EDITING STRATEGIES USED BY THE U.S. BUREAU OF LABOR STATISTICS IN DATA COLLECTION OVER THE INTERNET

Stephen H. Cohen
Director, Mathematical Statistics Research Center, Office of Survey Methods Research,
U. S. Bureau of Labor Statistics, Washington, D.C., 20212, USA
Voice: 1-202-691-7400, FAX: 1-202-691-7426
E-mail: Cohen_Steve@BLS.GOV

Abstract: The U.S. Bureau of Labor Statistics (BLS) is the principle Federal agency charged with disseminating statistics in the broad area of labor economics. In particular BLS publishes the principle Federal economic indicators on unemployment and inflation in the U.S. BLS attempts to provide data providers with as many options as possible for submitting their reports such as over the Internet. The BLS approach to Internet data collection is to provide a singular, manageable, secure architecture for all surveys.

This paper will explore the editing strategy used by BLS on the Internet. The paper will summarize the types of edits used in post collection review and the strategy employed to incorporate edits into web based instruments. Edit experience from two Internet survey instruments will be discussed including implementation strategies, edit failure experience, and comparisons with edit failures of data collected by standard protocols.

This is a slightly revised version of a paper presented to the UNECE Work Session on Statistical Data Editing in Madrid, Spain, 20-22 October 2003. Views expressed in this paper are those of the author and do not necessarily represent the official views of the Bureau of Labor Statistics or the Department of Labor.

I. Introduction

Human error, such as programming mistakes, miscalculations, keypunch errors, and interviewer misclassifications are a pervasive fact-of-life for surveys. Their contribution to the total survey error, however, may be controlled or at least somewhat mitigated by good survey practice, forethought and planning, and the advances of computer technology. All such efforts to seek out and correct these errors fall under the purview of *data editing*.

Editing may occur at almost any phase of data collection or analysis. It ranges from the almost routine activities of correcting typographical errors or out-of-range entries done by interviewers as they enter information into the computer from the field or telephone center up to the elaborate statistical checks performed by computers to identify misshapen blocks of aggregate data. In longitudinal surveys comparisons to previous results are made. But in all cases, the goal is the same: to identify and correct as much error as possible.

In mail surveys of establishments, data editing is performed during post-data-collection processing. In establishment surveys that utilize computer-assisted data collection technologies, however, some or all of the data editing can be performed during data collection. For example, in surveys that use computer-assisted telephone interviewing (CATI) or computer-assisted personal interviewing (CAPI), data editing rules, referred to as *edits*, can be incorporated into the CATI/CAPI instrument so that the interviewer is notified of a response that fails one or more edits. The interviewer can then probe for an alternative response or for verification that the response is correct.

In addition, the Bureau of Labor Statistics (BLS) uses a hybrid technique for its compensation programs—data is collected in establishments using paper and pencil techniques by trained field economists and entered after the interview on a laptop with the data uploaded to servers in the National Office. The laptop system has edits that must be addressed before the field economist can successfully call the schedule complete. Edits on the database then flag schedules that initially looked correct but presented edit problems when viewed against other data.

Edits can also be incorporated into computerized self-administered questionnaires, which are delivered to the respondent by mailing disks or CD-ROMs or transmitted electronically over the Internet's World Wide Web. There are two types of web-delivered instruments: downloadable or browser-based. When using a downloadable instrument, respondents download the instrument from a web server to their personal computers, install it to complete the self-administered form, and then transmit data back over the Internet to the survey organization. When using a browser-based instrument, respondents use a web browser communicating over the Internet to view the questionnaire and to enter responses to survey questions. When an Internet instrument is used to collect survey data, respondents--not interviewers--are notified when a response fails one or more edits.

Survey organizations that collect data from establishments are interested in web-based data collection for a number of reasons [6], [15], [16]. One of these is the economics. Web-based data collection reduces mailing costs, interviewer costs, and data entry costs. Also, the use of computer-generated electronic mail in conjunction with web-based data collection reduces the costs associated with providing advance notices and reminders to respondents. Multimode surveys that include the web allow respondents to select the most suitable mode for their circumstances, in a sense, reducing burden. Web-based data collection also augments the palette of the questionnaire designer by adding to it web technologies such as hyperlinks, color

manipulation, dynamic graphics, and multimedia players that can provide instructions visually and/or aurally.

The objective of this paper is to review the data editing strategies when survey organizations that collect data from businesses adopt web-based data collection methods. The BLS is in the process of incorporating web-based data collection into a number of its establishment surveys. BLS is doing this to achieve the following potential benefits:

- Control of costs
- Improved response rates
- Decrease in perception of burden
- Improved data quality and
- For surveys with multiple closings--smaller revisions between preliminary and final estimates.

From the viewpoint of survey operations, the purpose of data editing is to provide information to an *edit reviewer*, who (or which) processes this information and decides on an appropriate follow-up activity. In interactive post-data-collection editing, the edit reviewer is a clerk or subject matter expert, whereas in fully automated post-data-collection editing the edit reviewer is another computer program. For CATI and CAPI the edit reviewer is the interviewer, and for online web surveys the edit reviewer is the respondent. Because the edit reviewer is usually a person, behavioral scientists may be able to provide insight into the performance of data editing with respect to the cognitive processes involved and system characteristics that limit or enhance these processes.

From the viewpoint of data quality, on the other hand, the purpose of data editing is to detect potential nonsampling errors—item nonresponse errors when data are missing and measurement errors when data are identified as inconsistent, questionable, or technically impossible. Hence, survey researchers should also be able to provide insight into the effects of changes in editing strategies in terms of the effectiveness of these changes in reducing nonsampling errors. Unfortunately, all of the data needed to study data editing from the viewpoint of data quality are not readily available in many survey-processing systems but instead must be obtained from research experiments imbedded in operational surveys. Nevertheless, this viewpoint allows us to state several survey design issues associated with changes in data editing when survey organizations adopt web-based data collection methods.

II. Editing at Practices at the BLS

BLS processes each of its surveys on systems that are individually tailored to the survey design to ensure maximum efficiencies in the agency's operations. All BLS survey data are extensively edited to ensure the agency releases the most accurate statistics possible.

Comment:

In addition to verifying valid entries, each post processing systems includes some of the following type of edits:

- Linear edits (i.e., their graphs are straight lines)

- Non-linear edits
- Conditional edits
- Ratio edits
- Variables that can accept negative values (negative values are edited out)
- Quantitative edits
- Qualitative edits
- Univariate range checks
- Multivariate, record-level checks for consistency
- Other consistency checks (e.g., work schedule if full-time employee)
- Route checks or skip patterns and
- Check of generated values against external standards.

See [14] and Appendix A for a complete summary of edit practices at BLS.

As advances have occurred in technology, data collection has progressed away from collecting information by paper and pencil, then keypunching the data onto a database. Current technology permits the field representative to enter data directly on a computer and then transmit the data to a database. EDI allows us to copy respondent data files directly to a database. For surveys with only a few data elements, respondents can enter data directly into agency databases through touchtone data entry. Mail surveys are processed by keypunching the data, editing for keying mistakes and then editing for logical errors.

III. Bureau of Labor Statistics (BLS) Internet Data Collection Experiences

The Bureau of Labor Statistics (BLS) is responsible for collecting information on labor economics. BLS is organized into four program areas plus support offices.

Surveys directly fielded by BLS are establishment based. Collection methodologies vary by program. Federal-State cooperative programs initially field most schedules by mail with telephone follow-up for nonresponse. At the other extreme, most schedules in the compensation and price programs are initially fielded by personal interview due to the complexity of the data requirements with mail used mostly for updating data.

Many surveys are longitudinal giving a respondent a copy of at least his most recent response and allowing edits that can take previous responses as inputs. BLS has begun to give respondents the option of recording their responses on the Internet. The Current Employment Statistics Survey, Annual Refiling Survey, Multiple Worksite Report Survey, Occupational Employment Statistics Program, National Compensation Survey, the International Price Program, the Producer Price Program and the Occupational Safety and Health Survey all have or will shortly have capability to allow respondents to submit reports via the Internet. This type of arrangement then raises the question of what type and how many edits do we put in place to ensure that we do not have to call back the respondent with questions.

The BLS approach to Internet data collection is to provide a singular, manageable, secure architecture with the following characteristics:

- Same entry point for all BLS surveys

- Common look and feel across surveys
- Support for multi-survey respondents
- Multiple levels of security
- System behind its own firewall, outside of BLS firewall
- Access controlled by BLS-issued digital certificates and passwords
- Monitoring and risk assessment of only one infrastructure.

Here we detail the results of efforts in three BLS surveys to illustrate editing problems associated with different levels of complexity in data-collection requirements. Most BLS surveys are not mandatory. The Occupational Safety and Health Survey (OSHS) is the only mandatory BLS survey.

IIIa. Current Employment Statistics Survey (CES)

BLS has explored various mediums to capture the data that makes it as easy as possible for a respondent to participate. The CES has employed touch tone data entry for years to ensure timely capture of the data as efficiently and low cost as possible. CES has arrangements with large responders to receive electronic files containing the required data elements, generally in a standard format. The process greatly reduces respondent burden. Since 1996, CES has had a small portion of its sample report via the Internet.

The CES is a monthly survey of employment, payroll, and hours. The data are used to produce the establishment employment statistics used in the employment situation release, which is a key economic indicator. The sample of 300,000 business establishments provides data that are published after only two and a half weeks of collection. Details about unemployment by industry, state, and area, average hourly earnings and average weekly hours are published. Respondents are queried about five basic items:

- All employees
- Women employees
- Production or nonsupervisory workers
- Production/Nonsupervisory payroll
- Production/Nonsupervisory hours
- Commissions collected for the Service Providing industry
- Overtime hours collected in Manufacturing.

The CES has moved away from mail collection over the last dozen or so years. In use since 1986, telephone data entry (TDE) is the backbone of CES collection. 150,000 reports are collected by TDE. Respondents call a toll-free number and receive a computerized interview. Growth of the Internet and improvement of computer technology allows use of new approaches: Web, Electronic Data Interchange (EDI), and Fax Optical Character Recognition.

CES began collecting data on-line from the web in 1996. One aspect of web methodology is on-line editing—the visual interface allows interaction with respondent not available in other self-administered collection methods such as FAX or mail. TDE systems could edit data and

repeat failures back to respondent, however, CES program staff believe the cognitive burden resulting from such edits would be excessive.

Initially the CES Web system performed basic edits only (April '97):

- Logic errors--e.g. All Employment greater than Production Workers
- Range checks--Average hourly earnings between $1.00 - $150.00
- Validity checks--numeric entry, mandatory fields (All employees) completed
- Data entry errors.

The results from these initial basic edits were as follows:

- Approximately 40 percent of current web sample has failed at least one edit check
- Approximately 3 percent of web reports fail one edit check each month
- In 88 percent of all edit failures, the respondent corrected and submitted data during the same session.

After a respondent has entered all of the data elements, he/she submits the report. Once the respondent hits the submit button, the edits are performed. If there is a failure, an edit box appears. The edit messages show as highlighted text below the data section on the instrument (see Appendix B for screen shots). The respondents have to reconcile the error or enter an appropriate comment code and resubmit. For the two global edits, the data must be actually corrected to submit the report, e.g. number of women workers cannot exceed total employment and number of production/nonsupervisory workers cannot exceed total employment.

Because data can be quite variable from month to month, the program managers decided that most edits needed to be soft. Experience with analyst review of edits tends to suggest that most records that fail current machine edits are accepted by the analyst and ultimately used in estimation. Therefore, the only hard edits relate to "impossible" type data situations or entry of nonnumeric values in the data fields.

The current CES Internet application was enhanced in May 2001 to include an expanded array of edits:

- Expanded basic and longitudinal edits
- Critical values vary by major industry division
- Over-the-month changes compared for each data element and for several "calculated" averages.

Up to 21 separate edit checks are performed for each report (depending on the data items on the report form). These edits are performed on the server side.

Except for the two "hard" edits cited above, respondents must either correct so the edit condition is removed or select "comment code" from a drop-down list of common reasons for large fluctuations in the data.

The enhanced edits were patterned after the edit parameters currently used in the CES CATI system. The primary difference was the level of industry detail at which the parameters were set. In the CATI system, the edits were specific based on 2-digit SIC and size of firm. For the Web edits, the parameters were set at the major division level (with size of firm also taken into account).

Experience with enhanced edit in CES Web application: With the introduction of enhanced edits, the incidence of edit failures rose from 3 percent of reports each month to about 7 percent. This was to be expected since many additional edit checks were now in place. The percent of units entering a comment code likewise rose from about 6 percent to about 14 percent. By comparison the proportion of TDE self-reported units that enter a comment code is only 3 percent. During CATI interviews the percent of records with a comment code is about 12 percent. So it appears that Web reporters are self-reporting comment codes at about the same rate as interviewers are entering them to explain large fluctuations in the data and considerably more often than touchtone respondent self-report a comment code.

Respondent reaction to longitudinal edits: Our overall experience has been favorable. We provided no advance warning to Web respondents when we introduced the expanded edits. Nevertheless, we had no negative feedback when they were introduced. To further check on this, we conducted several debriefing calls to respondents that had triggered one or more of the enhanced edits. Respondents did not express any concerns about the edits and were able to navigate through the process either by correcting the data or providing a comment code.

CES staff believe that more research on the cognitive aspects of edits--how they are shown, what wording is used, what options are presented--would be valuable.

IIIb. Occupational Safety and Health Survey (OSHS)

The OSHS is a mandatory survey designed to yield detailed specific industrial incidence rates of workplace injuries and illnesses. The survey provides information annually on the number and frequency of nonfatal injuries and illnesses occurring in the workplace. It also reports on the demographics and case characteristics for the serious incidents, those that require time away from work.

The OSHS statistical system is built on the collection of recordable cases. A recordable case is any occupational death, regardless of the time between the injury and death or the length of the illness or nonfatal occupational illness, or nonfatal occupational injury which involves one or more of the following: loss of consciousness, restriction of motion, transfer to another job, or medical treatment (other than first aid). The Occupational Safety and Health Act required that employers subject to the act maintain a standard set of records covering their injury and illness experience. BLS then asks for reports on the summary of incidents at a sampled site with further sampling of individual cases for the event, source, nature, and part of body affected.

In the OSHS production data collection systems, we perform nearly 170 edits (167 in the 2002 system). The edits fall into the categories of validation edits (data are in the correct format), consistency (comparing data between data elements - does it make sense), and reasonableness (primarily for coding - can you really have a funeral director working in a furniture store?).

For the 2002 OSHS survey just fielded, each traditional schedule in 47 States includes a URL and password for a respondent to enter his data directly via the Internet. About 1,000 schedules were returned via the Internet. With this first release of an on-line Internet instrument, we didn't want to overwhelm respondents with many error messages. We made a conscious decision to keep the editing to a minimum flagging only invalid data that would cause the database to crash (where it mattered most) and get the respondents in and out quickly. The edits included in the Internet instrument ensure validity of the data, for example, the value entered in Annual Average

Employment must be numeric and the value entered in the Total Number of Cases with Days Away from Work must be numeric.

We expected that we would have to make about the same number of calls to the respondents as we do currently. The edits occur when the respondent presses SAVE (which they can do at anytime on any page) or they press CONTINUE to go to the next page. The edit failure rates for schedules received via the Internet were no different than those received as traditional mail schedules.

The processing system edits establishment summary data for differences in total reported employment from BLS records, data out of range and injury and illness counts unusual for the reported industry. The sampled information requested for individual injuries or illnesses that meet reporting threshold is edited for consistency between the nature and source of the injury versus injury outcome. For example, if the nature of an injury is a sprain the body part affected cannot be the brain or skull.

IIIc. National Compensation Survey (NCS)

The NCS is used to collect compensation data. Products derived from NCS data include the Employment Cost Index, locality and national occupational wage levels, national wage and benefit cost levels, benefit incidence, and benefit provisions. These products provide the following information:

- Wages by occupation and work level, for localities, broad geographic regions, and the entire United States
- Employer costs for wages and benefits, and the rate of change in those costs the Employment Cost Index
- Percent of workers receiving benefits and provisions of those benefits.

The NCS samples about 40,000 establishments over 154 PSUs. Within each establishment, a probability-proportionate-to-size sample of occupations is selected. Each occupation is classified by work level. Wage data are collected in all sampled occupations within the establishments. Benefit data are collected in about 40 percent of the establishments.

Establishments are in the sample for approximately five years with 20 percent replaced each year. For about 60 percent of the establishments, only wage data are collected; with updating occurring once a year. For the remaining establishments, wage and benefit data are collected quarterly. Data are collected on establishment characteristics, occupation characteristics including work level and wages with incentive payments, and employee benefits

First time establishments receive a personal visit from a BLS field economist. Subsequent contacts are via personal visit, mail, phone, or electronic data transmission. Field economists enter data into a laptop-based data capture system developed by BLS. The system can accept data from respondents electronically.

The NCS is developing software to permit the collection of wage data via the Internet. The Internet will be used only to update information previously collected. The addition of Internet reporting will begin on a test basis in the summer of 2003.

The software under development will have seven major features:

- Selected schedules are made available to respondents through the Internet
- Respondents see information from their previous schedules
- Respondents enter updated information
- Basic edits help respondents identify problem situations and clarify or correct data
- Directions and a full-featured help system guide respondents
- Data moves from Internet collection vehicle to the NCS Integrated Data Capture system (IDC) by the field economist
- Field economists import the data into the IDC schedule using standard techniques.

Edit messages will pop up and require respondents to double-check their input. The respondent will have a choice among the following actions:

- Fixing the failed data
- Canceling the changes
- Saving the changes anyway (system flags cells)
- Documenting the situation in a remarks box
- Accessing a help screen.

The initial NCS Internet Collection System will only edit for invalid entries. Warning messages will also pop up for questionable work schedule entries or missing text in text fields.

Edits range from invalid data field entries to negative values for current employment levels, wage rates, and hours worked. For numeric entries, valid values are 0 to 9.

Warning messages are displayed for the following situations:

- The average salary in the occupation has decreased
- The number of workers in the occupation has changed by more than ten percent
- The average salary has increased by more than ten percent
- The work schedule is less than one hour per day or more than 8 hours per day
- The work schedule is less than 32 hours per week for a full time occupation, greater than 40 hours per week for a full-time occupation, or is greater than 32 hours per week for a part-time occupation
- The Current Reference Date is more than 60 days from today
- New employment differs from previous employment by more than 20 percent
- A text field has missing data such as Company Name is missing

Once the respondent completes data entry, the information will be transmitted to a BLS field economist who will have responsibility for reviewing the data and correcting or documenting any additional anomalous data and loading it into the NCS database. In validating the input, the field economist will have the assistance of the full range of data edits that are built into the NCS data capture system. The field economist may need to contact the respondent via e-mail or telephone to resolve any discrepancies that are not explained by documentation. Data edits in the processing system check for occupation code against salary range coded, wage rate compared to

the expectation for the local market area, salary range expected for the industry code associated, and so forth. At estimation, outliers are reviewed for validity.

IV. Summary of BLS Experiences—What We've Learned

BLS began Internet data collection in 1996 with the CES survey. Now at least eight surveys have or plan to have Internet data collection facility in the very near future. Initiation of Internet data collection was targeted to specific establishments while most initial efforts on editing have been on ad hoc basis. Themes emerge from the BLS experience:

- Edits start out at a basic level not to discourage respondent participation. Initial editing ensures that the data submitted are valid (i.e. the numeric data), that mandatory fields have been entered (i.e. the minimum number of fields to ensure a valid response), some gross checks for impossible types of entries
- Edits on Internet applications are kept to a minimum to ensure respondents continue to cooperate
- Edits on the Internet applications are borrowed from post-collection edits
- Program managers prefer to accept data with unresolved errors rather than receiving no data at all
- Reaction of respondents to edits is not shared across the organization.

CES has expanded its edit list over time as the managers have become comfortable with respondents' reactions. OSHS has made Internet data collection an option for most respondents for the 2002 survey. Now is the time to look back at the issues so that we can develop more sophisticated systems that will ensure data quality at reduced cost while maintaining respondent cooperation levels.

V. Research Issues associated with data collection on the Internet

As you can see from the case studies, BLS has included edits in Internet data collection using a conservative philosophy--at a minimum, we receive data, which may or may or not pass basic edits, from cooperative respondents. We seek to build upon past research [1], [2], [7], [8], [9], [10], [11], [12], [13], and [16] to develop a more ambitious research program, in which we start moving the extensive post-collection edits into Internet data collection in order to reduce costs and increase quality, while maintaining respondent cooperation. Issues that need to be studied include:

Respondent behavior issues. The following are some issues associated with maintaining respondent cooperation:

- How many edits should there be in a web instrument, relative to the number of questions? At what point does the respondent consider there to be too many edits?

- How should edit information be presented to respondents?

- To what extent can CATI/CAPI edits and post-data collection edits be incorporated into web-based instruments? Which ones? What kinds?

- How complex can Internet edits be on Internet instruments? What criteria should be used to judge this?

- How should we word error messages involving variables that are calculated by the system and not entered by the respondent? What types of tools are needed to effectively deliver complex edit-failure feedback information to respondents?

- Are certain types of edits inappropriate for web instruments? What types of limitations should be considered?

- What types of information should we collect during cognitive pretesting and usability testing to guide the types of edits incorporated into Internet instruments as well as how well edit-failure information is being communicated?

- If respondents need additional help to resolve one or more edit failures, how should it be provided?

- Will respondents react negatively to being asked to resolve data-collection edit failures and also receiving a clarification call from the Agency?

Data quality issues. Issues concerning data quality and resource allocation can arise when large mail surveys change to or offer the option of web-based data collection. These issues arise from the fact that large mail surveys have high variable costs (with respect to number of respondents and number of survey cycles) associated with data editing because clerks and subject-matter experts review the edit information produced by post-data-collection edits. Editing at the time of data collection, on the other hand, by the respondents' reviewing edit information from web-instrument edits can have high fixed costs for programming and questionnaire testing, but the corresponding variable costs associated with data editing should be much lower than those for post-data-collection editing. The following are some of the issues associated with this situation:

- Can some (or all) post-data-collection edits be eliminated by incorporating them into the web instrument?

- When post-data collection edits are incorporated into a web instrument, what critical values should be used—same as those used in CATI or post-data-collection systems or different values?

- Will there be an "editing mode" effect—that is, different responses to edits in a self-administered instrument vs. to traditional call backs requesting clarification?

- How should the effectiveness of editing be measured, permitting comparisons of different modes of editing?

- What are the consequences of the current strategy favoring measurement error over nonresponse error? What criteria should be used to determine the optimum tradeoff?

- How does one prevent overediting, such as editing out a trend?

- To what extent will editing at the time of web-based data collection reduce the number of failed edits occurring in post-data-collection editing?

- If web-collected data contains fewer edit failures than mail-returned data, can survey resources devoted to post-data-collection editing be reduced without degrading data quality?

- Will BLS be able to publish estimates earlier because of earlier receipt of data and less time spent reviewing post-data-collection editing information?

- What types of research should be conducted to determine the effects of Internet data collection on data quality?

Overall strategy issues. Although survey practitioners would very much like to have "generally accepted practices" or "rules of thumb" for many of these design issues, we expect this to be virtually impossible given the variety of surveys, administrations, and trade offs related to data quality and response. Instead we think it would be more appropriate to develop a set of guidelines to aid decisions related to editing.

- What criteria should guide the determination of editing strategies in web surveys?

- What strategy should be employed in evaluating overall editing performance? How should one trade off edits in Internet instrument vs. post-collection processing?

- How should one prioritize edits in selecting the edits to be included in an Internet instrument?

- How can one transition new surveys to the web without implementing web edits in stages as has been the philosophy to date?

- What types of empirical analyses of edit results (Internet and post-collection) should be conducted to provide feedback into the improvement of questionnaires (so that respondent errors may be prevented in the first place)? What are some current experiences in this area and the type of benefits that may be realized?

- Are there sources for working papers on studies not in the published literature on Internet data collection that address data editing? Do committee members know what the private sector and/or European community doing in this area?

VI. Acknowledgements:

The author wants to thank the following staff for their help in gathering the background information for this paper:

Philip Doyle, BLS
William McCarthy, BLS
Dee McCarthy, BLS
Richard Rosen, BLS

References

[1] Anderson, Cohen et al, 2003. "Changes to Editing Strategies when Establishment Survey Data Collection Moves to the Web", Presented at the Federal Economic Statistical Advisory Committee March 2003

[2] Bzostek, J. and Mingay, D. 2001. "Report on First Round of Usability Testing of the Private School Survey on the Web." Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #42.

[3] Dumas, J. and Redish, J. 1999. *A Practical Guide to Usability Testing*. Portland, OR: intellect.

[4] Fellegi, I.p and Holt, D (1976). "A Systematic Approach to Automatic Edit and Imputation." *The Journal of the American Statistical Association*, 71, pp. 17-35.

[5] Economic Electronic Style Guide Team. September 28, 2001. "Style Guide for the 2002 Economic Census Electronic Forms" U.S. Census Bureau, Economic Planning and Coordination Division.

[6] Evans, E. Forthcoming. "QFR CSAQ Evaluation." Internal Memorandum. U.S. Census Bureau, Company Statistics Division.

[7] Nichols, E. 1998. "Results from usability testing of the 1998 Report of Organization CSAQ" Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #19.

[8] Nichols, E., Murphy, E, and Anderson, A. 2001a. "Report from Cognitive and Usability Testing of Edit Messages for the 2002 Economic Census (First Round)" Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #39.

[9] Nichols, E., Murphy, E, and Anderson, A. 2001b. "Usability Testing Results of the 2002 Economic Census Prototype RT-44401" Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #49.

[10] Nichols, E., Saner, L, and Anderson, A. 2000. "Usability Testing of the May 23, 2000 QFR-CSAQ (Quarterly Financial Report Computerized Self-Administered Questionnaire)" Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #33.

[11] Nichols, E., Tedesco, H., King, R., Zukerberg, A, and Cooper, C. 1998. "Results from Usability Testing of Possible Electronic Questionnaires for the 1998 Library Media Center Public School Questionnaire Field Test." Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #20.

[12] Rao, G. and Hoffman, R. 1999. "Report on Usability Testing of Census Bureau's M3 Web-Based Survey" Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #26.

[13] Saner, L., Marquis, K., and Murphy B. 2000. "Annual Survey of Manufacturers Usability Testing of Computerized Self-Administered Questionnaire Findings and Recommendations Final Report" Internal Memorandum. U.S. Census Bureau, Statistical Research Division, Usability Lab, Human-Computer Interaction Memorandum Series #30.

[14] Stinson, Linda L. and Fisher, Sylvia K. 1996. Overview of Data Editing Procedures in Surveys Administered by the Bureau of Labor Statistics: Procedures and Implications, Presented at the First International Computer-Assisted System Information Computing Conference at San Antonio, Texas.

[15] Sweet, E. and Ramos, M. 1995. "Evaluation Results from a Pilot Test of a Computerized Self-Administered Questionnaire (CSAQ) for the 1994 Industrial Research and Development (R&D) Survey," Internal Memorandum. U.S. Census Bureau, Economic Statistical Methods and Programming Division, #ESM-9503.

[16] Tedesco, H., Zukerberg, A., and Nichols, E. 1999. "Designing Surveys for the Next Millennium: Web-based Questionnaire Design Issues," *Proceedings of the Third ASC International Conference*. The University of Edinburgh, Scotland, UK, 22nd-24th of September 1999, pp. 103-112.

Appendix A – Editing at BLS

Stages of editing

1. Most systems at BLS allow editing to occur throughout the survey process—from collection through the estimation steps. Editing occurs at:

- Interviewer Level
- Regional Office or State Level
- National Office Level
  - Pre-estimation
  - Post-estimation.

2. BLS Editing system allows manual review and changes:

- At data entry, eg while reviewing paper schedules or notes
- During data entry
- After data entry.

3. Editing software:

- Requires substantial data cleaning during data editing process
- Requires manual fixes of machine-generated signals
- Has batch capability with manual resolution of errors.

4. Characteristics of various BLS Editing systems include:Data entry and data edit at the same time (usually performed only on sub-sections of data)

- Data to be entered with correction (high speed)
- Full-screen data entry
- Verification (i.e., enter data twice)
- Implied decimals to be typed (e.g., type 23 for 2.3)
- Data-entry statistics to be generated.

5. In order to facilitate data analysts, editing systems have various reporting features:

- Lists of missing reports
- Reports
- Log or trace files
- Tables
- Use of Logical operators
- External; Records.

6. Most editing systems in use at BLS have many features in common. For example, most editing systems have the capability to edit continuous data, decimal values,

character data, and some binary checks, such as would be used for categorical data. Some systems imbedded in the data entry process itself and include many of the typical CATI/CAPI editing features, such as the capacity to enter and edit data at the same time through the pre-programmed specification.

- Alphabetic or numeric characters
- Acceptable numeric ranges
- Consistency checks (e.g. marital status versus age)
- Routes and skip patterns
- External standards for comparison.
- Decimal in wrong location
- Wrong Yes/No choice
- Wrong response option
- Wrong measurement unit (week, day, month, year)
- Wrong numeric entry
- Changing decimal locations.

7. By contrast, some errors many not necessitate editing and are unlikely to have an impact upon data quality and accuracy. A common example of this type of edit would be the presence of typos that do not obscure the meaning of text

8. Systems employ a variety of edit types to check for logical relationships within the schedule and across the historical database. These types of edits include

- Linear edits (i.e., their graphs are straight lines)
- Non-linear edits
- Conditional edits
- Ratio edits
- Variables that can accept negative values (negative values are edited out)
- Quantitative edits
- Qualitative edits
- Univariate range checks
- Multivariate, record-level checks for consistency
- Other consistency checks (e.g., work schedule if full-time employee)
- Route checks or skip patterns
- Check of generated values against external standards.

9. Editing systems are designed to:

- System can accept logical operators (e.g., ‘and,’ ‘or,’ ‘not’)
- System can perform intra-record checks
- States, regions etc. have customized error limits
- Software performs graphical inspection of the data.

10. Editing systems can perform the following functions:

- Check edits for redundancy
- Check that edits are not contradictory
- Generate implied edits
- Generate external records.

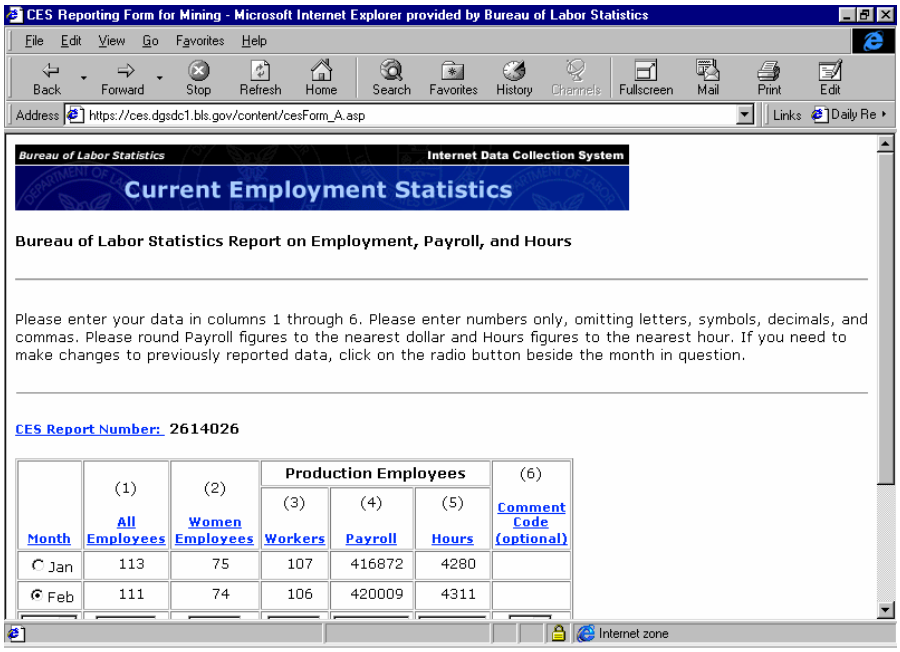
11. System can perform statistical edits based upon:

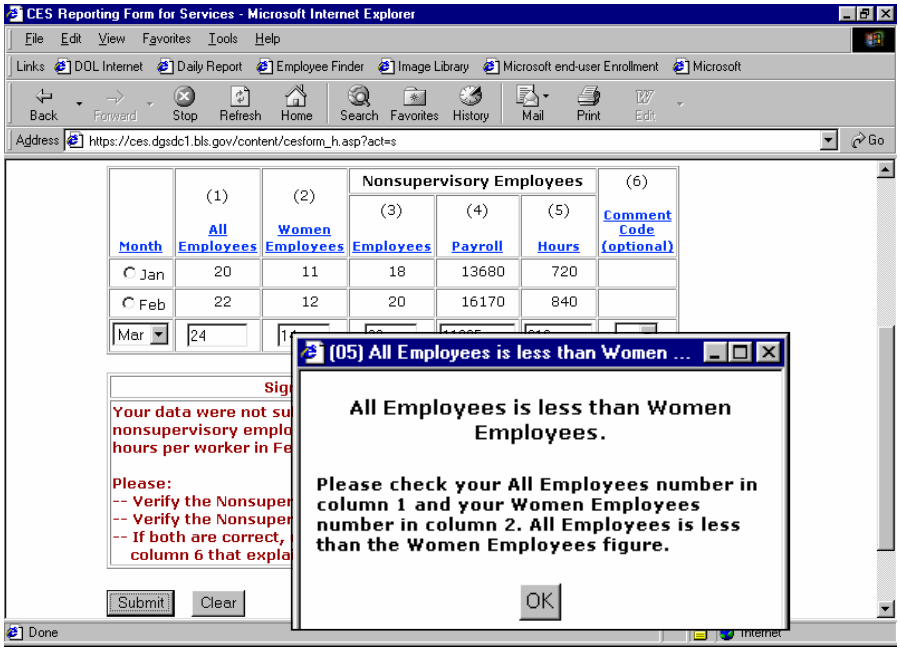
- Historical data
- Cross-record checks
- Univariate outlier detection
- Complexity of data structure.

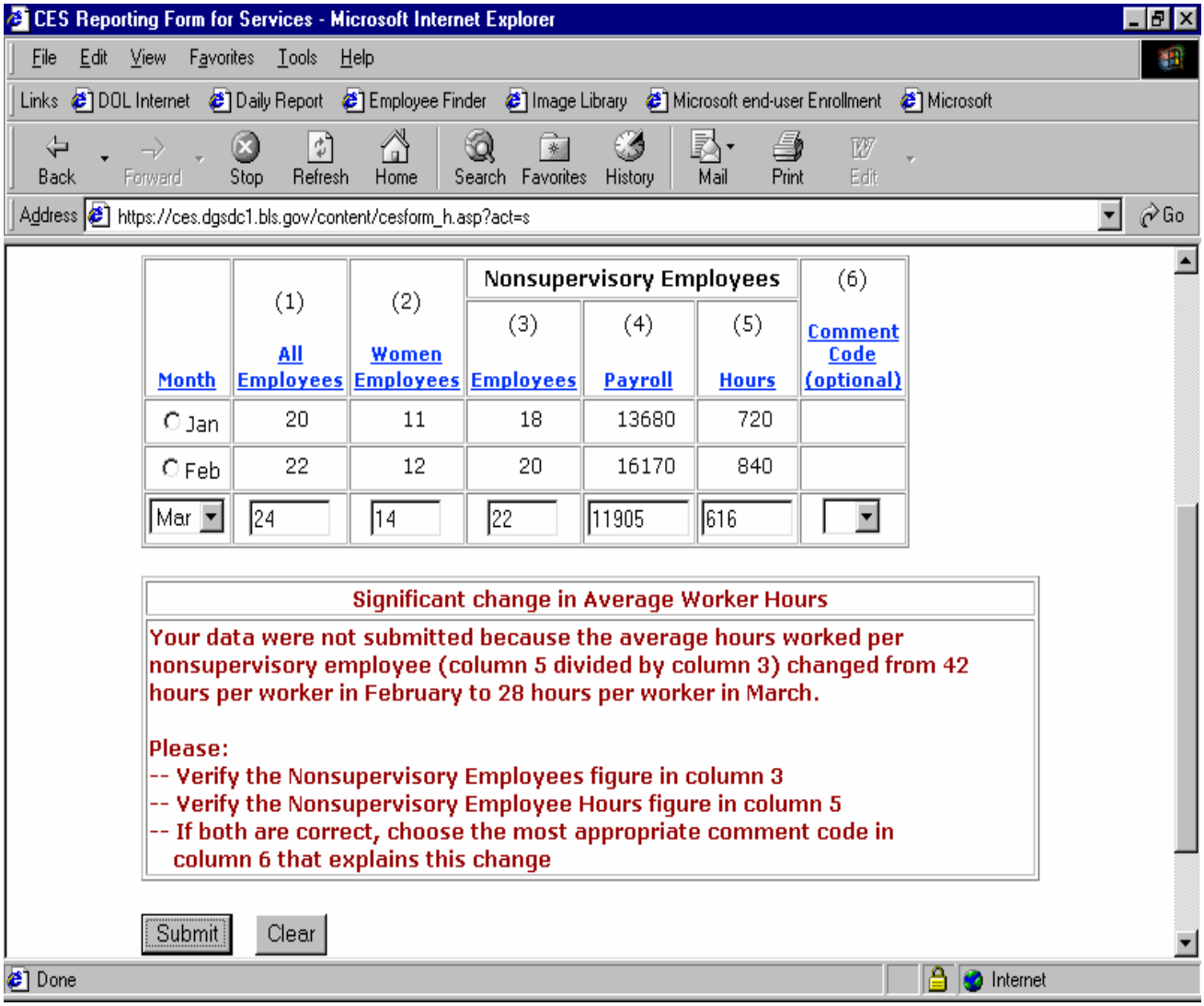
12. Editing system allows: Manipulation of hierarchical data

- Manipulation of complicated skip patterns
- Manipulation of subfiles
- Manipulation of cross-record checks
- Respecification of edits without other manual software changes
- Respecification of data without other manual software changes.

Appendix B CES Edit Screen Shots




CES Reporting Form for Services - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links DOL Internet Daily Report Employee Finder Image Library Microsoft end-user Enrollment Microsoft
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address https://ces.dgsdc1.bls.gov/content/cesform_h.asp?act=s
Go
Nonsupervisory Employees
(1) All Employees
(2) Women Employees
(3) Employees
(4) Payroll
(5) Hours
(6) Comment Code (optional)
Month
Jan 20 11 18 13680 720
Feb 22 12 20 16170 840
Mar 24
Your data were not su
nonsupervisory emplo
hours per worker in Fe
Please:
-- Verify the Nonsuper
-- Verify the Nonsuper
-- If both are correct,
column 6 that expla
Submit
Clear
(05) All Employees is less than Women ...
All Employees is less than Women Employees.
Please check your All Employees number in column 1 and your Women Employees number in column 2. All Employees is less than the Women Employees figure.
OK
Done


CES Reporting Form for Services - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links DOL Internet Daily Report Employee Finder Image Library Microsoft end-user Enrollment Microsoft
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address https://ces.dgsdc1.bls.gov/content/cesform_h.asp?act=s
Go
Nonsupervisory Employees
(1) All Employees
(2) Women Employees
(3) Employees
(4) Payroll
(5) Hours
(6) Comment Code (optional)
Month
Jan 20 11 18 13680 720
Feb 22 12 20 16170 840
Mar 24 14 22 11905 616
Significant change in Average Worker Hours
Your data were not submitted because the average hours worked per nonsupervisory employee (column 5 divided by column 3) changed from 42 hours per worker in February to 28 hours per worker in March.
Please:
-- Verify the Nonsupervisory Employees figure in column 3
-- Verify the Nonsupervisory Employee Hours figure in column 5
-- If both are correct, choose the most appropriate comment code in column 6 that explains this change
Submit
Clear
Done
Internet