Topic (iii): Electronic data reporting – editing nearer source and multimode collections

Discussants: Pedro Revilla (Spain) and Paula Weir (United States)

1. Introduction of topic

Electronic data reporting (EDR) offers the possibility of using built-in edits in electronic questionnaires previously not possible in paper questionnaires or other modes of data collection. This topic covers all issues relating to editing as it pertains to EDR. This includes strategies and methodologies with respect to implementing editing at the point of data collection and its relationship to some other modes of collection and to editing that occurs in the post-collection processing. Topics of interest include:

- the impact of EDR on the editing strategy;
- the optimization of the effectiveness of both editing at data capture and at post-collection survey processing;
- performance measures and indicators of editing at data capture and post-collection processing as it affects the overall survey quality;
- challenges and issues such as security and confidentiality, respondents' burden, response rates, timeliness, and incentives;
- the use of focus groups, usability or cognitive testing of data providers with respect to editing at data collection.

The nine papers for this topic, two invited and seven supporting papers, touch on all of these topics of interest. Together these papers span self-administered surveys as web based forms, as well as other methods of self-administered electronic reporting, computer assisted telephone interview and computer assisted personal interviews for surveys addressing business and household data. They describe systems, guidelines, principles, approaches, studies and lessons learned focusing on electronic data reporting and editing of data at the point of collection.

2. Summary of Papers

Invited Papers:

WP 21: The Impact of EDR on Long-Established Surveys: Statistics Austria's Experience in the Short-Term Production Survey – *Wolfgang Koller, Frederick Rennert, and Guenther Zettl, Austria*

Paper summary

This paper presents the experiences while introducing EDR in Statistics Austria. The Short Term Survey (Production) is presented as an example. For this survey, two parallel electronic reporting tracks are in use, in addition to the original paper questionnaire. The authors describe the initial situation in 1998/1999 and give an overview of the fundamental design decisions taken at that time. These have served as the basis for a general strategy in the areas of data collection and editing. Some of the most important software products - among them e-Quest and e-Quest/Web - and the design principles used in their implementation are presented in brief. Finally, the paper

will demonstrate to what extent and in which ways the monthly procedures for processing the collected data have changed in consequence, as compared to the "paper-only" era. One point to highlight is that while 50 subject matter employers were entrusted with processing the Short Term Survey in 1996, nowadays they number only 30.

Points for discussion:

1. How to integrate EDR with existing processing systems? How to achieve early integration of the various respondent tracks?
2. Can EDR reduce respondent burden?
3. How to weigh the number and type of validations?

WP 22: Designing Interactive Edits for U.S. Electronic Economic Surveys and Censuses: Issues and Guidelines– *Elizabeth Nichols, Elizabeth Murphy, Amy Anderson, Diane Willimack and Richard Sigman, United States*

Paper summary

The paper *Designing Interactive Edits for U.S. Electronic Economic Surveys and Censuses: Issues and Guidelines* from the US Census Bureau describes the iteractive editing approach used in computer self administered questionnaires (CSAQ) that are browser-based (on-line web-survey/Internet questionnaire) and used for collecting economic data from businesses. The authors offer preliminary guidelines for incorporating edit checks into CSAQs based on a their usability study. In addition, they provide their findings, themes and issues about their approaches in interactive editing.

The two questions most asked regarding editing are: How many edits? Should failed data be allowed to be submitted? Typically, Census does not require respondents to resolve all failures before submission. In those cases where fields are required to pass the edits, the responses are considered unusable otherwise, and nonconformance results in unit nonresponse. The overall philosophy of editing is based on 2 principles: 1) Let the user be in control to the extent possible, (representing the respondent's point of view) and 2) Obtaining some data, even failed data, is better than no data (representing the survey management perspective).

The Census Bureau conducted a usability study on edit-failure notification and respondent behavior using the think-aloud protocol with video tape recording of the session. This study yielded 14 guidelines for editing. Many of these guidelines are centered on the first principle of letting the user be in control to the extent possible. For example, respondents prefer immediate checks that are run when the user chooses, possibly iteratively, avoiding unsolicited pop-up messages that frequently are ignored by respondents. Respondents should also be informed of the policy for submitting data with unresolved edit failures.

The research and survey experience by the Census Bureau are highlighted by the following points:

- The use of edit checks has grown and averages 2 per item

- A reasonable number of edits will not necessarily increase respondent burden or lead to unit nonresponse
    - Respondent acceptance depends on several factors including perceived usefulness
    - Allowing submission of data with edit failures reflects more willingness by the Census Bureau to accept measurement error over nonresponse error
- Not all post-collection edits can be embedded in interactively (programming or technical issues or macro level edits)
- Priority should be placed on edits on items that are mission critical, along with guidance from subject matter experts.

Points for discussion:

1. How do we determine the trade off between measurement error and nonresponse error as they relate to interactive edit checks, and what type of research will assist us in that determination?
2. What is the efficient and effective balance of interactive edits and post-processing edits taking into account costs and effect on data quality?
3. What is a reasonable amount of interactive edits and how do we convey their usefulness to generate respondent acceptance? Should this explanation be incorporated into the edit failure message?

Supporting Papers:

WP 23: Evaluation of Data Collection via Internet for the 2004 Census of Population Test – *Danielle Laroche and Laurent Roy, Canada*

Paper summary

This paper presents the main results of the 2004 Census of Population Test conducted by Statistics Canada in order to measure the effect of a new data collection method via the Internet on the content and the quality of data. This method will be applied in the next 2006 Census of Population in Canada. Respondents will have the option of completing and returning their questionnaire via the Internet. The author describes the main characteristics of the electronic questionnaire such as non-response messages, error messages, automated skips and on-line help that were tested during the Test. A comparative study of the item non-response rate and the follow-up rate between the electronic and paper questionnaires is presented. The study shows that data collection via the Internet is more complete and less expensive.

Points for discussion

1. How to evaluate the trade-off between the number of error messages of the electronic questionnaire and the fatigue of the respondent?
2. What differences exist between the households that respond via the Internet and the ones that respond in paper? Would these differences have influence in the final results?

3. What kind of measures should be implemented in order to increase the take-up of the electronic questionnaire? Will the respondents get some kind of reward to fill in the electronic questionnaire?

WP 24: Data editing for the Italian Labour Force Survey – *Claudio Ceccarelli and Simona Rosati, Italy*

Paper summary

The paper *Data Editing for the Italian Labour Force Survey* (LFS) addresses the redesign of the LFS that utilizes a rotating panel design with mixed reporting modes of Computer Assisted Personal Interviewing (CAPI) and Computer Assisted Telephone Interviewing (CATI). The editing approach described is a combination of selective editing, and automatic editing. Because automatic editing identifies all incorrect records and is low cost and not time consuming, records are first processed through automatic editing, and then split into two paths—critical and non-critical. Critical records that are incorrect due to systematic errors are processed through a deterministic algorithm for imputation, and those that are incorrect due to probabilistic/random errors are imputed according to the Fellegi and Holt methodology. (Random errors have equal probabilility of occurring in different variables and not correleted with errors in other variables. Systematic errors are defined as non-probabilistic, generally due to some defect in the survey structure.) Noncritical records could be imputed automatically or not without reducing data quality. At present these are not automatically imputed, but are imputed through a deterministic algorithm whose implementation is time consuming, requiring re-edit of imputed values, and needs modification each quarter.

After data capture, data are edited using the automated system based on the Fellegi and Holt model. This can be performed every week to provide timely notice of errors due to the electronic questionnaire. Explicit edits relating to categorical variables are performed in the a version of the System for Editing and Automatic Imputation, SCIA, that was included in the CONCORD system, Control and Data Correction, according to the Fellegi and Holt methodology. Edits related to continuous variables are translated into a SAS program including if-the-else rules.

It is assumed for this survey that deterministic imputation is more suitable for correcting systematic errors (2.5% of records) while probabilistic methods are more specific for errors generated from random error models. The deterministic imputation assigns only one value determined a priori on the basis of other values or by experts. The probabilistic imputation assigns a value according to a stochastic model or using a donor unit considered to be similar. Approximately 95.5% of records require no imputation, and 3.5% require only one impute. It is believed that the increase of data quality due to editing is usually negligible, and implementing for large complex data sets may be too costly in timeliness and resources. It appears unnecessary to correct all detailed data but the information on error detection plays an important role in improvement of the electronic questionnaire and data collection, and provides timely feedback on the interviewers. Fatal errors can be identified in time so that errors can be eliminated.

Points for discussion:
1. It is stated that noncritical records could be imputed automatically or not *without reducing data quality.* How is data quality being measured in this case?
2. How are critical and noncritical defined?

WP 25: Electronic Data Collection System Developed and Implemented in Central Statistical Bureau of Lativia – *Karlis Zeila, Latvia*

Paper Summary

This paper describes the new Electronic Data Collection System (EDC), Web based system that has been implemented for 34 surveys. The web forms preserve the look of the paper questionnaires to the extent possible to ensure simple transition to the web for the respondents. This system is part of the overall Data Management System that is centered on common metadata base from which all processes are driven. By using the common validation rules description, any rule can be added to the data collection that is defined in the metadata base. Before sending data, data validation is performed on the respondents' side.

Points for discussion:

1. Are data that do not pass the edit rules allowed to be submitted?
2. How are respondents presented the information on data that do not pass the edit rules (immediately? Pop-up or list?) How much information is conveyed? Do the messages recommend to the respondent what action to take?
3. Based on the comments received from the respondents, what changes have been made to the system?

WP 26: Modernization of the Data Collection Systems at the CSO of Poland – *Krzysztof Kurkowski, Poland*

Paper summary

This paper presents the experiences in a project of modernisation of the data collection systems at the CSO of Poland. The beginning of the project was in 2004 and it should be finalized in 2006. Pilot projects for electronic forms created with the use of Internet technology have been conducted in recent years. During 2005 there are plans to implement 30 different electronic forms over the Internet. Within the perspective of the upcoming few years, the main channel of informational flow between the CSO and the reporting units is planned to be the Internet. A Reporting Portal is being developed to offer a new functionality related to collecting data from statistical units. Data entered from forms to the Reports Repository is sent to logical and computing audit system. In case errors are discovered in the Report by the logical and computing audit system, its clarification should take place, which is to lead to correcting the Report. In the case of the electronic channel, the notification via e-mail is used.

Points for discussion:

1. How to integrate the data editing strategy in a multi-modal data collection system?
2. If the main mode of data reporting is planned to be the Internet, how to increase the take-up of the electronic questionnaire?
3. How to ensure uniformity when different modes of data collection are used?

WP 27: EDR and the Impact on Editing – *Pedro Revilla, Ignacio Arbues, Margarita Gonzalez, Manuel Gonzalez, and Jose Quesada, Spain*

Paper summary

This paper discusses the possibilities of Web questionnaires in order to reduce editing tasks. The possibility to use built-in edits allows reporting enterprises to avoid errors as they are made. The elimination of data keying at the statistical agency directly gets rid of a common source of error. Hence, some of the traditional editing task could be reduced. The combination of built-in edits and selective editing approach appears very promising. The paper explores if after implementing correct Web edits, no traditional microediting will be needed. Some practical experiences in the Spanish Monthly Turnover and New Orders Survey are presented. A Web form is offered to reporting enterprises as a voluntary option to respond to the Survey. Free of charge tailored data is sent to the enterprises that choose the electronic option, in order to increase the take-up of the Web questionnaire. When an enterprise sends a valid form (i.e. passing the mandatory edits), it immediately receives tailored data from the server. These tailored data consist of tables and graphs showing the enterprise trend and its position in relation with its sector.

Points for discussion:

1. Many statistical agencies are offering Internet questionnaires as a voluntary option. Hence, a mixed mode of data collection is used. How global strategies should be designed? Should data editing strategies differ when using paper that when using an electronic questionnaire?
2. What kind of edits should be implemented on the Web? How many? Only fatal edits or fatal edits and query edits?
3. What kind of edits should be mandatory when using Web questionnaires?

WP 28: EDR and the Impact on Editing – A summary and a Case Study – *Paula Weir, United States*

Paper summary

This paper describes the growth of the various types of electronic reporting for 65 surveys conducted by the agency between 2003 and 2004. While significant growth has been experienced, only diskette/CD software applications, and Internet surveys provided editing at data collection beyond edits required for data capture. Even in these cases, editing is still being performed in the traditional data processing stages to accommodate unresolved edits from IDC respondents, as well as to perform macro edits on the integrated responses from all modes of

collection. Balance of the two phases of editing is viewed as optimal for improving efficiency and data accuracy without negative side effects on response rates.

One fully web-based survey that recently implemented and editing module was examined in terms of how the respondents used the edit feature in reporting. A respondent questionnaire on the edit process revealed that 25% of the respondents reported that they ignored the information provided regarding the failed data, and others did not understanding why the failure occurred. In addition, all respondents reported that navigation between the review screen and the main screen was not a problem, yet 20% reported flipping back and forth one or more times, and 33% reported that they record the information provided to a location outside the system for further review. These findings were compared to an error log that revealed on average respondents change screens from the main screen to the review screen more times than the number of edit failures, resulting in a screen change to failure rate of 1.3 changes per failure. Analysis of the logs by responses to the questionnaire provided evidence that some respondents were editing data based on partial information, thereby affecting the edit rule and resulting failures.

EDR expands the "self-administered" role of the respondents to include their interaction with the edit process and, as such, requires that new indicators on the performance of the edit process be constructed and analyzed. An edit strategy for each survey should recognize the conflicting goals of maximizing the use of the EDR option by respondents and minimizing the errors on the submitted data. The strategy must take into account when/if to use hard or soft edits, when to invoke the edit in the data entry process, how to present messages regarding the edit failures, and how to navigate efficiently to correct errors and/or submit the data.

Points for discussion:

1. Given the new role of the respondent in EDR with respect to the edit process, what new indicators of performance of the process should be constructed and analyzed? Should we restrict the EDR application to prevent misuse of editing by respondents or only implement edit rules that are not affected? How does this fit with the principle of letting the user be in control expressed in the US Census paper?
2. How do we construct an edit strategy that takes into account differences in surveys, as well as the overall data quality strategy for the survey or the final product?
3. Is there knowledge from other disciplines that we should seek that can guide us in providing effective edit failure messages that convey the correct meaning and desired action from the respondents?

WP 29: Electronic Data Reporting and Data Collection Edits at the National Agricultural Statistics Service – *Daniel Beckler, United States*

Paper summary

This paper discusses the NASS' approach to EDR, including how data collection edits are built and applied on Web surveys. The implementation of these Web data collection edits is compared and contrasted with edits utilized in NASS' other data collection modes. Most NASS data collections are multi-modal and include some combination of mail, face-to-face, and telephone. Most recently, NASS began making all of its surveys and censuses available on the Web. The

necessity to produce Web versions of surveys also brought the need for a system to develop them. EDR consists of the Question Repository System (QRS), a series of PERL scripts running on a Web server, and associated databases. Data collection edits are utilized for CATI, the Web, and face-to-face modes in order to improve data quality. These edits are generally subsets of post-data collection edits. Problems identified and corrected during data collection minimize the amount of imputation during post-data collection processing and reduce burdensome re-contacts. Compared to those data collection edits in CATI or the Web, there are fewer face-to-face edits. Moreover, CATI data collection edits are more numerous and complex than Web data collection edits.

Points for discussion:

1. How to integrate the Web into the overall data collection program? How extensive and how complex data collection should be?

2. How data collection edits should be implemented – specifically how many, how complex the edits should be, and how should error messages and visual cues to indicate which responses are involved be conveyed to respondents?

3. It is convenient to use hard edits (i.e., changes to reported data are mandatory) in Web data collection or only soft edits?