sciendo

Journal of Official Statistics, Vol. 35, No. 2, 2019, pp. 337–352, http://dx.doi.org/10.2478/JOS-2019-0016

Enhancing Survey Quality: Continuous Data Processing Systems

Karl Dinkelmann[1], *Peter Granda*[1], *and Michael Shove*[1]

Producers of large government-sponsored surveys regularly use Computer-Assisted Interviewing (CAI) software to design data collection instruments, monitor fieldwork operations, and evaluate data quality. When used in conjunction with responsive survey designs, last-minute modifications to problems in the field are quickly addressed. Complementing this strategy, but little discussed, is the need to implement similar changes in the post data collection stage of the survey data life cycle. We describe a continuous data processing system where completed interviews are carefully examined as soon as they are collected; editing, recode, and imputation programs are applied using CAI tools; and the results are reviewed to correct problematic cases. The goal: provide higher quality data and shorten the time between the conclusion of data collection and the appearance of public use data files.

Key words: Data quality; curation; tools; dissemination.

1. Introduction

Many survey research projects depend heavily on Computer-Assisted Interviewing (CAI) to program the design of data collection instruments, improve error checking, closely monitor fieldwork operations to counter nonresponse, increase response rates, and evaluate completed cases almost immediately after they are collected. More recently, several commentators have focused on post data collection issues, particularly with correcting nonsampling errors as an essential component to improve overall survey quality (De Waal 2013; Thalji et al. 2013). Even before CAI became a standard method of conducting many large national and cross-national surveys, the connections between data collection and data processing had become more collaborative (Biemer 2010). Principal investigators have a great incentive to process and analyze their data as quickly as possible in order to publish their results and to meet data sharing requirements now demanded by many funding agencies.

CAI added a very powerful dimension to this connection. It permitted storage of the variable-level metadata: variable names and labels, question text, universe statements, interviewer instructions, missing data definitions, and so on within the actual data collection instrument. Although not an early priority, CAI systems could repurpose this metadata for such things as public use documentation or to reuse the material when creating project reports.

Certainly, one of the main features of CAI systems is to perform data checking during the interview process itself. The programming logic built into the survey instrument

[1] University of Michigan, USA, Institute for Social Research, 330 Packard Street, Ann Arbor, MI, 48104, U.S.A. Emails: karldi@umich.edu, peterg@umich.edu, and mshove@umich.edu

prevents impossible or improbable responses and often carefully controls acceptable answers for demographic questions. For many years, national statistical agencies have developed internal controls to monitor and edit incoming data to standardize workflows and improve data quality (Bethlehem 1997). However, certain types of complex surveys, such as ones that collect family histories and have lengthy questionnaires, which severely test respondent recall, present significant challenges for any automated checking system. Respondents can easily misstate or fail to remember the dates of important events that may become evident only when the entire interview is completed. Data producers must also balance the quest for accuracy with the need to complete interviews within available budgets. Surveys with these characteristics often require considerable checking and editing after the data collection period ends.

Under such conditions, we suggest treating the post-data collection process in the same way as we now treat the planning and conduct of field operations. This article proposes a "*continuous data processing system*" to routinely evaluate inconsistent or illogical responses and make appropriate corrections. The model described below does not require new tools or systems, but uses the features of the original CAI data collection program to perform automated data checking, cleaning, and processing tasks at the same time that interviews are completed in the field.

The initial implementation of this system grew from data processing tasks connected with producing public use files for the National Survey of Family Growth (NSFG), a nationally representative survey conducted by the National Center for Health Statistics (NCHS) in the United States that gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health (http://www.cdc.gov/nchs/nsfg.htm). The NSFG uses CAI systems for all aspects of data collection and the transmission of completed interviews to a central project database. Internal communications protocols review incoming cases from the field daily to verify that each case has sufficient information to qualify as "completed" based on agreed project parameters. After verification, completed interviews are ready for the checking and editing operations.

Our goals for testing the system with the NSFG included addressing the following questions: (1) whether or not it was feasible to review completed records immediately after they were collected in the field; (2) was the new CAI programming application successful in correcting all responses affected by changes in the values of erroneous entries; and (3) how much time and effort would the implementation of this system save for data producers.

We begin by describing why continuous data processing systems are a useful tool for complex surveys, provide a comprehensive description of the system used for the NSFG, how successful we believe the system worked in this initial application, and finish with an assessment of the many data quality implications such systems may provide in the production of public use data files and documentation.

2. Continuous Data Processing Systems

2.1. Why is it Necessary to Change Post-data Collection Procedures?

While it is true that CAI software facilitates the collection and checking of data in the field, it is also the case that CAI programmed instruments are focused on completing interviews

as quickly as possible. Transforming raw data elements into public use files that secondary analysts can use effectively often requires several processing steps that may include:

- Preliminary consistency checking of completed interviews,
- Creation of new recoded/derived variables to facilitate analytic use,
- Imputation of item missing values,
- Generation of several types of weights dependent on the survey population,
- Variance estimation,
- Disclosure review to decide which variables are appropriate for public release,
- Changes to the data (e.g., top and bottom coding, swapping, perturbation) to further protect respondent confidentiality, and
- Creation of extensive documentation on the entire data life cycle to facilitate use by secondary analysts.

These processing steps are often lengthy and time-consuming because of the complexity of CAI-generated interviews. However, the CAI software makes it possible to create a systematic approach to a continuous data processing design that can contribute significantly to expediting processing tasks and satisfying the needs of funders, data producers, and interested researchers at the same time.

A continuous data processing system would perform many of the tasks listed above on a regular schedule so that interviewing and processing occur almost simultaneously. Such a system could conceivably edit and check cases immediately after completion, create recodes (i.e., derived variables calculated from raw data variables) and sampling error codes, calculate weights, and build the basic documentation files. Some tasks, such as imputation and disclosure review would take place only after enough data was collected to permit secondary analyses.

The creation and success of any continuous data processing system would depend upon close collaboration between the collector of the data and those who produce the public use or analytic data files. This collaborative effort must adhere to one of the basic principles about case editing: disturb the original data as little as possible.

One of the earliest attempts to set rules for case editing and apply them to a system for survey data appeared in a seminal article by I.P. Fellegi and D. Holt entitled "A Systematic Approach to Automatic Edit and Imputation" (Fellegi and Holt 1976). Their objective was to design an automated procedure for editing and imputing data that would alter the fewest possible values, maintain the frequency structure of the data file, and derive imputation rules directly from the editing rules. Believing that designing separate computer programs to edit and correct records would be costly and error-prone (perhaps as true today as it was in 1976!) they suggested an approach based on simple, logical rules created by subject matter experts.

In theory, it is now the case, some 40 years later, that the advent and continuous development of CAI software has made it possible to avoid a large amount of post-data collection processing and systematically improve data quality by simplifying data capture and editing tasks. This becomes possible when the CAI software encompasses both data collection and data cleaning operations.

Edit checks are routinely built directly into the software to reduce interviewer entry errors and to require respondents to rethink and correct erroneous or questionable answers. Common patterns of quality checking have emerged with these CAI systems. Completed

interviews are sent from the field to the coordinating center or survey headquarters on a daily basis. Each interview undergoes some type of automated review, and, if problems arise, interviewers and survey managers are in immediate communication to resolve them. Programmers can produce tables that check the values of key indicators in the survey.

Another method used in CAI programming to check recorded values is to create 'computed' variables (essentially recodes built right into the CAI programming structure) based on responses to original questions which can be transferred into final output files and serve as summary variables, saving time and effort in secondary analyses. For example, respondents may answer a series of questions about their race and/or ancestry that would then be condensed into a single 'computed' variable, which is stored and subsequently transferred to the output data file.

These CAI programming structures are especially valuable when a survey collects extensive respondent and family histories regarding work patterns, educational attainments, family formation, and health issues over extended time periods. In such surveys when reporting key family events, interviewers can expect that specific dates might not always be accurate, particularly when the event occurred many years before the date of the interview. Immediate checking of anomalous dates can often be incorporated into the CAI software programs through "hard" edit checks that force interviewers to review problematic or impossible responses with the respondent and correct the information before completing the remainder of the interview. However, survey designers also consider keeping such "hard" edit checks to a minimum so as not to increase the time it takes to collect the interview, cause a refusal, or increase respondent burden. The tradeoff often involves using "soft" edit checks that permit the interviewer to review a particular response but move on to other questions if the respondent does not provide adequate clarification. (Soft checks are also used when a given response is unlikely/improbable yet could still be possible).

Recode programs provide a third opportunity to check possible reporting errors. Post processing recodes can either use raw and/or CAI-generated 'computed' variables in their creation. Once the actual code for generating these post-collection recode variables is complete and thoroughly checked, any cases not meeting the specified conditions may indicate some discrepancy with the data as it was originally collected.

However, the costs of all of these computer-assisted checks may be considerable in both programming effort (as well as testing that they all work correctly) and in the extensive subsequent reviews necessary to ascertain the nature and extent of the problems that they might uncover. Testing of such programs can begin when sample cases are input into the CAI program or if a formal pretest is part of the data collection process. Even with such rigorous testing, it is not always possible to collect a broad enough range of responses to guarantee that the CAI programs are error-free. Having respondents recall events, which happened many years earlier, may present formidable obstacles to ascertain the validity of CAI checking programs.

2.2. *What Steps are Necessary to Implement a Continuous Data Processing System and What Implications Would it Have on Data Quality and Data Dissemination?*

This approach, involving both human and machine interaction, permits completed interviews to be carefully examined as soon as they are collected; identifies problematic

cases; determines resolutions; and, most importantly, *applies data edits directly in the CAI software*. As described below, these data editing and cleaning operations, because they work in close connection with the logic and rules programmed into the CAI instrument, will reduce errors and improve the efficiency of subsequent programs to create derived variables and imputed values, particularly with regard to correcting erroneous date and time values.

The NSFG consists of two main data collection applications, one for women and another for men. The female instrument has more than 8,200 internal consistency checks, programmed within the application to assist the interviewer with inconsistencies found during the course of the interview. The male instrument has more than 3,700. Routing for both instruments is highly dependent on respondents' reporting of events over time. Time and date calculations made within the CAI application use the system time of the data collection laptop during the interview. These date calculations are then used with programmed consistency checks to create new variables throughout the instruments. To facilitate working with data coded in months and years within the application, the concept of the "century month codes" was used. A century month is based on a coding system where the value of 1 is assigned to the month of January 1900 and increments by one for each succeeding month. The following formula translates actual months and years into century months:

$$\text{Century Month} = 12\ (\text{Year} - 1900) + \text{Month}$$

For example, February 2018 would equate to century month 1417.

Accurate reporting of events is necessary for proper routing though the instruments. In addition to internal consistency checks, the female instrument attempts to assist respondents by using a life history calendar to anchor key events to aid in the recall of dates of pregnancies and contraceptive usage to answer them more accurately.

However, despite having more than 11,000 internal consistency checks to aid the interviewer and a life history calendar to assist the respondent with capturing dates correctly, errors happen. To allow us to apply edits to the instrument after data collection finished, we had to turn off the dynamic nature of looking at the computer's system date. This was done by adding additional code to the date processing portion of the CAI system logic to ensure date calculations during the data editing process would be based on the date the interview was completed (instead of dynamically looking at the computer's system date). This allows us to programmatically apply edits to the survey data and systematically reprocess the rules of the instrument. When these edits are applied, it forces downstream rules within the CAI application to update any other areas that would be involved within a given edit. This can sometimes result in 20 or more constructed variables updated from one variable edit applied.

3. System Implementation

The overall model proposed for continuous data processing is illustrated in Figure 1.

The development of a practical continuous processing system should commence even before the start of data collection. Principal investigators often hire survey organizations to collect, clean, and process data for them. As questionnaire specifications are prepared

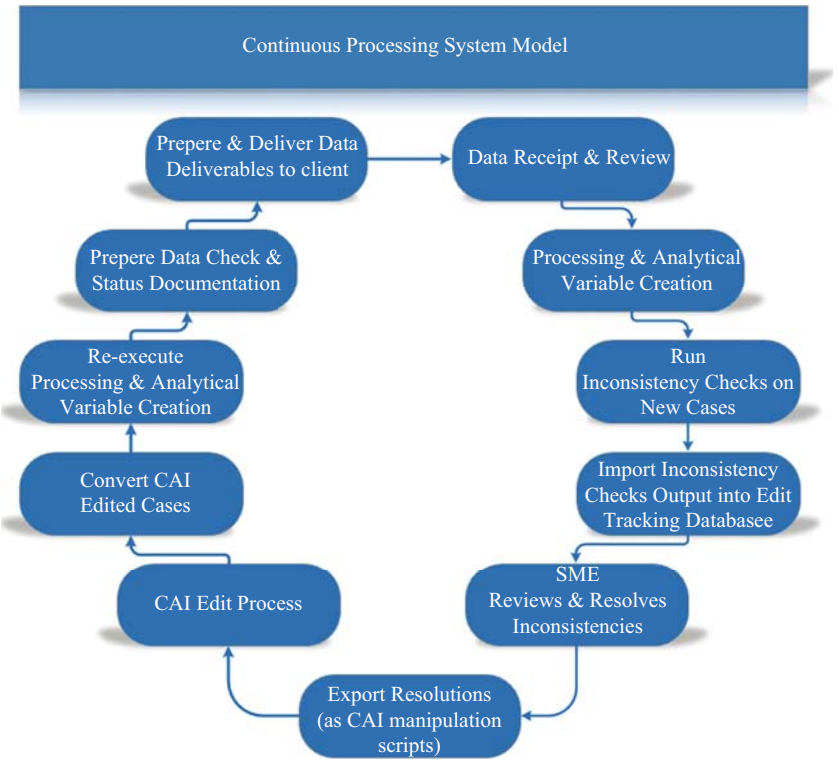

Fig. 1. Continuous processing system model.

and tested, the two key players who design and conduct the survey should meet and decide which types of quality checks they will perform on an initial set of cases. Ideally, these decisions would be in place before any data is collected. Data processing teams could conceivably modify quality checking routines as completed interviews arrive back from the field and they learn which sections of the questionnaire require enhanced review. After a relatively short period, both researchers and data processors would finalize quality checking procedures and methods for dealing with any unexpected anomalies.

At the same time, the data processing team would write and test the post-processing recode programs on this same set of initial cases. The research team would view the results and suggest alterations to recode specifications if additional conditions needed to be included in the programs to cover unexpected reporting situations. A set of system processing rules would follow this procedure. The goal of this initial, potentially intense set of interactions between researchers and data processors would be to integrate both data editing and recoding into a single system that would operate automatically as each new batch of cases arrived in the coordinating or data center. After a time, researchers would only need to review those cases that did not fit the set of agreed upon rules that they had created with the data processors.

Once these preliminary steps are completed and rules and procedures established, the continuous data processing system would operate in production mode with real cases from

Fig. 2. Edit tracking database template.

the field. Checks are then done on an ongoing basis when cases can be evaluated soon after their collection providing the best opportunity for evaluation and resolution.

The heart of this continuous data processing system is the CAI instrument itself. It is the foundation upon which the data editing process is built and consists of the following elements, steps, and procedures that are integrated within the CAI environment.

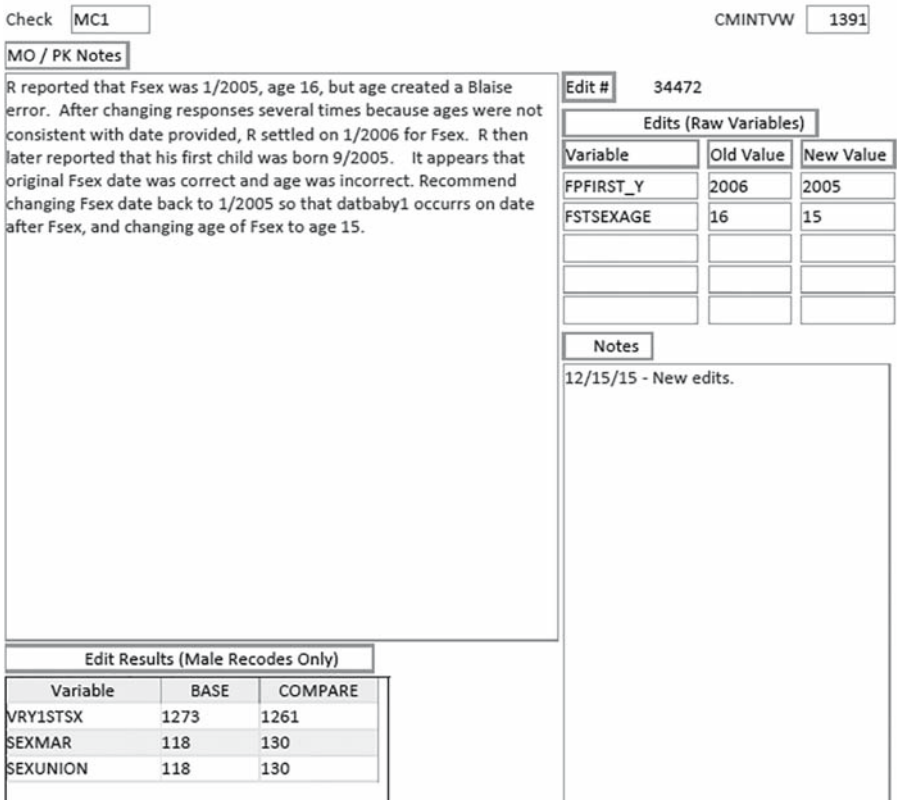

Fig. 3. Subject-matter Expert editing recommendations.

Data Edit Step		Resulting Outcome
1.	Isolate problem cases.	Database with cases to edit.
2.	Apply data edit(s) & re-execute the CAI software rules.	Database with edits applied; with off-route data removed & updated constructed variables.
3.	Read the cases from step 2 to determine newly on-route but empty items.	List of variables placed on-route and are empty.

Fig. 4. Editing steps and outcomes.

A separate "Edit Tracking Database" exists to both review and resolve inconsistencies in overall record logic or individual variables. Subject-Matter Experts (SME) examine problematic cases using both available data and paradata including interviewer comments, case notes, the review of specific interviews through an instrument keystroke playback, and, in some cases, re-contacting the interviewer as quickly as possible for additional information. Solutions are captured in the Edit Tracking Database using a series of forms. Figure 2 shows an example of one of these forms with some of the values for certain variables expressed in century months as described earlier.

Figure 3 shows an interview flagged for editing, the recommendation for editing, and the pre-editing and post-editing values for both raw and computed/recoded variables. In this case, the respondent reported inconsistent information about month of first sex and

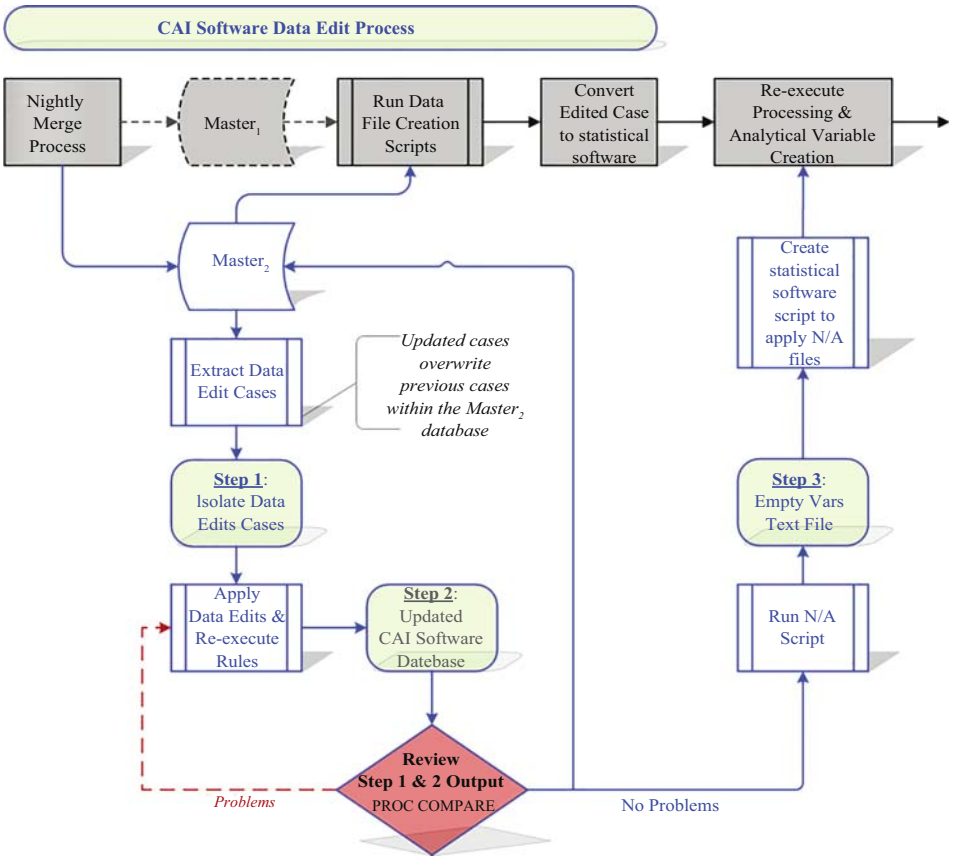

Fig. 5. Capturing the entire process.

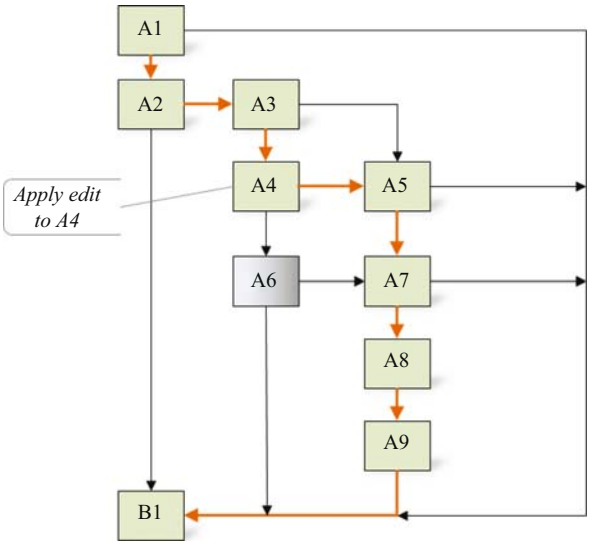

Fig. 6. Sample case question flow.

month of birth of first child. After review by a SME, the date of first sex was corrected as indicated in Figure 3.

The Edit Tracking Database is then queried to dynamically create a series of scripts used in the editing process within the CAI software which not only corrects the original inconsistency, but also values for all other variables affected by the change as noted in Figure 4.

The entire process can be diagrammed as shown in Figure 5.

The 'Nightly Merge Process' captures all cases that interviewers completed that day. These cases are exported to two identical data files: Master_1 and Master_2. Cases identified for review because of the editing checks are extracted and placed into a separate file for adjudication (Step 1). Project staff reviews each case, dynamically exports edit scripts, applies the necessary edits, and re-executes the CAI software program rules in order that the logical flow of the questionnaire is maintained. Re-executing the rules of the CAI

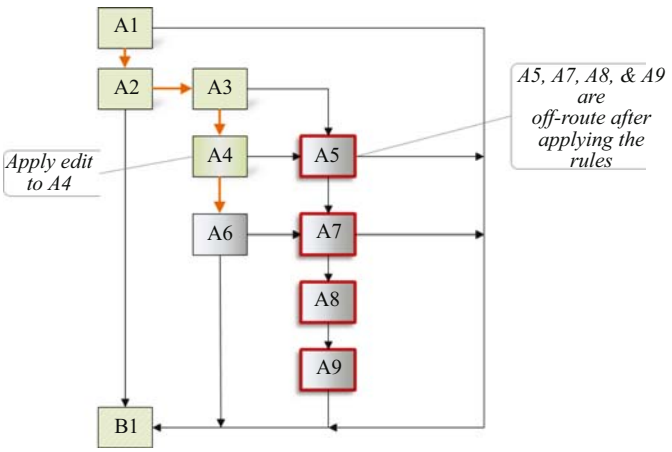

Fig. 7. Sample case rerouting.

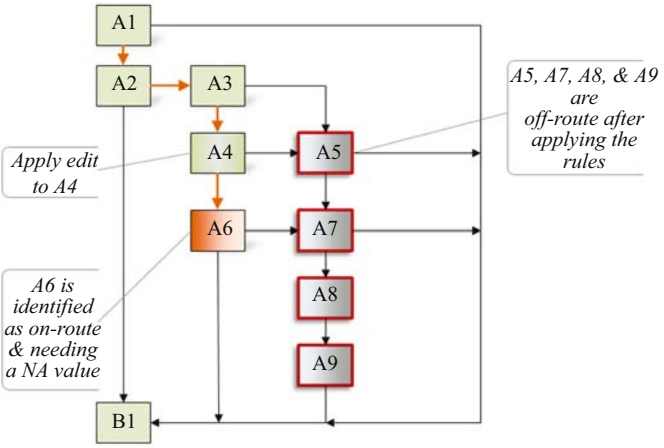

Fig. 8. Sample case editing.

software is done programmatically within the edit scripts. This insures unnecessary calculations or questions that have become off-route are removed and downstream calculations and/or questions that have become on-route are recalculated or identified as missing variables and assigned "not ascertained" (shown in Figure 5 as "N/A") in the process. For example, let us examine the case when a respondent originally answered affirmatively to Question A4. She is then routed to Question A5 and is not asked Question A6, as shown in Figure 6.

If the review process determines that she intended to answer Question A4 negatively, she would be routed to Question A6 instead. Because of this change, the CAI software programming rules would put A6 on-route while A5 and A7 to A9 would be placed off-route. This is illustrated in Figure 7.

The edit would change the value of A4 from affirmative to negative, alter the value of A6 to "not ascertained" since it is now on-route but has no value since it was not asked during the interview, and change the values for A5, A7-A9 to missing since these questions are now off-route.

This procedure is captured in Steps 2 and 3 of Figure 5. Step 2 could be repeated if the review of the new edits indicated a mistake in the code correcting any original values.

Once the review steps are completed, the edited cases are copied back into the Master_2 file and subsequently output from the CAI software and into an ASCII data file with accompanying syntax files that will read the data in such proprietary statistical software packages as SAS and SPSS. It is important to note that the Master_1 file is untouched in this process. It continually collects all of the *original* raw data from the field. This permits data managers to refer back to the original data whenever necessary, should questions arise later about any of the cases that have been edited in Master_2.

Storage of multiple databases should not be a problem for most surveys. While both the Master_1 and Master_2 files are updated at regular intervals, they are cumulative. Earlier versions do not require permanent backup and can be deleted. Once data collection is complete, one copy of the Master_1 data file and one copy of the Master_2 data file will be permanently archived.

4. Data Quality Implications

This editing system will enhance data quality in several ways:

4.1. Enable More Rapid Corrections in the Survey Instrument

Performing checks as cases come in from the field will not only catch potential errors in data collection but might also uncover potential errors in the instrument itself. A consistent pattern of questionable or erroneous values for a particular variable could indicate an error in programming or routing. Correcting such errors as early as possible will minimize the number of cases that must be adjudicated when the final data files are constructed. This illustrates how having a continuous data processing system as part of the normal workflow of a project can also improve data collection activities as such processing checks can actually affect how the instrument is implemented in the field.

4.2. Provide More Consistent Responses in Cases Where the Data Collection Instrument is Particularly Complex

CAI programming allows the construction of very complex instruments that often contain large numbers of "calculation intensive" computed variables – variables actually created by the instrument itself to record information, such as dates reported by the respondent which cover a long period of time. For example, when a survey collects extensive respondent and family histories regarding work patterns, educational attainments, family formation, and health issues over extended time periods the reporting of key family events interviewers and data producers can expect that specific dates might not always be accurate, particularly when the event occurred many years before the date of the interview.

Edit checks would identify and correct probable inconsistent records quickly avoiding the need to do so at the end of the data collection period, when it might be more difficult to uncover details about particular cases.

4.3. Place Active Data Processing Work as a Central Element of the Overall Survey Data Life Cycle

Opportunities to edit and clean data become less effective as the time span between the collection of a case and its review grows. Data collection and data processing should not be two separate stages that occur at very different times, but should occur simultaneously to quickly adjudicate problematic cases. A continuous data processing design will permit comprehensive descriptions of all data checking and cleaning operations from the start of data collection, providing secondary analysts with additional information for them to judge the quality of the data at their disposal.

4.4. Minimize "over Editing" of Data

After data collection ends, data producers often have a tendency to review any suspicious values not caught by the CAI instrument itself during post-collection checking and cleaning operations. The review may involve thousands of interviews, some going back several months or even a year or more. Retrospective editing from this time perspective is

very labor-intensive and is filled with uncertainty, especially when the interviews were collected much earlier. No matter what editing procedures are used, the review of all problematic cases at one time often encourages reviewers to feel that they must address every single case and make editing decisions, even if they do not have enough information to do so. In such cases, the question of whether or not the quality of the data is improved is open to debate (De Waal 2013). Checking cases soon after they are collected for consistent and logical reporting of life history events provides better and less costly opportunities to resolve them since more information is available to make informed decisions. In some instances, interviewers themselves can be recontacted to take advantage of their knowledge since they would have recently completed the case (Seiss et al. 2014). Even if one considers accuracy to be the most important aspect of data quality, survey researchers agree that timeliness and accessibility are also equally key components of quality (Biemer and Lyberg 2003).

4.5. Maximize the "cost-error Optimization" Ratio

The overall quality of the data depends significantly on how project resources are spent. Sufficient resources are necessary for all aspects of the survey data life cycle. Doing extensive data processing work after data collection ends might result in the expenditure of excessive funds and resources on data cleaning operations. If there is a large number of problematic cases to resolve, even if they might only involve a single variable or two, the result could be an unnecessary delay in the release/dissemination of public use files for the research community.

The costs and time involved in editing must always be balanced by the perceived improvements made to the statistical integrity of the final data file itself. Often referred to as "cost-error optimization", data producers should seek a balance in editing operations that seek out systemic problems, but avoid the temptation to check all values for all cases in hopes of producing a dataset devoid of error. Such a goal, of course, is never possible, but the power of modern survey instruments and technologies may make it difficult to decide where the "trade off" occurs. The data file may have 99% of all cases reviewed and cleaned, but the remaining 1% could easily take an inordinate amount of time to resolve. With limited resources, projects often must determine how to deal best with the cost-error optimization ratio. When is the best time to terminate cleaning procedures? Using CAI as part of a continuous processing operation allows projects to determine the kinds of consistency checks they will do. Data managers can concentrate on resolving only those cases. In effect, the system decides where the "trade off" occurs based on a specific set of rules developed by project researchers.

Project staff must always consider the "cost-error optimization" factor when performing these investigations. Test interviews entered by project staff or by real interviewers in a pretest should produce a set of rules and procedures that will determine which areas of the survey instrument and/or key variables are checked when the survey moves into full field production.

A key issue in this process is to determine the involvement of interviewers in the overall editing process. When a particular completed case exhibits unusual anomalies, field supervisors can contact interviewers directly as soon as possible to investigate and correct

possible errors. Recent research has indicated that those most closely involved in the data collection process are more likely to resolve inconsistencies with greater accuracy than other members of the survey research team (Sana and Weinreb 2008).

Project staff must balance the costs involved in having interviewers recheck cases against the potential loss of collecting additional interviews. A continuous data processing system requires a set of clear rules that determines when an interviewer becomes involved in a case, when the case is adjudicated in the main office or coordinating center, and when the inconsistency should remain on the data file. An effective system is not predicated on identifying and seeking to resolve every error or inconsistency. Its objective is to define which anomalies should receive further investigation and to provide a means of doing so at the least cost that will preserve as much of the original data as possible.

4.6. Encourage Faster Data Processing Times

Making the data checking and cleaning operations a continuous process will enable data producers of such complex surveys as NSFG to adjudicate interviews as they emerge from the field. If performed on a regular schedule (e.g., weekly or monthly or even quarterly), many cleaning operations could be completed before the data collection period ends. In a typical two-year data collection period for NSFG, there are an average of 59 female and three male interviews per month flagged for post-collection edits. Completing the editing process at this early stage can also result in reduced errors overall and improved efficiency in subsequent programs, that is, the production of derived variables (recodes), variable modifications due to disclosure review, and imputation. These additional processing steps can proceed more quickly, resulting in quicker turnaround times for the appearance of public use files and happier secondary analysts.

The implementation of a continuous data processing system with the NSFG rests on an ongoing collaboration between the survey organization that collects and processes the data and the principal investigators at the National Center for Health Statistics (NCHS). Subject matter experts at NCHS receive error-checking reports on a regular basis, evaluate proposed solutions that the survey organization provides based on a review of each case including any comments provided by the interviewer, and make final decisions on all edits. This process, which takes place while data collection is ongoing, lessens the amount of time devoted to this task after data collection ends. If the system is implemented in the same time as the collection of data is monitored, it can result in the release of public use data files several months earlier than originally anticipated.

A continuous data processing system also provides more flexibility in scheduling new releases of data. Since new cases are consistently reviewed, checked, and updated, they can be maintained in a single data repository. This facilitates the creation of different types of data files, for example, for different time periods or for specific kinds of respondent groups as data accumulates sufficiently to encourage analyses of new topics or subpopulations.

5. Total Survey Error (TSE)

Any quality enhancements derived from implementing a continuous data processing system directly relate to such nonsample aspects of the total survey error paradigm as usability/interpretability, relevance, accessibility, and timeliness/punctuality (Groves and

Lyberg 2010). Biemer (2010) cites the existence of quality reports and profiles as evidence that these concepts are attracting greater attention by survey managers. Yet they only exist for relatively few major surveys and focus more on discussions of response and imputation rates, but very seldom on other components of TSE, largely because guidelines and requirements do not yet exist (Groves and Lyberg 2010).

Increasingly, major surveys such as the European Social Survey (ESS) provide formal reports when data files and documentation are released for public use. Yet these documents still focus primarily on such topics as coverage, sampling, and nonresponse adjustment. The authors of the ESS quality report for Round 6 recognize this emphasis in their own work and go on to state that "the equivalent and comprehensive report for future ESS rounds should cover all or at least more aspects of the survey life cycle: from translation and sampling to data cleaning and processing. This extension is necessary to assess the overall quality of the produced data" (Beullens et al. 2014). A developed continuous data processing component that creates comprehensive documentation throughout the data collection process can become an integral part of the TSE evaluation and provide data users with a fuller understanding of the survey's "fitness for intended use".

6. Summary

This article has argued that the production of public use data files from complex surveys that rely on computer-assisted data collection software would benefit from the implementation of continuous data processing systems. Testing such a system with the National Survey of Family Growth allowed us to investigate some key questions about how successful it might work and what obstacles it might encounter.

Our first goal for testing the system focused on the feasibility of reviewing records soon after interviews were completed. It is common practice in survey research that all records are automatically checked for completeness, as well as plausible values on certain key variables. It was relatively easy to expand these checks to search for more subtle inconsistencies that would normally be resolved much later during the post-collection period. This work did involve additional time and effort, but became part of a regular monthly error-checking routine. We believe that implementing this enhanced review was successful, but it required full cooperation between the data producer and project investigators to adjudicate problematic cases on a timely basis.

Our second goal was to test the validity of using the CAI program, created to collect the data as efficiently as possible, to correct errors uncovered *after* interviews were sent back from the field. We considered this process as a novel development in CAI programming uses. Would the program correct erroneous values and make appropriate changes to values on subsequent questions if necessary? Our examination of all altered values suggested that the programming changes worked as intended. In particular, the program successfully created "inapplicable" or "not ascertained" values based on changes made to key variables that affected the routing of the questionnaire into different paths.

Finally, and perhaps the most difficult outcome to measure, were the costs and benefits of implementing this continuous processing system. The costs included the CAI programming changes, the monthly checks of all records, determination of which records to change and assigning appropriate values to each item, checking the results, and

replacing records with corrected values when the data producer and project investigators agreed on the change. The benefits included saving time in the post-data collection phase by adjudicating problematic records as early as possible, simplifying the recoding and imputation processes by eliminating inconsistent inputs, and focusing all staff on the importance of thinking about the creation of public use files as an integral component of the project from its inception. The key overall factor in measuring the success of this system may very well be the degree of cooperation and commitment to work on the task continuously. Since, in most cases, resources are always stretched, it is often easier to decide to pursue this kind of checking when the project is focused solely on producing public use files. We believe the system implemented for the NSFG improved the quality of the end product, but every survey with similar characteristics may decide differently.

While they are not an integral part of responsive survey design, we suggest that continuous data processing systems may add a new component to recent examinations of the effectiveness of such designs (Tourangeau et al. 2016) and shares similar characteristics. It allows data producers to administer the post-data collection process in the same manner as the planning and conduct of field operations. Just as principal investigators review the data coming in from the field and make adjustments to rework existing questions or formulate new ones and as survey managers follow sampling strategies and constantly review interviewer assignments to maximize response rates, so too data managers and processors should review and, where appropriate, correct erroneous data values. When continuous data processing happens while field operations are ongoing, we begin to mesh the survey production and data processing environments, moving them away from their long history of separation and closer to a unified process.

Utilizing the advantages of CAI programming as an integral part of a continuous data processing system can have significant advantages: the production of higher quality data, expedited availability to the research community and greater flexibility in addressing topics that are more timely and relevant to current research agendas.

7. References

Bethlehem, J. 1997. "Integrated Control Systems for Survey Processing." In *Survey Measurement and Process Quality*, edited by L. Lyberg, P. Biemer, M. Collins, E. De leeuw, C. Dippo, N. Schwarz, and D. Trewin, 371–392. New York: Wiley and Sons, Inc.

Beullens, K., H. Matsuo, G. Loosveldt, and C. Vandenplas. 2014. *Quality report for the European Social Survey, Round 6*. London: European Social Survey ERIC. Available at: http://www.europeansocialsurvey.org/docs/round6/methods/ESS6_quality_report.pdf (accessed September 2018).

Biemer, P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74(5): 817–848. Doi: https://doi.org/10.1093/poq/nfq058.

Biemer, P. and L. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons.

Biemer, P., D. Trewin, H. Bergdahl, and Y. Xie. 2017. "ASPIRE." In *Total Survey Error in Practice*, edited by P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West, 359–385. Hoboken, New Jersey: John Wiley & Sons, Inc. Doi: https://doi.org/10.1002/9781119041702.ch17.

De Waal, T. 2013. "Selective Editing: A Quest for Efficiency and Data Quality." *Journal of Official Statistics* 29(4): 473–488. Doi: https://doi.org/10.2478/jos-2013-0036.

Fellegi, I.P. and D. Holt. 1976. "A Systematic Approach to Automatic Edit and Imputation." *Journal of the American Statistical Association* 71: 17–35. Doi: https://doi.org/10.1080/01621459.1976.10481472.

Groves, R.M. and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74(5): 849–879. Doi: https://doi.org/10.1093/poq/nfq065.

Groves, R.M., W.D. Mosher, J. Lepkowski, and N.G. Kirgis. 2009. *Planning and Development of the Continuous National Survey of Family Growth* . National Center for Health Statistics. Vital Health Stat 1(48). Available at: https://www.ncbi.nlm.nih.gov/pubmed/20141029 (accessed May 2019).

Sana, M. and A.A. Weinreb. 2008. "Insiders, Outsiders, and the Editing of Inconsistent Survey Data." *Sociological Methods Research* 36: 515–541.

Seiss, M., E.A. Vance, and R.P. Hall. 2014. "The Importance of Cleaning Data During Fieldwork: Evidence from Mozambique." *Survey Practice* 7(4). E-ISSN: 2168-0094. Available at: http://www.surveypractice.org/article/2864-the-importance-of-cleaning-data-during-fieldwork-evidence-from-mozambique. (accessed September 2018).

Thalji, L., C.A. Hill, S. Mitchell, R. Suresh, H. Speizer, and D. Pratt 2013. "The General Survey System Initiative at RTI International: An Integrated System for the Collection and Management of Survey Data." *Journal of Official Statistics* 29(1): 29–48. Doi: https://doi.org/10.2478/jos-2013-0003.

Tourangeau, R., J.M. Brick, S. Lohr, and J. Li. 2016. "Adaptive and Responsive Survey Designs: A Review and Assessment." *Journal of the Royal Statistical Society, Series A* 180(1): 203–223. Doi: https://doi.org/10.1111/rssa.12186.

Received August 2016
Revised July 2018
Accepted September 2018