constitution Ave NW & 20th St. NW, Washington, DC 20551

The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors.

Abstract
In complex surveys spanning diverse populations, the interviewed sample can generate data with a wide range of quality. Some respondents forget information, provide information in incorrect places, misunderstand questions, or even refuse to provide data at all, while others will supply answers that almost perfectly represent their situations. In the context of a highly-detailed household financial survey, we investigate respondent-level determinants of data quality. Do specific demographic groups provide higher quality data than others? Are any respondent traits, including financial literacy, particularly indicative of higher quality data? Results suggest that greater care could be taken for respondents with less confidence in financial topics in order to generate increased data quality for financial questions.

Key Words: financial survey, data quality, financial literacy

Surveys are used to gather a variety of data, some objective and factual, others subjective and based on opinion. For survey questions eliciting objective data, there exist factually correct answers to the questions. For instance, if a survey asks how much money is currently in the respondent's checking account, a factually correct answer is possible, though respondents may not be willing or able to provide that factual response. When, then, do respondents report the correct answer? Put another way, when do respondents provide data of high quality, data that most accurately reflect the real situation of the respondent?

The accuracy of a response can be difficult to ascertain, as the truth is often unobservable to the researcher. After all, if one had the correct data already, one would not need to conduct the survey in the first place. To determine the quality of data provided by respondents, researchers typically rely on measures believed to be highly correlated with, or indicative of, accuracy or data quality. These measures can then be analyzed for relationships with survey characteristics to improve survey design. For example, such a measure could be useful when experimenting with different question wordings, question order, or survey layouts, and trying to understand the tradeoffs involved with the various options.

In this paper, our goal is instead to determine how data quality is related to respondent demographic characteristics, and we propose a new measure for doing so in the context of

a household survey on financial topics. If certain respondent characteristics that are evident early on in the interview process are highly correlated with data quality, it could be possible to increase data quality by immediately giving respondents who tend to give lower quality data additional explanations and definitions of financial terms, and thus increase the accuracy of survey data.

1. The Survey of Consumer Finances

The analysis in this paper uses data from the Survey of Consumer Finances (SCF), a nationally representative survey of American households conducted every three years. Sponsored by the Federal Reserve Board, it gathers data on assets, debts, income, and demographic characteristics through face-to-face or telephone interviews.[1] The survey is designed to enable researchers to essentially create a financial balance sheet for each household, which can then be analyzed in conjunction with income and demographic features, such as race, age, or household composition.

1.1 Data Editing

The SCF covers a broad range of economic and financial topics that can be challenging in an interview setting. As such, interviewers are encouraged to probe and provide clarifying comments throughout the survey. All data from each interview, including interviewer comments, are reviewed by project staff, who can then conduct data editing as appropriate, particularly for objective, factual questions. There are two main motivators for editing data.[2]

First, inconsistencies in responses can arise due to purposeful redundancies built into the design of the SCF. For example, data on pension income is gathered in the work history section, as well as the income section. This allows a data editor to examine the two values and see if they are inconsistent with one another, in which case the incorrect data can often be fixed with a reasonable amount of certainty.

The second motivator is interviewer comments. The SCF is collected using CAPI and CATI, and the software allows an interviewer to make comments at any point in the survey. Sometimes these comments record the interviewer's observations and explanations, while other comments offer clarifications from the respondent. These comments can be used by the editor to enter any missing data, such as an unreported savings account remembered later by the respondent or unusual pension details that may not have been correctly recorded.

Data editing is only performed when doing so is reasonably expected to increase data quality, and thus it is a natural focus of any data quality analysis.

1.2 Types of Edits and Question Metadata

A particular data edit could be implemented based on varying levels of certainty or judgment.

---

[1] The survey uses a dual frame sample, where 1,500 people are selected based on their predicted wealth, and another 3,500 to 4,000 are selected based on geography. See Bricker, et al. (2017) for more details on the design and content of the SCF.

[2] See https://www.federalreserve.gov/econresdata/scf/files/ASA2011.1.pdf for more details on data editing on the SCF.

Sometimes these edits involve moving values from one variable to another. For example, the SCF asks for balances of checking accounts separately for each account, but some respondents will mentally combine them and report them as one, and the interviewer might leave a comment that this response should be split. These, and other edits based directly on explanatory comments left by the interviewer, would involve very little judgment on the part of the editor.

Other edits require more judgement. For example, a comment might indicate that the respondent remembers, at the end of the assets section, that they have another account of some kind that they rarely use, but do not remember whether it is a checking or a savings account. The editor would then have to decide where to place this account. Another example could be a pension with an associated account amount from a previous job, but the respondent cannot recall any details about it. The editor would have to decide if they could label it as a certain type, and whether or not they have good reason to estimate the amount in the account.

In the SCF, every variable has an associated 'shadow variable' that contains a metadata code from which we can (a) identify if the data has been changed from the respondent's original response, and (b) infer the level of certainty or judgment required to make the change. Table 1 lists the most common values and their meaning.

Table 1: Selection of shadow variable values and their meanings

Shadow Variable Value	**Meaning**
0	Respondent not required to answer question, no data
1	Original data from the respondent, unedited
2	Data was moved from another section of the survey to the current location
5	Data was edited, but did not require any judgement by the editor, likely due to a verbatim comment from the respondent
10	Data originally located in the associated variable was moved to another location in the survey
13	Data was edited, and judgement was required from the editor, likely due to an ambiguous comment or a data inconsistency
1041	Data was imputed using a range provided by the respondent
2050	Data was imputed, and original response was 'Don't know'
2053	Data was imputed, and original response was 'Refused'
2098	Data was imputed, and original value was missing

The vast majority of these are either 0 or 1. 0s indicate that the survey variable's question was not asked of the respondent, and thus will be empty. 1s indicate that the question was asked of the respondent, and their original answer is in the associated variable.

Values that are missing due to an answer of 'Don't know' or 'Refused' from the respondent or a data edit are imputed (shadow values of 2050, 2053, 2098). Likewise, if a respondent or a data editor enters a possible range of values for a question that data point will also be imputed (shadow value of 1041).

Higher values of the shadow variable indicate that some sort of editing or imputation was performed on the value in the associated variable.

2. Measures of Data Quality

Common measures of data quality have both positive and negative properties in the context of the objective personal financial data collected on the SCF, particularly if we are most interested in the degree to which the data accurately reflect the respondent's real-life financial situation.

Item nonresponse (Fricker and Tourangeau, 2010), also called the fraction of missing information (FMI) (Wagner, 2010), is defined as the percent of a case that a survey respondent does not answer. Missing data cannot accurately reflect reality, and so by definition has no quality. Thus, this measure can serve as a proxy, such that the higher the percentage of missing data, the lower the data quality.

In many surveys, questions that a respondent should answer but chooses not to (for a variety of reasons) are left missing for data users. However in other surveys, including the SCF, this data is edited or imputed. Editing and imputation can also be used to overwrite responses that may have been erroneously reported or is unlikely to reflect the respondent's real-world situation. Thus, *change rates* can have a broader scope than item nonresponse, encompassing more sources of low data quality. The greater the number of edits and imputations, the lower the data quality (Davern, Rockwood, Sherrod, and Campbell 2003). Because of this broader scope, it is reasonable to hypothesize that the edit and imputation rate better reflects data quality, since it allows for more situations where data does not reflect reality: both item nonresponse, and responses that are detectably inaccurate, given the context of the rest of the interview and comments. For the rest of the paper, this measure will be referred to as the change rate. That is, the rate at which the original values provided (including nonresponse) by the respondent have been changed. In the next section we will propose a refinement of this change rate that can be computed from the SCF's metadata.

Two other common measures are not applicable in the context of the SCF. *Response consistency* is often employed for panel surveys on questions that should remain the same over time, such as race, but other surveys can employ a similar measure. For example, the Current Population Survey (CPS) performs re-interviews on a subset of respondents and calculates the response variance (Fricker, Tourangeau, 2010). The higher the variance, the lower the implied data quality. The SCF is not a panel survey and does not employ re-interviews, so it would be difficult to perform such a measure.

Length of open-ended answers is one final measure of data quality. Longer length response is generally associated with higher response quality (Smyth, Dillman, Christian, Mcbride, 2009). The SCF asks a very limited number of open-ended questions, but the nature of the questions does not generally lead to particularly lengthy answers, so it is questionable whether or not simple length would indicate data quality in the SCF.

3. Methods

For the remainder of this paper, we will examine two measures of data quality: one conventional measure, and a newly-designed measure exploiting features of the SCF's metadata, which differentiates the different types of edits made to data.

3.1 Change Rate

First, we follow past literature and analyze a measure of data quality based on occurrences of edits and imputations. We define the *change rate* as the percentage of a respondent's data that was changed due to editing and imputation. We performed a comparison of the pre-edited dataset and the post-edited dataset. A change was said to have occurred whenever a datapoint took different values in the two datasets. The percentage of a respondent's case that was changed was calculated by counting the number of changed values in that case and dividing by the number of values that existed in either the pre-edit or post-edit dataset.

A potential problem with such a measure is that it tends to attribute lower data quality to complex cases, for which respondents have more opportunities to mis-categorize particular aspects of their wealth, whether by reporting particular assets (or debts) in the wrong part of the survey, double-counting, or omitting assets during the interview.[3] For example, reporting a home equity line of credit during a question on home equity loans, or mixing personal and business assets instead of reporting them separately. Because complex cases tend to be for those who are highly educated and affluent, the result is that highly educated respondents appear to give lower quality data. This seems counterintuitive, however it might be possible. Still, this doesn't seem to be the intended meaning of data quality

3.2 Low-Certainty Edit Rate

To account for complexity, we propose a new data quality measurement, a refinement of the change rate that looks specifically at the type of change, which we call the low-certainty edit rate. This measure is defined as:

SUM(SHADOW-VAR >=13)/SUM(SHADOW-VAR >=2)

This low-certainty edit rate takes advantage of a natural division in the shadow variable codes. Codes 13 and higher require judgement on the part of the editor or imputation, and thus there is less certainty that the resulting values of the associated variables accurately reflect reality. Specifically, the proposed measure is the proportion of edits that are made with low-certainty. Unlike the change rate, this measure does not seem to be inherently driven by case complexity when its relationships to different demographics are examined.

4. Results

Our analysis focuses on how these data quality measures vary by household income, net worth, and a variety of demographic measures. We present a series of charts displaying

[3] These errors are often discovered by the interviewers over the course of the interview and noted in comments.

median Change Rates (orange lines, right axes) and Low-Certainty Edit Rates (blue bars, left axes) for sub-groups of SCF respondents.[4]

First, we investigate if data quality varies across the distribution of household wealth and income. The two data quality measures are displayed for five percentile groups of the household net worth distribution in Figure 1A, and by income in Figure 1B.

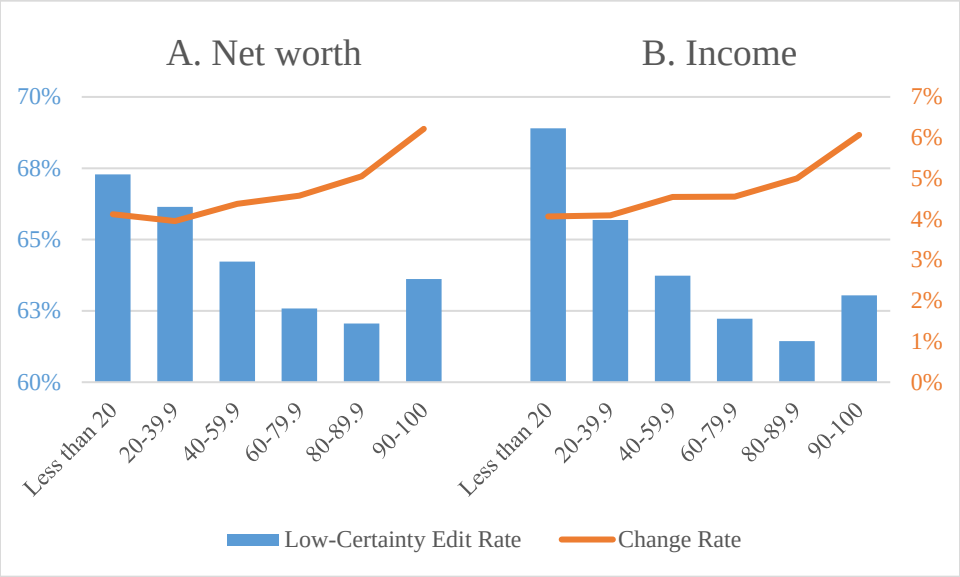

Figure 1: Low-Certainty Edit Rate and Change Rate by net worth and income percentile groups

The change rate exhibits a stark positive relationship with wealth and income. As net worth increases, the Change Rate increases, implying that data quality decreases. This is consistent with the fact that wealthier families have potentially more complex financial situations, which may require more editing as there is more opportunity for details to be captured in comments, or for assets or debts to be mis-categorized. Therefore, elevated change rates for wealthier families may not actually reflect lower-quality data reported by those respondents.

The Low-Certainty Edit rate, however, could circumvent this problem. This measure does not appear to be dominated by increasing case complexity for wealthy families. In fact, the low-certainty edit rate *declines* with net worth and income, reaching a minimum in the 75th to 89.9th net worth group, indicating those groups provided the highest quality data. This relationship is more expected and explicable given this demographic breakdown. Wealthier respondents could be expected to be more familiar with their finances and the financial concepts covered more generally, which would make it easier for them to provide higher quality data. Such arguments, however, do not explain the drop-off in data quality for the highest net worth group.

[4] We present medians due to the skewness in the distribution of change rates. Patterns by the various population cuts presented here are qualitatively similar to those using means.

A similar pattern is seen in income percentile groups, displayed in figure 1B. The Change Rate increases as income goes up, however the Low-Certainty Edit Rate goes down, again reaching a minimum in the second-highest income group.

We now proceed with an analysis of demographic characteristics, beginning with the role of the age of the respondent displayed in figure 2A. Once more, the Change Rate seems dominated by case complexity, with the 65-74 year old group having the highest change rate. This group, consisting of respondents that are about to retire or are recently retired, have had a much longer period of time to accumulate wealth in various assets, including retirement accounts, yielding a more complicated financial state than other age groups.

The Low-Certainty Edit Rate, on the other hand, shows that data quality is lowest in the youngest and oldest age groups.

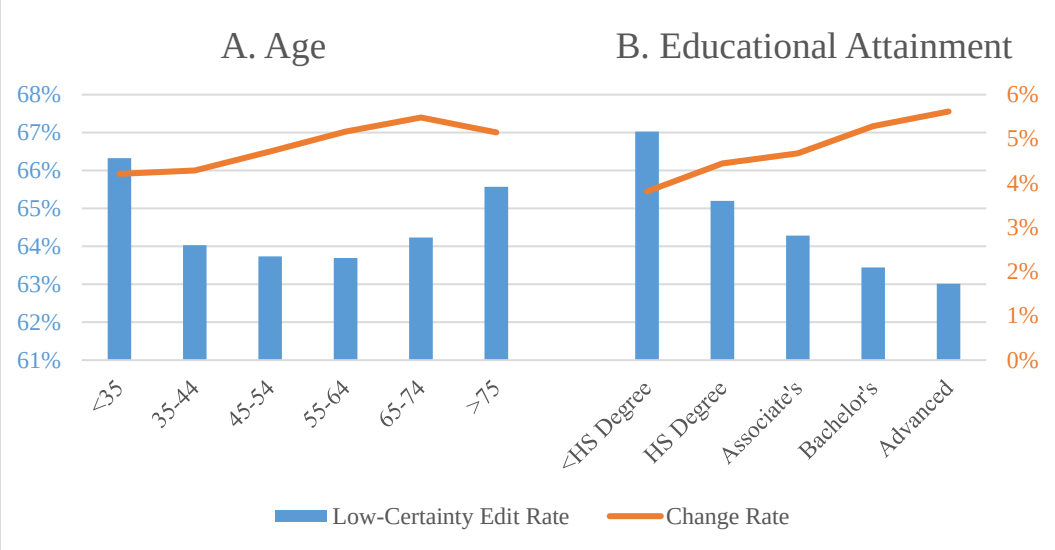

Figure 2: Low-Certainty Edit Rate and Change Rate by age group and highest educational degree attained

Figure 2B displays the data quality measures by education level, defined as the highest degree attained by the respondent. This breakdown highlights most clearly the discrepancy between the more traditional data quality measure and our Low-Certainty Edit Rate.

The Change Rate increases monotonically with education, implying that the more educated the respondent, the lower quality data they provide. Since higher education is also highly correlated with greater wealth and income, it again appears that greater case complexity is dominating this data quality measure.

The Low-Certainty Edit Rate, on the other hand, shows a more sensible relationship between data quality and education. More educated respondents, who we might expect to have greater comprehension of the questions, seem to give higher quality data.

Given these strong patterns by education, we now consider knowledge specific to the survey content, in the case of the SCF, financial literacy. The SCF gathers information on financial literacy in two ways. The first is by asking respondents a set of three objective questions designed to determine their understanding of stock investment, inflation, and

interest rates. The second is to ask respondents to rate their own level of personal financial knowledge on a scale of 0 to 10.

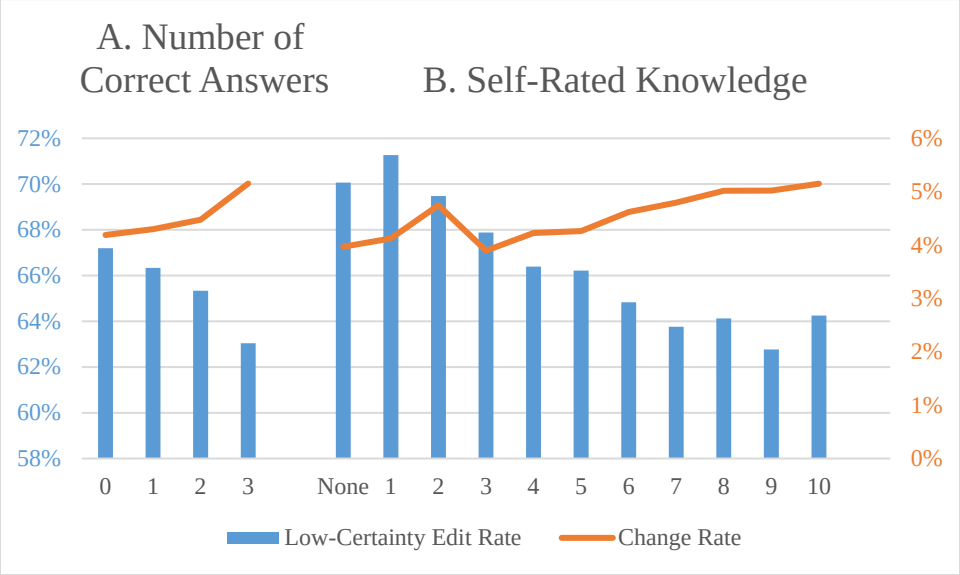

Figure 3: Low-Certainty Edit Rate and Change Rate by the number of financial literacy questions answered correctly out of three and the respondent's self-rated financial knowledge on a scale of 0 to 10

Results by the objective measure of financial literacy are shown in figure 3A. The two data quality measures show opposite relationships with financial literacy, consistent with the patterns seen when divided by education group. The most financially literate group had the highest Change Rate, but the lowest Low-Certainty Edit Rate. The self-rated financial knowledge measure (figure 3B), reinforces this finding. Those who rated their personal financial knowledge as a 1 had a Low-Certainty Edit Rate of 71%, whereas those who rated themselves as a 9 had Low-Certainty Edit Rate of 63%. Similar to the previous chart, the Change Rate has an almost perfectly opposite relationship, with it growing larger as the respondent's self-rated financial knowledge rating goes up.

5. Discussion

Traditional measures of data quality are useful, but in a household financial survey like the SCF, seem to be dominated by factors other than data quality. In particular, the Change Rate, or the percent of variables that are edited or imputed in the data, appears to more closely reflect case complexity. If we interpret the Change Rate as an indicator for data quality, we would infer that educated respondents provide lower quality data, while less educated people give higher quality data, a counterintuitive result. Respondents with a high Change Rate also belong to groups that tend to have more complex data—high income, wealthy, older households—which could lead to a more difficult interview overall.

However, by examining the rate of low-certainty edits, a measure that does not seem to be dominated by case complexity, more sensible and nuanced relationships become visible. Net worth and income groups near the top of the distribution, but not the very top, give the

highest quality data according to this measure. The youngest and the oldest give the worst quality data.

Most notably, the most educated and financially knowledgeable respondents tend to provide the highest quality data. This is consistent with the idea that respondents with domain-specific knowledge—in this case, financial literacy—will find the task of answering the survey to be easier and provide higher quality data. In particular, our results are consistent with satisficing theory, which predicts that respondents are less likely to provide merely "good enough" (or lower-quality) responses when they have higher ability (Krosnick 1991).

Future research into other demographic trends relating to data quality would certainly be useful, but there are also immediate applications. For example, given the relationship between data quality and self-rated personal financial knowledge, survey design could potentially be improved by asking a question on domain-specific knowledge near the beginning of the interview and instructing interviewers to provide extra assistance to respondents who rate themselves as having low knowledge. Such a policy could perhaps reduce the amount of low-certainty edits and increase overall data quality.

References

Bricker, Jesse, Lisa Dettling, Alice Henriques, Joanne Hsu, Lindsay Jacobs, Kevin B. Moore, Sarah Pack, John Sabelhaus, Jeffrey Thompson, and Richard Windle. 2017. "Changes in U.S. Family Finances from 2013 to 2016: Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin* 103(3).

Davern, Michael, Todd H. Rockwood, Randy Sherrod, Stephen Campbell. 2003. "Prepaid Monetary Incentives and Data Quality in Face-to-Face Interviews: Data from the 1996 Survey of Income and Program Participation Incentive Experiment." *Public Opinion Quarterly* 67(1): 139-147.

Fricker, Scott and Roger Tourangeau. 2010. "Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys." *Public Opinion Quarterly* 74(5): 934–955.

Krosnick, Jon A. 1991. Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5: 213–236.

Smyth, Jolene D., Don A. Dillman, Leah Melani Christian, Mallory Mcbride. 2009. "Open-Ended Questions in Web Surveys: Can Increasing the Size of Answer Boxes and Providing Extra Verbal Instructions Improve Response Quality?" *Public Opinion Quarterly* 73(2): 325-337

Wagner, James. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74(2): 223-243.