Identifying Data Problems and Improving Data Quality in the Survey of Consumer Finances

Catherine Haggerty, Micah Sjoblom and Steven Pedlow
NORC, University of Chicago, 55 East Monroe, Suite 2000, Chicago, IL 60603

Abstract

Since the Survey of Consumer Finances (SCF) delves into every financial detail of a household's finances, maximizing data quality is a constant challenge. Staff at NORC at the University of Chicago (NORC) and economists at the Board of Governors of the Federal Reserve (FRB) have collaborated in the design and implementation of a refined data quality process and review cycle. All interviewer commentary including call record entries, interviewer comments recorded during questionnaire administration and interviewer debriefing notes recorded post-interview are reviewed for potential quality issues related to respondent identification and questionnaire administration. Along with economists at the FRB, the NORC team evaluates the questionnaire data and identifies potential errors and anomalies that require follow-up. These evaluations are incorporated into a timely and systematic feedback loop delivered to interviewers. This feedback will trigger remedial actions designed to address quality deficiencies (e.g., broadcast memos regarding proper protocol, self-directed electronic training, and supervisor-led training) and provide an on-going assessment of interviewer performance. We will describe the processes used to identify data quality issues, our data quality improvement protocols, and data quality measures over time.

Introduction

Survey researchers work to ensure high data quality in a number of ways. First, researchers concern themselves with selecting a method of collecting data that makes sense for the topic and population of study and with regards to the ease or difficulty of understanding the concepts under investigation. They then turn their attention to crafting straightforward questions with simple language that will be easily understood by the

population of interest. The best practice in survey research is to pretest the questions to ensure they are understood and elicit the data they are designed to capture. Once the questionnaire is carefully constructed and tested, the next hurdle in data quality is fine tuning the means of collecting the data. In both self- and interviewer-administered data collection, it is essential that the paper or electronic instrument includes clear and simple instructions so as to facilitate the proper navigation throughout the instrument while strengthening the precision of collected data. Interviewers administering the questionnaire need to be adequately trained to have enough context to provide clarification when a respondent asks a question about a particular questionnaire item; they must grasp active listening skills to detect when the respondent does not understand the question; and they need to clarify responses through careful probe selection in an unbiased way so as to elicit a fitting answer to the question. Even with good question construction and well trained interviewers, post production examination of data often reveals some degree of problems.

Background

The SCF has concerned itself with data quality and engaged in detailed data quality activities since its inception. Early on, data quality activities consisted of the careful inspection of hardcopy questionnaires which included reading marginal notes that explained unusual questionnaire responses and family situations. These notes informed edits to the data which improved overall data quality. When the instrument was programmed and administered via Computer Assisted Personal Interviewing (CAPI) both interviewer comment boxes that could be accessed at any time during the interview and a debriefing section completed at the end of the interview were added to the electronic instrumentation to allow interviewers to record clarifying information just as they did previously within hardcopy questionnaire booklets.

While the SCF has always hosted an in-person training held over multiple days to prepare interviewers to collect details regarding family finances, incremental improvements and expansion of the training has become an integral part of our data quality program. The formal interviewer training program has evolved to include a pre-training project overview home-study exercise, a four-day in-person training program prepared for adult learners which includes interactive learning modules and simulated questionnaire practice. . Additional training activities follow the in-person training, including modules that: 1) reinforce fundamental interviewing skills interviewers must possess before they start work, 2) provide remedial exercises that target problems identified during case level questionnaire review, and 3) provide advanced learning opportunities prepared for interviewers demonstrating readiness to perform specialist activities.

The SCF has always validated a percentage of the interviewers' work. Over time inspection was expanded to include the close examination of paradata to detect potential data quality problems. These activities include questionnaire timing, missing data counts, and keystroke counts. Additionally, Benford's Law is applied to the data to help identify potential data falsification.

This paper describes the comprehensive data quality program which is comprised of interviewer training, examining and evaluating both paradata and questionnaire data immediately after it is collected, and the remediation activities resulting from problem identification.

Literature

As noted by Groves (1989), researchers and survey managers give much of their attention to interviewer training as a way to minimize interviewer error and variability from the beginning, with an ultimate goal of "nothing less than the elimination of the interviewer as a source of measurement error" (p. 358). The goal of training is to teach and provide practice in survey techniques and standard procedures so that there is consistency in the delivery of the survey, and variation in the data due to the interviewers is minimized (Lyberg and Kasprzyk, 1990, p. 252).

While competing theories exist on the most important factors for interviewer selection and the best questioning style for accuracy, certain findings about training length and content are generally accepted. Fowler (1991) makes the fairly intuitive statement that minimally trained interviewers will likely be unable to gather reliable data, while more fully trained interviewers (with at least two or three days spent in training) have the ability to gather data in a standardized way (p. 278). Billiet and Loosveldt (1988) found that relatively untrained interviewers had significantly higher non-response rates than trained interviewers on sensitive questions for which the study expected generally higher item non-response. However, Fowler and Mangione's (1983) important work makes clear that whether or not sufficiently trained interviewers collect data with the low measures of interviewer variability and bias they are capable of depends largely on supervision and monitoring.

As survey data collection modes have changed over the years, so have the methods and information available to evaluate interviewer performance. Traditional methods of interviewer monitoring, such as supervisors listening in on Computer Assisted Telephone Interviewing (CATI) surveys or accompanying in-person interviewers, can still provide a unique perspective on the interviewer's interaction with the respondent, and the relatively new technology of Computer Assisted Recorded Interviewing (CARI) achieves a similar effect. Re-interview and validation re-contacts remain useful ways to check interviewers' work. The biggest changes are seen in the growing use of computerized analysis of the data and information about its collection, called paradata.

Groves (1989) notes that interaction coding from interview tape-recording or monitoring can be put to a variety of uses. It can measure how often or to what extent the survey guidelines are being followed, a reflection of data quality. It can identify segments of the questionnaire that often prove difficult for interviewers or respondents. Finally, it can provide data on how individual interviewers are performing, which can allow for objective feedback and assistance (p. 389). Fowler and Mangione (1983) found that interviewers who tape-recorded their interviews and received feedback from the evaluation of a sample of these tended to have smaller interviewer effects than those who

were evaluated based on the apparent quality of their completed interviews or those who were evaluated based on such measures as costs per interview.

According to Forsman and Schreiner (1991), re-interview can be a useful tool both for estimating response error and for evaluating field work to reduce error. A variety of designs are possible, from simple validation re-contacts to full second interviews, after which the respondent may be asked to reconcile his or her response differences between the original and re-interview. The design will depend on the purpose of the re-interview; a few simple questions should suffice to detect the fabrication of an entire interview, while a much longer series of questions with a reconciliation process may be required to try to determine "true" values from which to calculate bias in the original data. An intermediate design including some subset of the original interview questions may be enough to approximate differences in interviewer effects during the field period, which can identify interviewers with higher variance for retraining or increased supervision. Another interesting possibility is to use other data, such as suspicious levels of vacancy or unemployed respondents, to determine which interviewers will become the focus of re-interview efforts.

Paradata can include interviewer characteristics, number of calls or time spent on a particular case, detailed records of keystrokes in computer-assisted surveys, and other measures related to the data collection process. Kennickell et al. (2009) considers paradata to be "of two types: (1) process data recorded as a by-product of the work to conduct a survey (e.g., call records, interview length, missing data codes, etc.) and (2) context data that are obtained separately or with a specifically targeted effort" (p. 1). Laflamme, Maydan, and Miller (2008) propose that paradata can be used to detect and correct data collection issues in a timely way, and to balance data quality with cost (p. 630).

Increasingly, paradata's potential for evaluating interviewer performance and data quality is being explored. Safir, Black, and Steinbach (2001), although their results were inconclusive, pursued an interesting direction in examining the relationship between interviewer characteristics such as experience and rates of gaining cooperation with item non-response rates and lower reporting of sensitive items. Mockovak and Powers (2008) used audit files, which recorded interviewer actions in the questionnaire instrument, to analyze training effectiveness, interviewer efficiency, and questionnaire design and usability (p. 1386). They were able to make conclusions about how interviewers were using the navigation and other tools in their instrument, and designed a refresher training to address inefficient habits.

Statistical analyses of the data themselves can also be fruitful for improving data quality during the collection period. A few studies have addressed using Benford's law, a pattern of first digit distribution that applies to many data sets, to detect unlikely or fraudulent survey data. Swanson, Cho, and Eltinge (2003) discuss using Benford's law in the Consumer Expenditure Survey to find possible instances when either the data was falsified by the interviewer or the respondent was guessing values. Cho, Eltinge, and

Swanson (2003) explore ways to use the results of Benford's law comparisons to better allocate resources for other data quality checks, such as re-interviews.

Groves (1989) points out that it is rare to find an error that all interviewers make or that any interviewer makes every time (p. 359). Dealing with such a variety of interviewers, respondents, and surveys, it is fortunate that there are so many ways to understand and improve data quality. It seems that the best approaches are combinations of techniques. For example, Fowler and Mangione's (1983) interviewers did best when they were both well-trained and tape-recorded, but relatively poorly with just the training or just the tape-recording. Swanson et al. (2003) and Wang and Pedlow (2005) caution that Benford's law can identify unusual data, but other tools must be used to determine why it is unusual.

Finally, Haggerty and Kennickell (2012) point to the importance of creating a culture of quality within survey teams which includes a strong collaboration between the survey organization and the funding organization. Bricker and Kennickell (2013) point to improvements achieved early in the field period of the 2013 SCF as a result of a comprehensive data quality program. From the methodologists who fine tune the survey questions to those who create and test the CAPI instruments, from the clerks who respond to requests for information to the interviewers who collect the data from respondents, from the field managers on the front line with interviewers to the Central Office staff who interact frequently with the client who funds the work and ultimately uses the data to inform policy, when all of the actors work together with the same well defined and shared vision of the purpose and means to achieve the goal improvements in quality can be attained.

Data Quality Methodology

The SCF uses numerous forms of data monitoring, evaluation and improvement over the course of data collection.

Data Monitoring. NORC uses both traditional and nontraditional forms of data monitoring; we employ the following data monitoring methods:

- *Validation*. At least ten percent of each interviewer's completed interviews are re-contacted to verify that the interview was conducted with the respondent. Several factual questions are asked to be sure the data were accurately recorded, and we confirm that an incentive was paid to those respondents flagged to receive an incentive.

- *Missing data counts.* Because there are many highly sensitive questions which require the capture of dollar amounts, interviewers need to constantly encourage and persuade respondents to provide answers to these questions. Cases with too many missing values are flagged for review and often do not contain enough data to be acceptable for FRB analysis.

- *Questionnaire timings.* Interviews taking fewer than 35 minutes to complete are considered executed too fast to allow the collection of high quality data; these

cases are closely examined, always validated, and sometimes deemed unacceptable by the FRB.

- *Character keystroke counts in selected questions.* The debriefing questions require the interviewer to clarify extraordinary responses, provide a synopsis of the family's finances, and to explain or clarify unusual family situations. If the keystroke count is low we suspect the explanation is inadequate and the case is flagged for a more thorough review.

- *Benford's Law.* Each week all questionnaire data is compared to Benford's Law, also called First-Digit Law, and refers to the frequency distribution of digits in many real-life sources of data. In this distribution, the number 1 occurs as the leading digit about 30% of the time, while larger numbers occur in that position less frequently: 9 as the first digit less than 5% of the time. Interviewers whose cases do not conform to Benford's Law are subject to close scrutiny so as to identify potential data falsification as early as possible.

Data Evaluation. Until this current round of data collection, economists at the FRB exclusively reviewed all questionnaire data and provided case level feedback to interviewers. The FRB continues to engage in a full review of the questionnaire data and all commentary associated with each case, but NORC currently engages in reviewing a subset of the data collected that has been triggered by flags using SAS code programmed by the FRB. Each week, the cases completed in the prior week are checked for 15 potential problems. Examples of the kinds of problems are interviewing the wrong household member, unknown homeowner status, incorrect dollar value(s), and households with zero income. The notes contained in the call records, interviewer commentary recorded inside the instrument and debriefing documentation are reviewed for explanations of the triggered flags. Often there are clear reasons for the unexpected data recorded but when an explanation does not exist we ask the interviewer if they recall any details regarding the unexpected responses. When we are unable to resolve the potential issues prior to data delivery, we notify the FRB to prioritize the review of cases with unexplained potential problems.

Data Improvement. Activities to improve data quality are varied and custom tailored to the experience and skill level of the interviewers. In the previous section, we described ongoing evaluation of questionnaire data at NORC and the FRB. This evaluation results in information regarding problems in the data that are both systemic and idiosyncratic; these problems are addressed in the following ways:

- *Weekly newsletters/memos.* Our weekly memos are limited to two pages and present three to four simple explanations of ways to deal with knotty questionnaire items, addresses a request for policy decisions related to unexpected situations encountered in the field, and explains protocols that may have been misunderstood. Exhibit 1 includes examples of a few memos. We utilize large font, bright colors, icons and cartoon-like characters to make the content attractive and interesting. We strive to keep the messages short and simple so that the newsletter will be read and the content memorable.

Exhibit 1: Memos via Moodle

Remedial training videos. All of the content delivered at our in-person training via video and PowerPoint is available to interviewers to re-visit throughout the field period. If the outcome of the NORC or FRB questionnaire case review indicates that an interviewer would benefit from retaking a particular module, the supervisor directs the interviewer to go to a module within a week of the notification. The following week the supervisor and interviewer discuss the module to be sure that the interviewer understood the points made and the intent of the lesson.

▪ *Common error videos.* Several dozen, short, self-directed, electronic training videos designed to address the most common data collection errors were produced prior to the start of data collection based on our repeated experience with the SCF. All interviewers are directed to view these videos, one per week,

starting at the beginning of data collection. The supervisor follows-up with each interviewer regarding the video lesson during weekly one-on-one meetings.

- ▪ *Advanced learning videos*. Several videos were produced aimed at interviewers ready for advanced learning to prepare them to handle collecting data from respondents with complex finances. The videos address known strategies for working with respondents to properly collect complex financial information within the existing CAPI instrumentation.

Exhibit 2: Video Lessons via Moodle

Exhibit 2 displays screenshots from our on-line learning center depicting the choice of videos and the short descriptions accompanying a video. You will also notice that each topic is accompanied by a quiz that interviewers can take to be sure they absorbed the key points of the lesson. Both the memos and videos are accessed via Moodle. Moodle, an open-source online classroom software used by

many colleges and universities allows for efficient development and delivery of online learning content.

- ▪ *Mentoring.* Remedial Training is also delivered through one-on-one meetings between the supervisor and interviewer held via telephone and outside the standard weekly meeting. These meetings are held with interviewers whose issues are typically idiosyncratic to them.

- ▪ *Coaching.* Small group meetings devoted to providing additional guidance to larger groups of interviewers experiencing difficulty in specific areas for improvement are held on an ad-hoc basis and attended by interviewers across regions. In addition to the supervisor, a senior interviewer may also be asked to join the meeting to discuss his or her own tips for dealing with the challenge being discussed.

Observations from Ongoing Data Monitoring and Evaluation Activities

In the methodology section of this paper we described several data monitoring activities we are engaged in. Here we share some of data from our data monitoring activities.

Benford's Law examines the distribution of first digits in any collection of numerical variables. Exhibit 3 shows the pattern that occurs naturally.
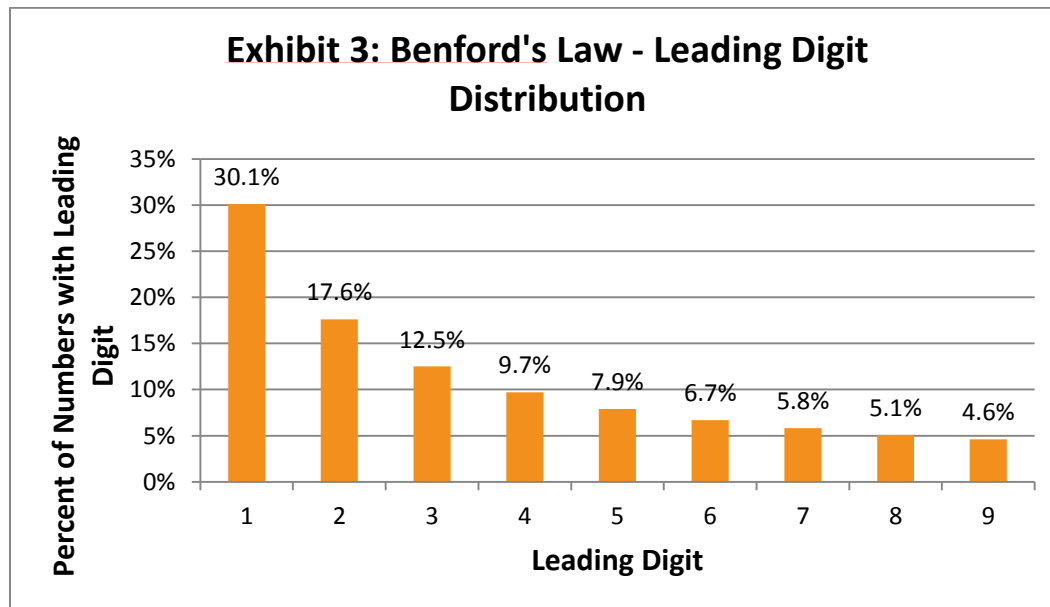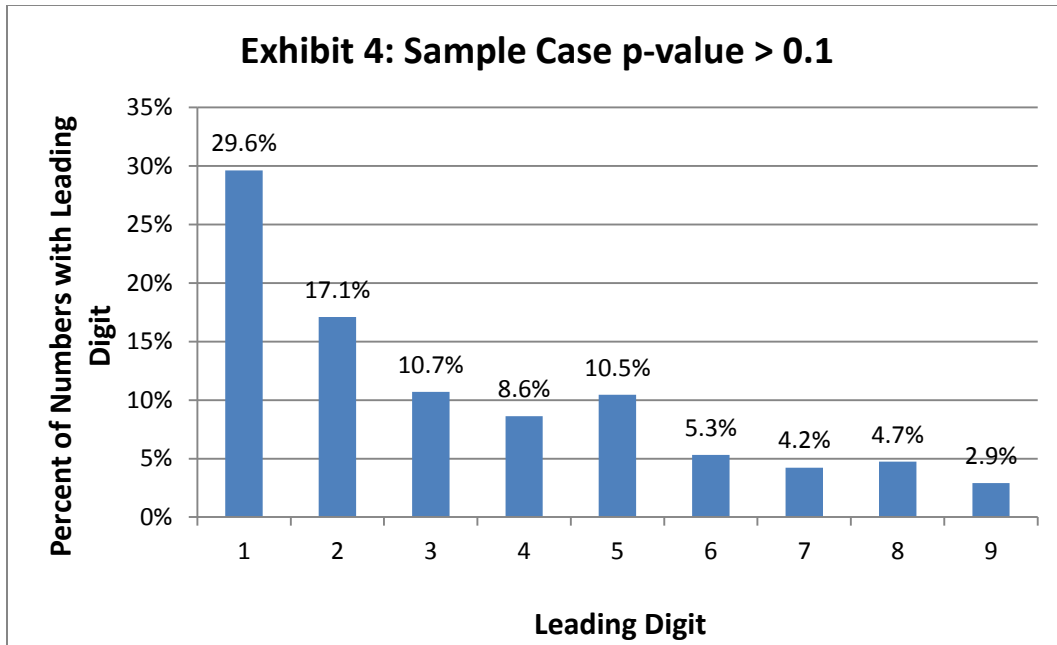

Exhibit 4 shows the pattern found in one particular case and which is typical in our survey. We often see additional "fives" due to rounding. Interviewers with cases with the lowest p values are flagged each week for examination.

Exhibit 4: Sample Case p-value > 0.1

Another data monitoring activity is the examination of the length of interviews. Exhibit 5 shows the elapsed time of interviews distinguishing those within and outside an acceptable range.

Exhibit 5: Elapsed Time per Interview in 1 Hour Increments: 0-5 Hour Range
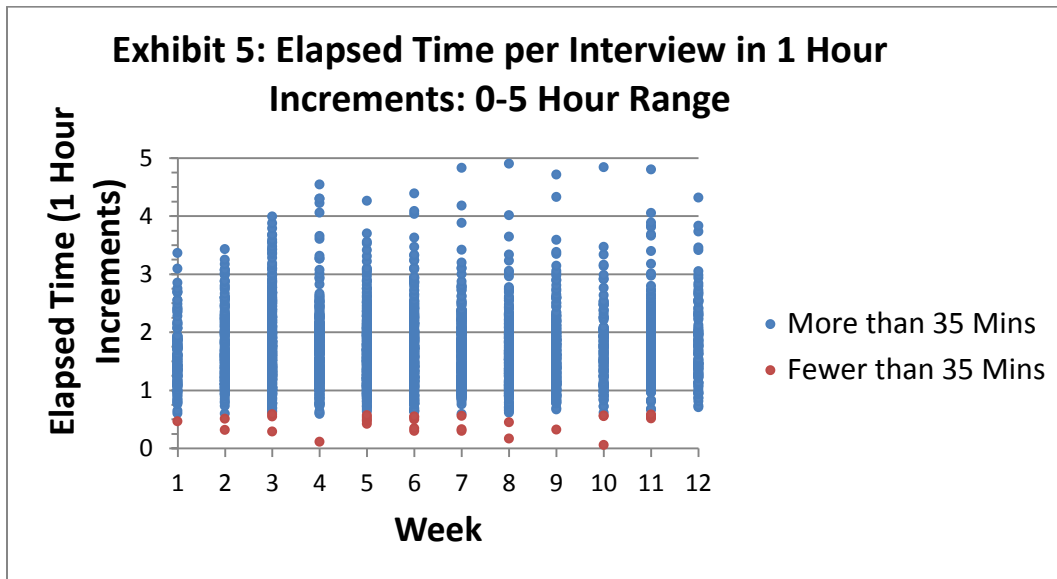
Also in the methodology section, we described data evaluation activities that included checking for potential problems in the data triggered by a set of flags. Almost half of the 3,250 (46%) cases completed through week 12 were flagged for one or more potential problems. Flags are constructed using 1 – 15 separate observations. Multiple flags are devoted to single issues and there are several flags that have not yet been triggered in the first fourteen weeks of data collection.

Through week 12 the problem flags, and frequency of the flags, are found in Table 1:

Table 1: Frequency of Problem Flags Through Week 12

| Description of Potential Problem	Frequency
Debriefing too short	1258
Edit check triggered	803
Institution not coded properly	428
Dollar value of 1	236
Potential wrong R (sum of 4 flags)	116
Business listed in section R but not in section F	106
Interview shorter than 35 minutes	32
Home ownership (5 flags)	23
10 flags	< 10 cases
7 flags	not triggered

Through week 12, our case review allowed us to resolve almost three quarters of the issues by either reading the commentary associated with the cases or by clarifying identified issues directly with the interviewers. However, we decided to select four potential problems to examine more closely including: 1) Dollar value of 1 (at least one of 15 questions has a value of "1" for this flag to be triggered); 2) Institution type not coded properly (this flag is triggered when the interviewer codes the recorded financial institution as an "other specify" instead of selecting the proper code); 3) Income is reported as zero (this flag is set when one of three questions in the income section indicates that the household income is zero); and 4) Wrong respondent was selected to complete the questionnaire. Exhibit 6 shows the frequency of these errors over the first twelve weeks of data collection.

Exhibit 6: Flagged Cases
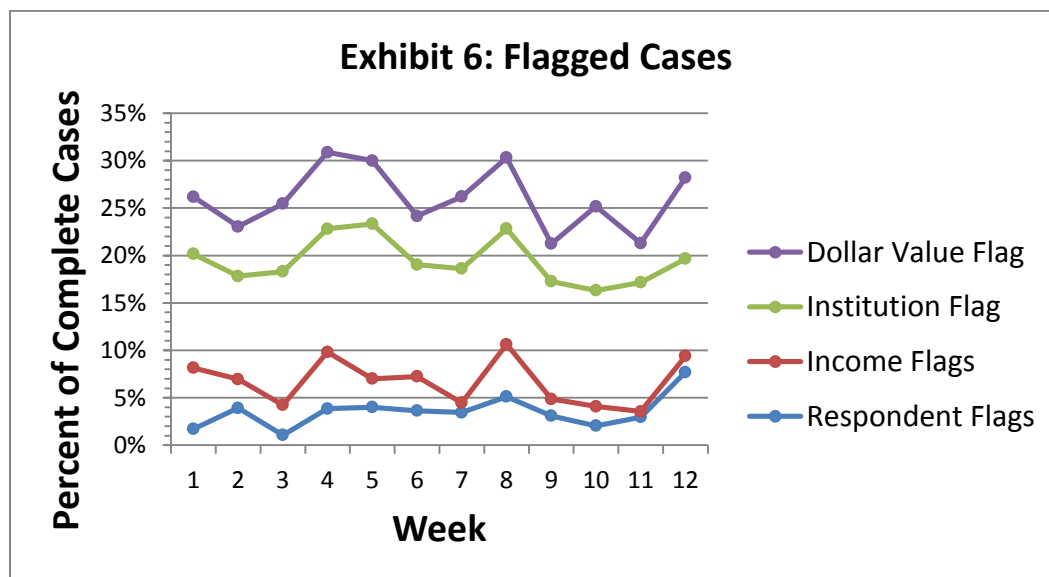
While the distribution of these errors seems to be similar over the first 12 weeks of our field period, we investigated variation by interviewer type: new hire, NORC experienced interviewer without previous SCF experience, NORC experienced interviewer with prior SCF experience and elite NORC interviewers with expert-level SCF questionnaire administration experience.

Exhibit 7: Income Flag & Interviewer Experience Type
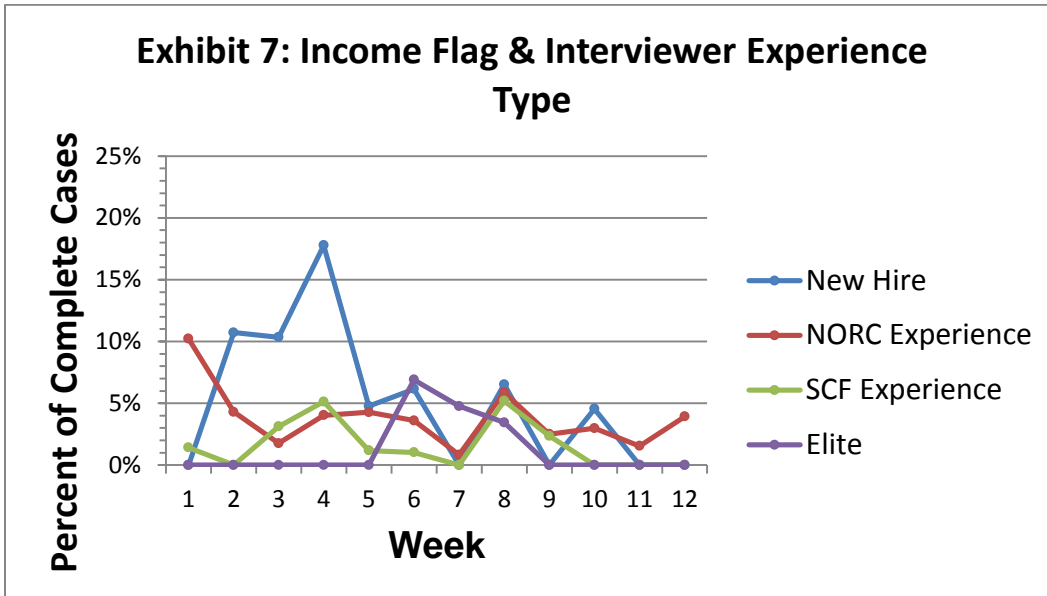
In Exhibit 7 we see that NORC experienced interviewers start out with the highest percent of income error flags. While the errors are much reduced, they still produce the highest percentage of errors weekly. We see the biggest improvement with the new hires. Our SCF experienced and elite interviewers have the least of these problems; this is what we would expect to see.

Exhibit 8: Respondent Flag & Interviewer Experience Type
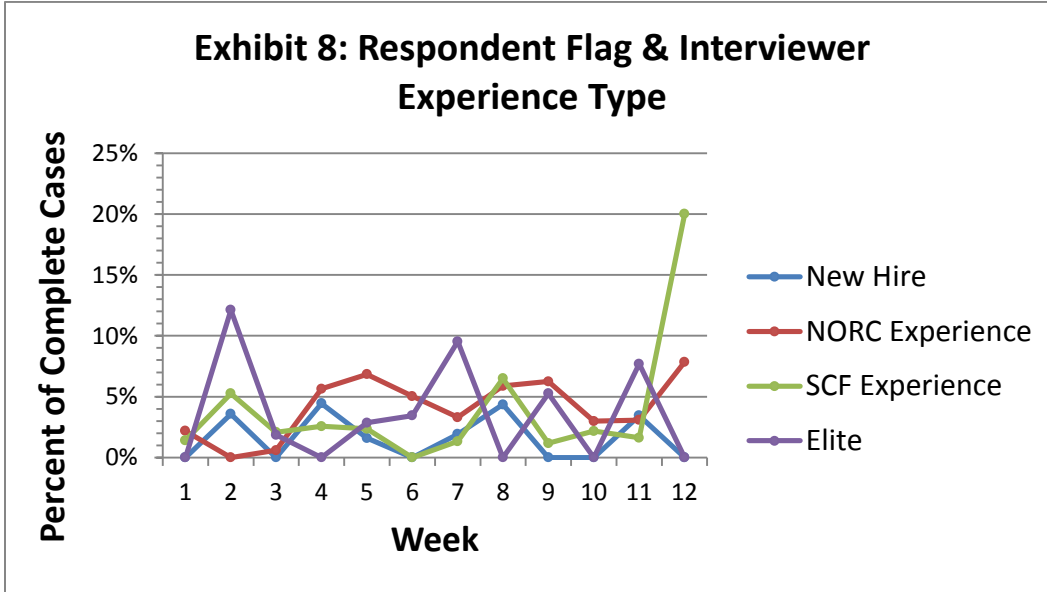
When looking at the distribution of errors for respondent selection across interviewer type in Exhibit 8, we were surprised: The new hires tripped this flag the least.

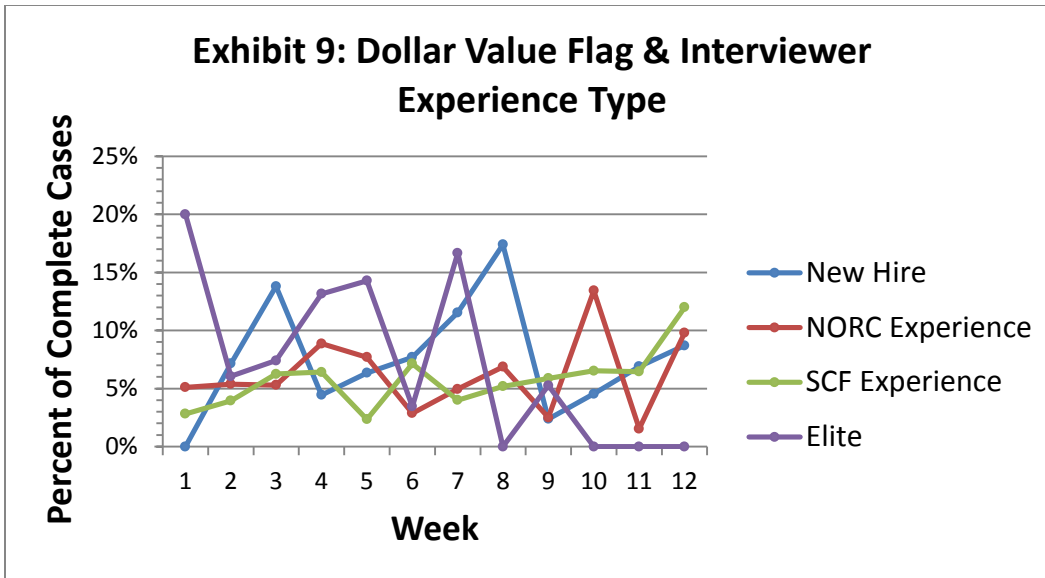**Exhibit 9: Dollar Value Flag & Interviewer Experience Type**

Exhibit 9 shows that the elite interviewers are at zero for the last two weeks but at week 12 the SCF experienced, NORC experienced, and new hires are all still about at the same rate of errors.

Exhibit 10: Institution Flag & Interviewer Experience Type
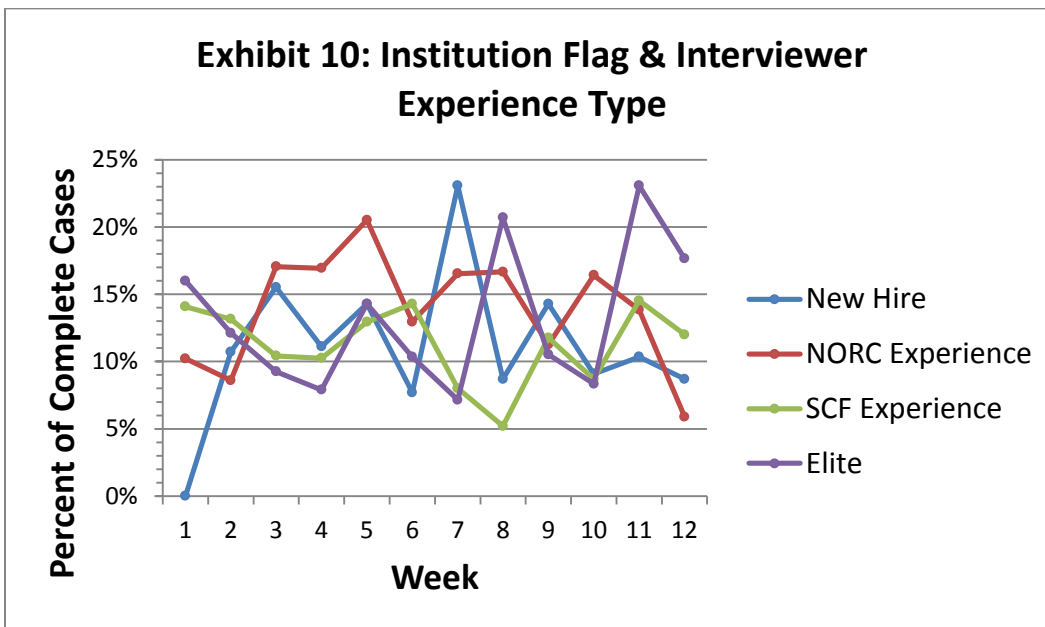
While there is a lot of movement across all the interviewer groups in Exhibit 10 we see that improperly coding institutions is more of a problem for the elite, NORC and SCF experienced interviewers than it is for the new hires.

Finally, in the methodology section we described data improvement activities. These activities included remedial, new and advanced learning lessons. By week 14, we had

released 14 Field Interviewer Memos. These memos included 36 substantive articles/lessons. The articles fall into the following three categories:

- New Instruction. The information in the article is likely unfamiliar to most FIs as they were not introduced to it at training. Example: *'No one knows it all' – breakdown of what List Sample information comes from which source and who has access.* Nine articles/lessons fall into this category.

- Extended Training. A concept introduced at training is now fleshed out further and given context from the data we have seen this round. Example: *'Confidentiality Concerns' – reminder about not using identifying information in the call records.* Twenty-three of these articles/lessons fall into this category.

- Project Updates. Non-instructional updates from the project meant to inform the FIs on matters regarding the SCF 2013. Example: *'Read all about it' – 2013 SCF Press Release by the FRB and its link.* Four articles/lessons fall into this category.

Discussion and Recommendations

From the design, development, and pretesting of the questionnaire, throughout interviewer training and data collection, and into post-production analysis, there are always opportunities to reduce and understand shortcomings in data quality. We seek to prevent errors by testing the questionnaire for ease of use and understanding and by training interviewers thoroughly. We detect potential problems by closely examining the data and by increasing the body of information available surrounding it. We remedy issues by giving feedback to field staff.

This paper focused on both the examination of data throughout the field period while there was still time to take corrective action, and an on-going continuous learning program designed to reduce errors. It should be a fundamental requirement of all surveys to carry forward what is learned about patterns of error and to seek effective preventive and remedial strategies so that future surveys are improved upon. Quality protocols should also be an important ingredient in all surveys as they are an essential part of a process of continual improvement.

Bibliography

Billiet, J. and Loosveldt, G. (1988). Improvement of the quality of responses to factual survey questions by interviewer training. *Public Opinion Quarterly,* 52, 190-211. Retrieved from http://poq.oxfordjournals.org/

Bricker, J. and A. Kennickell. (2013) "Shared Understanding and Data Quality in the SCF," *JSM Proceedings*, Government Statistics Section, Alexandria, VA: American Statistical Association (forthcoming, 2013).

Cho, M., Eltinge, J., & Swanson, D. (2003). Inferential methods to identify possible interviewer fraud using leading digit preference patterns and design effect matrices. In 2003 Proceedings of the American Statistical Association.

Forsman, G. and Schreiner, I. (1991). The design and analysis of re-interview: An overview. In P. B. Biemer, R. M. Groves, L. E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.), Measurement errors in Surveys (pp. 279-301). New York, NY: Wiley.

Fowler Jr, F. J., & Mangione, T. W. (1983). The role of interviewer training and supervision in reducing interviewer effects on survey data. In 1983 Proceedings of the American Statistical Association.

Fowler Jr, F. J. (1991). Reducing interviewer-related error through interviewer training, supervision, and other means. In P. B. Biemer, R. M. Groves, L. E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.), Measurement errors in Surveys (pp. 259-278). New York, NY: Wiley.

Groves, R. M. (1989). Survey Errors and Survey Costs. New York, NY: Wiley.

Haggerty, C. and A. Kennickell. (2012). The Survey of Consumer Finances: Creating a Consistent Culture of Quality. In H2R Proceedings, Techniques and Methodologies Section. Alexandria, VA: American Statistical Association.

Kennickell, A., Mulrow, E., & Scheuren, F. (2009, November). Paradata or process modeling for inference. In Conference on Modernization of Statistics Production, Stockholm, Sweden.

Laflamme, F., Maydan, M., & Miller, A. (2008). Using paradata to actively manage data collection survey process. In Proceedings of the Survey Research Methods Section, American Statistical Association (pp. 630-637).

Lyberg, L. and Kasprzyk, D. (1991). Data collection methods and measurement error: An overview. In P. B. Biemer, R. M. Groves, L. E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.), Measurement errors in Surveys (pp. 237-257). New York, NY: Wiley.

Mockovak, W., & Powers, R. (2008). The use of paradata for evaluating interviewer training and performance. In Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Safir, A., Black, T., & Steinbach, R. (2001, August). Using paradata to examine the effects of interviewer characteristics on survey response and data quality. In Proceedings of the Annual Meeting of the American Statistical Association.

Swanson, D., Cho, M., & Eltinge, J. (2003). Detecting possibly fraudulent or error-prone survey data using Benford's Law. In Proceedings of the American Statistical Association (pp. 4172-4177).

Wang, Y., and S. Pedlow (2005). Detecting Falsified Cases in SCF 2004 Using Benford's Law. 2005 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, pp. 3652-3657.