Web-based Data Collection in the Current Employment Statistics Survey

Richard J. Rosen, Christopher D. Manning, and Louis J. Harrell, Jr.

Christopher D. Manning, BLS, 2 Massachusetts Avenue, N.E., Ste. 4860, Washington, DC 20212
Manning_C@BLS.GOV

Disclaimer: Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics

Key Words: E-mail, Internet, Information Superhighway, World Wide Web, WWW, Establishment Surveys

Introduction: Since 1995, the Current Employment Statistics (CES) Survey has tested the use of World Wide Web based data collection. CES is a monthly panel survey of about 380,000 businesses conducted by the Bureau of Labor Statistics (BLS). The Web collection system has evolved from a "proof of concept" to a production prototype, to a full production system. The current system runs on a secure server and is password protected. The system uses e-mail for various respondent contact purposes such as advance notice to report and nonresponse follow-up. The collection instrument contains embedded logic and range edits to help ensure the accuracy of the data. Thus respondents participate directly in the edit reconciliation process. CES currently has about 100 Web reporters.

This paper compares the Web collection system to Touchtone Data Entry (TDE), reviews the development of the system, describes the methods used to enroll units into the system, and discusses monthly response rates. We also discuss how the editing component has improved the data quality.

Background on CES Survey: CES is a monthly establishment survey of employment, payroll, and hours data. CES data, widely viewed as a major economic indicator, are published each month after only two and a half weeks of collection. This restricted collection period places a huge burden on the collection method (Harrell, 1996).

Traditionally CES data were collected by mail. However, in 1984, in an effort to improve response, we began examining alternative collection methods. During the last 14 years we have developed and instituted a variety of automated collection methods which have improved response, improved data quality, and reduced costs.

We are well on our way to a complete transformation to automated collection methods. Only 56,000 respondents (out of the 380,000 sample) still report via mail. Touchtone Data Entry collection currently accounts for 240,000 reports, while another 25,000 are collected by FAX. The rest of the reports are collected by a variety of methods including Electronic Data Interchange (15,000 respondents), Computer Assisted Telephone Interviewing (5,000 respondents), and other electronic media. Web data collection is the next step in the evolution of the automated data collection process. It is quick, has relatively low marginal costs, and improves data quality by allowing the data to be edited before submission.

Comparison of TDE and Web Methodology: The survey methodology of TDE and Web data collection is somewhat similar. Both methods start with a sample control file, rely on monthly advance notice and non-response prompting messages as reminders to report, and have a data-reporting component.

Touchtone and Web data collection starts with a sample control file containing respondent contact information such as name, address, phone number, and FAX number. The Web sample control file also contains the respondent's e-mail address along with the normal contact information. As the collection cycle begins, both collection methods depend on monthly advance messages to remind respondents to report. Research done for the development of TDE has demonstrated the efficacy of advance notice messages for maintaining response rates (Rosen et al., 1991). TDE respondents receive a monthly advance notice reminder message by FAX or postcard. Web respondents receive an e-mail message.

When TDE respondents are ready to report, they call an "800" toll-free number that activates a data-entry interview session. They use the numerical digits on their telephone keypad to respond to the questionnaire. Each entry is read back for respondent verification. A toll-free phone number is provided for problem reporting and inquiries. Instead using their telephone, Web respondents "surf" to a secure Web site and, after being validated as a CES reporter, access the on-line collection form to enter the data using their computer keyboard. The collection form is written in a combination of Hypertext Markup

Language (HTML) and JavaScript. It contains an image of the questionnaire, survey instructions, and hypertext links to definitions. An e-mail address is provided for problem reporting and inquiries. Once the requested data items have been entered, the respondent clicks the "Submit" button. Before submission, automated edits are performed on the data and any failures are noted on the screen. Data that pass the edits are immediately transported to BLS.

Direct editing of data is a distinctive characteristic of Web self-reporting. Data validity edits have been implemented using JavaScript. For example, if the respondent enters an All Employees figure less than the Nonsupervisory Employees figure, an error message is generated. The respondent turns the error message off, enters the corrected information, and submits the corrected report.

Finally, as the monthly deadline for data collection approaches, TDE respondents receive a phone call, fax, or postcard as a second reminder to report. The choice of mode depends on the size of the firm. Web reporters receive an e-mail message.

Touchtone has been our most successful collection method to date as over half of the current CES sample (240,000 out of 380,000) is collected by TDE. Since the majority of the sample has been transitioned to TDE, timeliness of collection and data quality have increased while collection costs have decreased. We feel that the success of Web collection will mirror and ultimately surpass that of TDE because Web collection contains all of the strengths of telephone collection methods while at the same time eliminating many weaknesses. For example, it allows the user to enter data using an intuitive, graphically interesting interface. Based on anecdotal evidence, we believe the visual interface may reduce cognitive burden on respondents compared to phone based reporting methods.

Web has substantially lower costs than TDE. Cost reductions are obtained by the elimination of toll charges incurred through supplying a toll-free number for respondents to call. Further cost reductions are obtained through sending advance notice and non-response prompts via e-mail as opposed to phone calls, postcards, and FAXes. And automated editing lowers costs while at the same time improving data quality.

We have also attempted to find ways to provide positive feedback to our respondents. Our current Web system provides hypertext links to other related Web sites, giving the respondent access to survey data products. This type of feedback is more difficult to provide using phone based reporting methods.

Hardware and Software Evolution: BLS has been testing Web-based data collection since 1995. In 1995, a "proof of concept" prototype was fielded using Sun Sparc 10 server. The National Center for Supercomputing Applications HTTPD server software was used. The system was written to work with the Mosaic browser. UNIX script files were used for data manipulation.

In 1996, the system was moved to an NT server. The server software was upgraded to Netscape's Secure Commerce server. A digital ID from Verisign, Inc., was implemented to upgrade the security of the site. The digital ID is an important technology that allows the user to confirm that they are reporting their data to BLS. The site was designed for use by browsers supporting HTML 3.0 tables.

Encryption is an important component of the NT based Web data collection system. The connection from the respondent's PC to the server is encrypted using the Secure Sockets Layer protocol. When the data arrive at the server, they are temporarily reencrypted. A collection moves the encrypted data inside the BLS firewall, decrypts the data, and reformats it for use by the estimation system.

In 1997, data validity edits were implemented. The JavaScript edits were written to be accessible both by Netscape and Microsoft browser users. In addition, a test version of the system was developed for the Microsoft Internet Information Server.

In 1998, the server operating system was upgraded to NT 4.0. The new operating system is intended to offer improved protection from potential denial of service attacks.

Capacity: Our current server is configured to offer 18 concurrent threads, meaning that we have the ability to allow up to 18 users to simultaneously report data. Given that the average length of a Web interview is about 2 minutes, how many cases could we collect on our current hardware without degrading performance?

Standard models exist for estimating the capacity of telephone systems (Bajorek, 1998). The Erlang B model is one of the most common. Two parameters are needed to use the model: the percentage of calls that receive a busy signal because all lines are in use, and the amount of traffic received during the busiest

hour of the busiest day. These values can be easily derived from our TDE experience. We would expect that respondents using the Web would exhibit the same reporting patterns.

Using the Erlang B model, assuming a 1% blocked call rate and 17% of call minutes being received in the busy hour of the day, our server can handle 1,835 cases per day. If we assume the same reporting patterns for Web respondents, the server could handle 10,794 cases per month. This capacity should be sufficient for the next few years.

Enrollment of the Web Sample: Approximately 100 respondents report each month using the Web collection system. This base of Web reporters has grown gradually as the Web system has evolved from a production prototype into a full production system.

1996 Enrollment: During April - November of 1996, as the system reached the production prototype stage, a sample of 514 CES TDE respondents was contacted by telephone and questioned about their Internet capabilities and their interest in reporting on-line. We contacted touchtone respondents because we had their contact information on file and because they were already familiar with the CES survey.

The touchtone respondents were screened on criteria such as their access to the Web, use of a compatible browser, and e-mail abilities from their computer desktop. Those that were qualified were further screened on their interest in converting to Web reporting. Those that were qualified and interested were considered eligible for Web conversion and were mailed a specially designed 'Web Conversion' information packet that explained how to report their data on-line. They were also given a password and were asked to begin reporting via the Web once they had reviewed the information packet. About half of those sampled (51%) were Computer and Data Processing businesses because it was expected that those types of businesses would be the most likely to have Internet and e-mail access. Other service industries were also contacted (17%) to see if characteristics were similar. Finally, based on a study by the Bureau of the Census that found interest in reporting via the Internet among government respondents (Sweet and Russell, 1996), State and Local Government establishments (32%) were also contacted.

As expected, the Computer and Data Processing businesses netted the largest number of Web-eligible respondents (See Table 1). Approximately 14% of those businesses contacted were qualified and interested in reporting data on the Web. While none of the other Services businesses we contacted were eligible, 10.4% of the State and Local Government establishments contacted were qualified and agreed to convert to on-line reporting. Overall, 10.7% of all respondents contacted were converted to Web reporting.

Table 1. Web Sample Enrollment Results

	Number Responded		Percentage Converted to Web Reporting	
Year	**'96**	**'97**	**'96**	**'97**
Computer & Data Processing Services (SIC 737)	265	44	14.3%	22.7%
Other Service SICs	86	72	0%	27.8%
State/Local Government	163	169	10.4%	7.7%
Total	**514**	**285**	**10.7%**	**16.0%**

1997 Enrollment: As the Web Data Collection system evolved into a full production system, we contacted another 535 TDE respondents to inquire about their Internet capabilities and interest. This took place in August and September of 1997, about a year after the first phase of enrollment. The sample make-up was similar to the first phase, including Computer and Data Processing businesses (17%), other Services (28%), State and Local Government (54%), and other industries (1%).

A mixed mode of contact was used in this second phase. E-mail was the desired method of contact; however, its use was restricted to 5% because of a limited number of e-mail addresses on file. Therefore, we chose FAX as the main contact mode (84%) because of its prevalence, speed, and low cost. The remainder of the sample (11%) was contacted by telephone.

For those contacted by FAX or e-mail, a specially designed message was sent that enabled respondents to screen themselves on Web eligibility. The reporters were asked to respond if they had Web and e-mail access on their desktop PC, if they could send and receive e-mail outside of the company, and if they had Netscape 2.0 or MS Internet Explorer 3.0 or above as their Web browser. Respondents were informed they could respond by e-mail, FAX, or

telephone. Of the 535 respondents contacted, 285 responded. The results are summarized in Table 1. As before, the Computer and Data Processing businesses netted a number of Web-eligible respondents (22.7%). But it wasn't the largest percentage this time. Surprisingly, almost 28% of other service establishments contacted were qualified and interested to begin reporting data on the Web. Of the State & Local Government establishments contacted, 7.7% were Web eligible and agreed to convert to on-line reporting. Overall, 16.0% of those respondents contacted were converted to Web reporting.

Our results lead to some interesting observations. The percentage of respondents that were eligible and agreeable to report on-line increased by 50% from 1996 to 1997 (10.7% to 16.0%). This is not surprising. Industry projections are that Internet usage doubles every year, which it has since 1988 (Anderson, 1997). We only enroll respondents who have Web access from their desktop PC. We could obtain an additional 2% increase to the conversion rate by allowing participants to use the Web from another PC. It should be noted that most CES respondents are clerical support staff with the firm. This may account for the apparent low proportion of respondents with Web access. However, given that Web access is rapidly growing from a small base, firms appear to moving from a consensus that Web access should be provided to only certain personnel.

FAX and e-mail solicitation proved to be highly effective modes of contact. Both are significantly less expensive than a phone call, making them a viable tool for Web enrollment. We also found a high correlation between having an e-mail address and having Web access. A late-1995 in-house study found little linkage between the existence of e-mail and Web access. However, now that Web is ubiquitous, if a frame of e-mail addresses can be obtained, e-mail solicitation can be a highly effective means of enrolling new Web respondents.

Client Side Editing: In April, 1997, we incorporated on-line edit reconciliation into the Web production system. The edits, written in JavaScript and HTML, are performed on the client side and precede any data submission.

Before the introduction of the client side edit checks, data submitted via the Web were reviewed by a data specialist for basic edit errors before being used in any estimates. If an error was found, the data specialist would call the respondent by telephone in order to reconcile the problem. Now, the basic edit functions are performed on-line before the data are submitted, eliminating labor-intensive edit and reconciliation activities. When the respondent clicks the 'Submit' button, data integrity edit checks are performed instantaneously. If no errors are detected, the data are immediately submitted. If a problem is detected, an error message is generated and appears to the respondent in the form of a pop-up box (see Figure 1). The error message is expressly tailored to point out which data items need to be reviewed and corrected, such as *"Please check your All Employees number in column 3 and your Women Employees number in column 4. All Employees is less than Women Employees".*

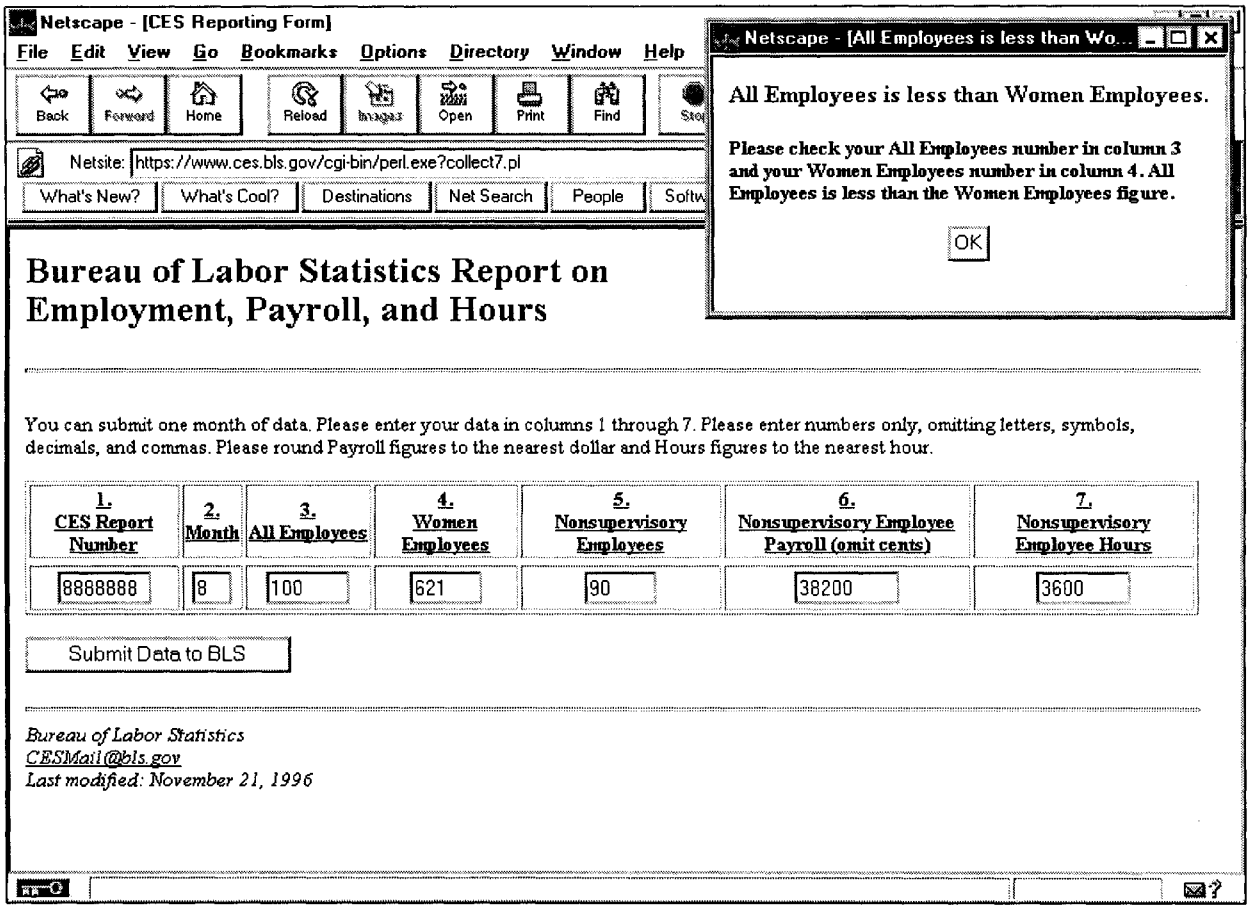
A button is provided so the respondent can close the pop-up box after reading the error message; however, the edit failure box will close automatically after approximately 10 seconds. Once the message box is closed, the respondent can correct and re-submit their data.

The current edits check for data validity, logic, and ranges. The validity edit checks ensure that all data entry is numeric and that mandatory fields (such as the month and the number of All Employees) are not blank. The logic edit checks test the relationships between the data items to ensure that all logic is maintained. For example, it is not possible for the number of women employees or production workers to be larger than the number of all employees because women employees and production workers are subsets of all employees. The range edit checks ensure that the month figure is between or equal to 1 - 12, and that the average hourly earnings figure (payroll divided by hours) is within a basic acceptable range.

As of August 1998, approximately 4% of all Web reports fail one edit check each month. This is basically the same as the edit failure rate seen when these edits are applied to TDE data. Since their integration into the system in April 1997, 30 different Web reporters (30% of the Web sample) have experienced 43 total edit failures. However, most errors are concentrated among a few reporters; ten respondents (out of the 30) have committed more than one error and account for 53% of all errors to date. There is no common quality that these ten respondents share besides the fact that eight of them are in the same industry (Services).

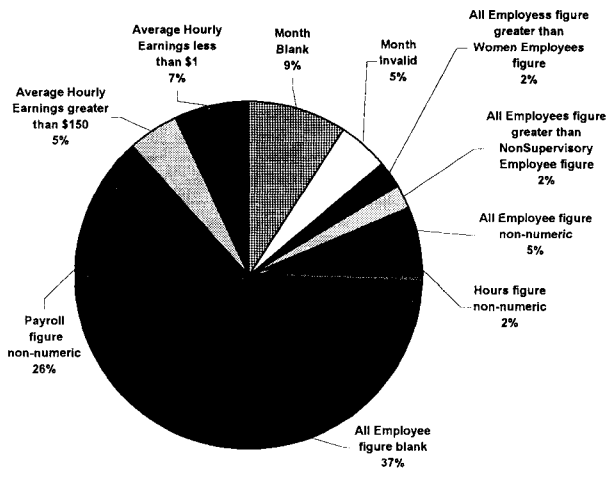
In 88% of all edit failures, the respondent was able to correct and submit the data on their own during the same session. Another 3% returned to the Web site later in the day and submitted a corrected report.

Figure 1. CES Collection Screen with Edit Note



The remainder submitted an e-mail message describing the problem and was contacted by a staff member to resolve the problem. Figure 2 breaks down the frequency of occurrence of edit failures to date.

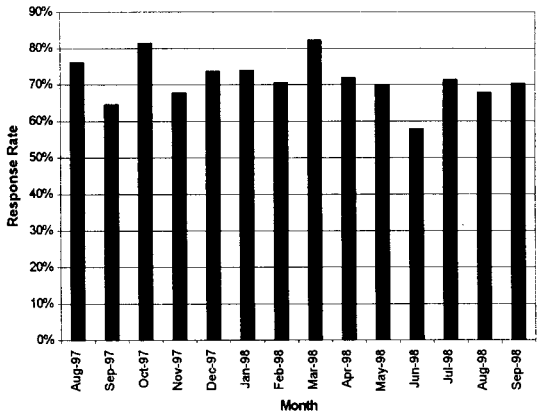
Figure 2. Edit Failure Frequency of Occurrence



The most frequent mistake the respondents make is leaving a mandatory field blank. The next largest problem is a non-numeric figure in the payroll entry. This is probably a case of the respondent entering a dollar sign, comma, or decimal point. In terms of actual data logic and range errors, about 16% of the records that were flagged as an edit failure did not pass the basic logic or range edit checks (such as Women Employees greater than All Employees or Average Hourly Earnings less than $1).

Response Rates: We have been tracking response rates achieved through Web based collection. The response rates are similar to those we have obtained in TDE collection. Figure 3 shows response rates as of the deadline for publication of the preliminary monthly estimate since January, 1997. Response rates for preliminary estimates have usually ranged between 60 and 70 percent, with the average initial response rate totaling 68%.

Figure 3. Web Response Rates: Preliminary Monthly Estimates



The Future of Web Data Collection in the CES:
The CES will continue to collect data by multiple modes for many years to come. We expect the percentage of the CES sample that is collected by Web to grow significantly in the next few years. We now have the infrastructure and methodology in place for large scale use of Web-based data collection. The biggest current obstacle is respondent access to the Internet from their desktop. If Web access continues to grow exponentially, this will become less and less of an issue.

As the Web sample grows, we will continue to research and implement many system enhancements. For example, we would like to integrate automated generation of the e-mail messages with the collection system. Other future enhancements include longitudinal editing (measuring month-to-month changes in data), use of streaming video for user help, possible use of intelligent agents such as the Microsoft Office animated help characters, implementation of enhanced e-mail messages, and video support for help desks. We would like to enhance the e-mail messages to include icons that would facilitate access to the server, and could provide video and audio clips to respondents.

Eventually we see the Web and TDE systems integrated into one, single system. A single sample control file will record the type of messaging required for each respondent (such as FAX vs. e-mail). A standard data record will be produced and uploaded to the estimation system.

An additional area where Web technology could be useful is in Electronic Data Interchange (EDI). EDI is the electronic submission of data records from machine to machine in a standardized format. Currently EDI relies on private networks for transmission of data. Private networks offer security but charge a fee for their use. However, the Internet is relatively free, and Web servers are now available that support secure file transfer over the Web. This will allow EDI respondents and statistical agencies to further cut costs by merging these reporting techniques.

References:

Anderson, Christopher (1997). "Doubling Games" *The Economist*, May 10, 1997. See also, http://www.economist.com.

Bajorek, Chris (1998). "Ask Dr. CT: How to properly calculate the number of lines your CT system requires." *Computer Telephony*, May 1998, pp. 112-114. See also, http://www.erlang.com.

Clayton, R.L., and G.S. Werking (1998). "Business Surveys of the Future: The World Wide Web As A Data Collection Methodology", *Computer Assisted Survey Information Collection*, in print, pp. 543-562.

Harrell, Louis, R.L. Clayton, and G.S. Werking (1996). "TDE and Beyond: Data Collection on the World Wide Web", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 768-773.

Rosen, R.J., R.L. Clayton, and T.B. Rubino (1991). "Controlling Nonresponse in an Establishment Survey." *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 587-591.

Sweet, Elizabeth, Russell, Chad (1996), "A Discussion of Data Collection Via the Internet", *Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 774-779.

Werking, G.S., and R.L. Clayton (1991), "Enhancing Data Quality Through the Use of Mixed Mode Collection", *Survey Methodology*, June 1991, **17**, No. 1, pp. 3-14.