Working paper No. 37

English only

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iv): Data editing by respondents and data suppliers

ELECTRONIC DATA REPORTING–MOVING EDITING CLOSER TO RESPONDENTS

Supporting Paper

Submitted by the Energy Information Administration, DOE, United States[1]

I. INTRODUCTION

1. As more surveys that historically were paper/mail surveys have begun offering the option of electronic data reporting, the potential for editing data at data capture has greatly increased. This paper examines that editing potential for electronic data reporting through computer self-administered questionnaires (CSAQ) via web surveys, downloadable software, and e-mail attachments. The presence and the extent of fatal and query edits that are implemented at the initial data entry and capture, versus those implemented in the traditional data editing stage is dependent on: 1) the amount of development resources dedicated; 2) the sophistication of the electronic option selected; 3) the security of the transmission that is required; 4) the quality of the data that is required; and, 5) the amount of respondent burden that is acceptable. Some specific examples in business surveys will be discussed to illustrate the balance between these factors and optimization of the total survey operation.

II. OVERVIEW—THE IMPUTUS

2. The Energy Information Administration (EIA) conducts approximately 65 surveys on all aspects of the energy industry. Most of these surveys are business surveys that vary in frequency from weekly to every four years. The 65 surveys vary in mode but historically, like the majority of business surveys, most are mail-out/mail-in paper and pencil surveys. The U.S. Paperwork Elimination Act of 1998 was an encouragement for surveys to move into more electronic modes but little progress had actually been made in the first years following the Act. Some surveys continued to offer electronic data reporting options that had been developed years previously,

[1] Prepared by Paula Weir (Paula.Weir@eia.doe.gov)

before Windows, but its success was limited to large firms, or isolated surveys. These early electronic options required mailing software, user's guides, etc., to respondents for initial installations and version updates. A few surveys provided respondents a diskette that contained survey forms on which respondents could enter their data and mail back the diskette containing the completed survey form. A number of other surveys provided a diskette that contained an application that the respondent installed on their PC, entered data on the form via the software, and then submitted the data electronically. Only a few surveys, particularly those weekly surveys with quicker turnaround, used Computer Assisted Telephone Interview (CATI). On the other side of the response time spectrum, a Computer Assisted Personal Interview (CAPI) was used for a much less frequent consumption survey. In 2001, only two surveys had moved to internet data collection. However, this survey-mode profile changed dramatically after September 11, followed by the anthrax scare. The main post office, Brentwood, which received most government mail was shut down. Mail delivery came to a halt. No longer was timely receipt of survey forms possible. Furthermore, respondents strongly preferred not to receive mail from government locations anymore. Facsimile and e-mail became the short-run solution to this overnight crisis. As it is often said, real change occurs when there is a crisis, and the true driver is the urgency or compelling context for change. This was certainly the case here. Many of the survey forms that had been placed on EIA's web site as PDF files for the purpose of reference/documentation now were being printed by respondents, filled in by hand, and faxed back. Other respondents spontaneously began sending e-mails, some with self-designed Word or Excel attachments. While for the short-run posting more surveys in PDF format on EIA's web site allowed surveys to keep operating, this crisis approach did not represent the most efficient electronic collection method. It was clear that respondents were ready to accept more electronic methods, especially methods for which they had a comfort level.

III. ELECTRONIC METEHODS IMPLEMENTED

3. One alternative method to faxed PDF or e-mailed attachments that was implemented fairly quickly by some offices made use of formatted Word files, and Excel files. Along with many of these was the option of secured or unsecured file transfer. These formatted Excel or Word versions of the surveys were placed on EIA'web site, in addition to the PDFs. At the top of the survey form, respondents were instructed to submit the form by mail, fax, e-mail or secured file transfer (SFT), HTTPS. Respondents were advised that e-mail is an insecure means of transmission and that the data are not encrypted. The instructions recommended that the respondent use a secure method of transmission, HTTPS, which is an industry standard method to send over the web. To encourage secured reporting, a link was placed directly on the survey form that directed the respondent to the secured transmission. Once the first survey form was formatted, and SFT developed, and the option posted on the web, the process became fairly routine to implement for other surveys. The SFT option was then also offered for submitting PDF versions.

4. The surveys that implemented the formatted files were successful because respondents felt comfortable with this option. First, not only were the files easily accessible on the web, they used formats/software familiar to the respondents. This made electronic reporting convenient. Secondly, the files were form image, so they looked like the current survey forms, also making reporting electronically simple. And thirdly, concerns regarding the confidential nature of the information were mitigated by the secure file transfer option. This third aspect addressed the confidentiality and privacy concerns often reported in the literature as respondents main reason for not choosing internet reporting. In addition, the respondent did not have to install anything but, if desired, they were able to store a copy of the filled in form on their own disk or hard drive, or print a copy for filing. From the respondents' viewpoint, this method was convenient, simple

and safe. From EIA's viewpoint, data were received more quickly by eliminating mail time. Forms were more readable because respondents had keyed the data into the formatted files. Total or net values were calculated for the respondent as data were entered on the spreadsheet, so some potential errors were avoided. However, received electronic forms were printed and re-keyed into the survey processing systems because importation of the files was not as easy as originally thought to be.

5. Importation of the files was difficult because importation was dependent on the standardization of the format of the files. Some respondents removed the protection on read-only cells, and changed fields. Importation was also dependent on the expectation of only one survey response for each submission. Some respondents who reported for more than one survey creatively combined multiple survey responses into one file for submission. More recently, passwords have been instituted to restrict permission to unprotect the worksheets to prevent some of these problems. However, other than automated totals or net values, no other interactive intervention at data entry, such as editing has been implemented. One group of surveys is currently working to develop an economic and secure way to transfer proprietary data back to respondents by e-mail to address issues with the data such as edit failures. This edit approach though is more comparable to the more traditional edit stage of processing.

6. Previous attempts using software that was installed on the respondent's computer have had mixed success. These electronic data reporting options required version control which included: a) distributing the version, user's manual, and install directions, and b) migrating to the new version while supporting the old, and then, c) retiring the old version. Respondents also had technical issues with loading software to their computer. Electronic submission of the reported data back to EIA was an issue with respect to security and changing firewall specifications. The import capabilities of the software, particularly for larger respondents, though did allow respondents to streamline reporting for frequent surveys, making this approach fairly successful. One particular application for reporting petroleum data, the Petroleum Electronic Data Reporting Option, PEDRO, has recently been updated. At this time, respondents are still mailed the software, but development is underway for respondents to download the software as needed. The respondent can import data if desired or enter the data, and edit their data by clicking on the checkmark icon, and refresh/rerun edits at any point. The application contains edits for control information, such as blank address fields, or invalid state code, as well as edits the survey data. These fatal edits include totals equaling sum of details, presence of data for related cells, and line imbalances. The edit failures are displayed as a pop up box, while highlighting the problem cell on the form, displaying all failures one by one. One line provides the edit failure message, and a blank second line is provided for the respondent to provide an explanation. Also, a summary screen shows all the failures with descriptive text. When the respondent submits the data they are encrypted and transmitted to an FTP server to an account and folder for PEDRO. The new version of PEDRO's communications system contains FTP related scripting done via an active-x component that connects via an anonymous user (with Write access only) to the PEDRO FTP server and transmits the survey data to survey specific group folders on the ftp server. Data are stored on the server until the survey runs a check-and-gather program that creates text files that are then imported into the survey processing systems. Logic was incorporated that determined which version of a survey form was used based on the report period. The new and old version will be run until all respondents have upgraded.

7. A second application, the Electronic Filing System (EFS) has also been recently updated. The survey collection previously was performed by mailing out application diskettes to respondents who used the application to fill in the survey form and then mailed the diskette back to EIA. The new version of EFS is a PC based application that the respondent downloads from

EIA's website. The respondent installs the application on their PC. Editing is performed as the respondent enters the data for select fields. For example, for the cells that are compared to previous year's values, the previous year's value is displayed below the active data field. When the data differ by a pre-defined percentage, a footnote screen automatically appears, requiring the respondent to footnote the discrepancy. The respondent can either type in an explanation, click the estimated data button, or change the data to be able to move to the next data field. Other edits include mandatory fields, consistency checks requiring the presence of complementary data. Before submission, the respondent chooses the validate form button to generate a discrepancy report which lists each potential error found in the current data submission. All errors that are not footnoted have to be corrected or footnoted before submission of the data. The discrepancy report ensures that all complementary fields are completed, data are within established ranges compared to last period's data, totals match sum of details, as well as, data satisfy other integrity checks. Some of the flagged items are labeled errors, fatal edit failures, meaning the value must be fixed and others are labeled as a warnings, query edit failures, indicating the value may be fixed by using a footnote to explain. Data are submitted to EIA by: 1) creating a single file that is e-mailed to EIA, 2) creating a diskette that the respondent mails to EIA, or 3) printing the completed form and mailing it or faxing the form to EIA. E-mail submissions are performed by using the EFS which guides the respondent through the e-mail process of attaching the file and sending the data using the respondent's e-mail software. These data are not considered proprietary, so standard e-mail security is acceptable in this survey. The non-proprietary nature of the data also enable date editing rules that use values beyond the respondent's current submission to be stored with the software that is distributed to the respondents.

8. Internet surveys, internet data collection (IDC), have become the alternative approach for electronic collection. While this method is viewed as more cutting edge, it requires more resources and time to develop, test and implement. Two main groups of Internet surveys were developed at EIA that are web browser based. These surveys tended to have "forms" design issues while other electronic collection preserved the paper design. They also were faced with different processing software interface and security issues. The main implementation problems for this approach tended to be the start-up problems of the distribution of passwords and the initial logon problems by some respondents. The usage of this reporting option for surveys that have implemented IDC has been steadily growing, ranging from 13% to 100% of the respondents across the surveys providing this option. Most of this growth has occurred in the last year, and has provided great opportunities to improve the editing process starting with the respondents editing their own data at data entry.

9. The IDC systems implemented so far at EIA use three-tier systems architecture. The **client tier** is the web browser. One system used for two different surveys required no software on the client machine. The other system used for seven surveys does require a one-time download to Oracle's J-Initiator. The **middle tier** is either IBM's WebSphre running on an IBM z800 partition, or Oracle's 9iAS Application Server running on a Unix platform (Sun server). In both systems this tier is behind the firewall but screened from the LAN. The **third tier**, the database tier, is Oracle RDBMS running on IBM z800 in screened subnet. Data are secured in transmission by https. The systems used HTML and Java for portability. The architecture was primarily driven by the need to use legacy hardware to keep costs down. The only purchase required was JBuilder, the visual tool for writing Java. Oracle tools were used for platform independence. The first system was developed in NT and ported first to an IBM S/390, and then z800 without code change. In addition, this system could also be run on a Unix box if desired. Both systems allow for submission and resubmissions and establish two-way communication between respondents and EIA. This communication allows EIA to also send and receive notices and comments regarding the surveys.

10. The IDC systems provide the respondents the ability to save their data at any point prior to submission. They are also provided the ability to export data to another application (such as Excel) and are provided online help, notification of progress, and confirmation of submission. Internal users have access to features and functions beyond those provided to external/respondent users of the systems.

11. The surveys implemented so far in the IDC systems contain edits that are classified or prioritized as either fatal or warning (query). The edits include:

- Range checks that check for numbers within a boundary, such as month must be between 1 and 12
- Null checks that define where null values are unacceptable
- Comparisons that compare current year's data with the previous year's data
- Validity checks that define valid codes for certain cells
- Consistency checks that require complementary data entry also
- Alpha/numeric checks
- Computational checks that require that a field be compared to a function of another field
- Common elements checks that require if a value is present, it is also present on other schedules.

12. The respondent has the option to run the edits by clicking on the Run Edits button or by clicking the "Submit" button asking for the data to be made available to EIA. The edits are run and the respondent is informed of the failures that should be reviewed and resolved. The respondent is provided an Edit Report to use in cleaning up any anomalies. This error-log contains a record for each item that failed the edits. After viewing the error log, the respondent can make changes to the data and rerun the edit check, and if passed, the error will no longer be listed in the log. The user can add a comment to the form stating the reason that the data value should be acceptable even though it failed the edits. Both the survey staff and the respondent are able to view the log simultaneously, enabling the two to discuss the contents. Additional edits are performed offline after all data are received from all modes of collection at the more traditional editing stage. This allows integration of current information to be used in the editing process and optimization of the editing and survey processing. The results of these offline edits are not viewable to the respondents.

13. Because the survey processing systems were already in place for most of these surveys, some retrofitting had to be done to receive the new data input from the IDC. For at least one survey, a new application was necessary to synchronize the processing system data base with the imported feed file from the Web based system at the end of each reporting cycle. Other surveys had similar requirements to input the data to the processing systems.

14. The table below provides an overview of the electronic methods being employed and their editing capability.

Electronic Method	Number of surveys using method and range of % respondents using method	Editing within electronic collection?
Unformatted e-mail	5 surveys 10-90%	no
Unsecured transfer Word or Excel file	39 surveys 1-100%	Automatic totals but no real edits; e-mail notices planned
Secured transfer Word or Excel file	27 surveys 1-70%	Automatic totals but no real edits; e-mail notices planned
Diskette/CD software (e-mail, fax or mail back)	3 surveys 3-57%	yes
PEDRO (mail CD, install and electronic submission)	10 surveys 1-27%	Roughly 60% of the surveys
EFS (Download software, install and e-mail, fax, mail completed form)	1 survey 100%	yes
Internet	12 surveys 2-99%	All but one survey

15. As shown in the table, not all surveys offer both secured and unsecured transfer of formatted files. The secured transfers were a more recent option offered. Both of these factors contribute to a smaller usage of secured transfers. Surprisingly, of the surveys offering both transfer options, 86% of the surveys had more respondents choosing unsecured transfer, but the number of respondents choosing secured is growing. This finding is interesting in view of the frequent reference that security concerns are respondents' primary concern about reporting via the Web. Unfortunately, the electronic efficiency of formatted files as a reporting mode has not been exploited beyond more timely submissions. At this time, editing has not been built into these formatted files, taking advantage of the potential for editing through software provided menu selections to validate the data. For some surveys, however, development is underway for an economic and secure method to e-mail proprietary data back to respondents regarding their submission and questionable items. The use of diskette/CD electronic forms with submission by mail or e-mail has been limited to just a few surveys but these surveys have been successful in moving editing to the data capture phase as the respondent enters data. Similarly, the EFS and PEDRO software that are installed on the respondents' computers have fatal and query editing performed by the respondent, including comparisons to previous period's data, but accomplish this with an open approach to providing the data necessary for comparisons according to their non-proprietary nature. The IDC surveys have also successfully accomplished editing by respondents including comparisons using previous period's data. The use of secured transmission, firewalls, and the three-tiered architecture have combined to satisfied proprietary requirements while providing efficient editing, reduced respondent burden, more timely and higher quality data.

IV. CONCLUSION

16. Spurred by the inability to receive in a timely manner hard copy mail at EIA, and concerns by respondents at receiving mail, electronic reporting at EIA made great advances in the last two years. In turn, electronic reporting by respondents has enabled more editing to take place at data capture by the respondents. While some electronic methods being used have not capitalized on this potential, others have. The diskette/CD or downloaded software and Internet

surveys are making the most use of editing at data capture. The sophistication of the editing performed is dependent on the editing rules and the data values used in the edit rules. Fatal editss are clearly the easiest and most commonly implemented. Edit rules that depend on fixed values or within form reported values are the next most commonly implemented. Edit rules that depend on external values such as previous period's report require that data be stored or accessible to the respondent at data capture, or quickly returned to the respondent in an interactive mode, and, therefore, are more difficult to implement for execution at data entry. Diskettes/CDs and downloaded software can contain the values required but are complicated by the confidentiality of the data and the need for data security. Internet surveys have been developed that provide the data needed to the individual respondents for editing while protecting that information from other respondents. Even though more and more editing is being performed by the respondents using electronic collection, editing is still being performed in the traditional data processing stage for non-electronic respondents, and other levels of edits performed across the integration of responses. It is important in all these surveys that 1) the edits performed on the electronic reports are the same edits performed on respondents using other collection modes, and 2) data from all collection modes are integrated and further level of edits performed to optimize the editing process.

References

[1] Clayton, Richard , Werking, George, "Using E-mail/World Wide Web for Establishment Survey Data Collection, BLS, http://stats.bls.gov/ore/pdf/st950030.pdf

[2] Nichols, Elizabeth and Sedivi, Barbara, "Economic Data Collection via the Web: A Census Bureau Case Study", U.S. Bureau of the Census, http://www.websurveyor.com/pdf/census.pdf

[3] Nicholls, W.L. II, Baker, R.P., & Martin, J. (1997) "The Effect of New Data Collection Technologies on Survey Data Quality' in L. Lyberg, P. Biemer, M. Collins, C. Dippo, N. Schwarz, & D. Trewin (editors) *Survey Measurement and Process Quality.* New York: Wiley.

[4] http://www.eia.doe.gov/oil_gas/petroleum/survey_forms/pet_survey_forms.html

[5] http://www.eia.doe.gov/cneaf/electricity/page/forms.html

[6] http://www.eia.doe.gov/cneaf/electricity/edc/contents.html