TDE AND BEYOND: DATA COLLECTION ON THE WORLD WIDE WEB

Louis J. Harrell, Jr., Richard L. Clayton, George S. Werking, Bureau of Labor Statistics
Louis J. Harrell, Jr., BLS, 2 Massachusetts Avenue, N.E., Ste. 4860, Washington, DC 20212
Harrell_L@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

Introduction: Electronic mail (E-mail) and World Wide Web services are increasingly available within businesses and may be exploited for survey data collection. The Bureau of Labor Statistics has developed a prototype World Wide Web collection instrument for a pilot test of Web collection in the monthly Current Employment Statistics (CES) survey. Respondents receive electronic mail requesting that they enter their data in a World Wide Web page. The data are immediately edited and transmitted to the survey agency's computer.

This paper reviews current Web features relevant to survey data collection, describes the prototype CES Web collection system versus Touchtone Data Entry (TDE), and identifies considerations in the development of a Web survey data collection system. The strengths and weaknesses of Web collection are compared to other automated collection methods in terms of quality, timeliness, and costs.

Background: Each month, the CES survey collects employment, payroll, and hours data from a sample of almost 400,000 business establishments. The CES operates in a Federal-State cooperative system where each State collects, enters, edits and transmits data for the national estimates.

The CES data are published after only two and a half weeks of collection, placing an extreme burden on collection methods. Until the last few years, the CES was collected entirely by mail. Now the CES is in the midst of a complete transformation to automated collection using Computer Assisted Telephone Interviewing (CATI) as a transition to TDE collection. Over the past decade, the CES has developed methods to obtain high response rates for preliminary estimates. The CES now focuses on a process of offering additional features designed to streamline operations, improve quality and reduce costs. The development of Web collection is a natural part of this evolutionary process.

Why Use the Web? The Web offers an intuitive interface, low cost, and a standard, easily managed data record format. Web collection embodies all of the strengths of telephone procedures while at the same time eliminating many of the weaknesses. It allows the user to enter data using an intuitive, visually interesting interface. Cost reductions are obtained through automated editing, also allowing improved data quality. Links to other related sites can be provided, giving the respondent access to survey data products.

Other approaches were considered and rejected. Initially, we contemplated using an E-mail system, where the respondent would enter comma delimited data items. The E-mail message would be parsed and the data record would be extracted. The attraction of this idea is that E-mail is more prevalent than Web service. However, it does not have the uniformity and user-friendliness of the Web. Another idea was the use of diskettes, or Computerized Self Administered Questionnaires (CSAQ). This was rejected on the basis of cost and timeliness. Thousands of diskettes would have to be distributed and processed each month.

Web Survey Methodology Compared to TDE: Under TDE, respondents receive a monthly advance notice message sent via postcard or automated out-bound FAX. This message replaces the arrival of the survey form as a reminder. Data collection is performed by dialing the TDE system and entering data as requested by the digitized verbal prompts. Lastly, non-respondents receive telephone or FAX prompts on specially designated days conforming to the availability of their payroll records (Werking, 1991).

The CES Web survey collection cycle parallels the TDE respondent contact process. It begins with a sample control file containing the respondents' E-mail address in addition to the normal respondent contact information of name, address, and phone number. The collection form is a standard "Web page" containing an image of the questionnaire, survey instructions, definitions, and hypertext links to definitions. An E-mail address is provided for problem reporting and inquiries. As the collection cycle approaches, the respondent opens their E-mail to find a reminder,

"surfs" the net to the CES homepage, accesses the data collection screen, and fills in the requested data. The moment the respondent clicks the "submit data" icon, the data are transferred to the survey agency. Schedules are checked-in and, at predetermined time periods, E-mail nonresponse reminder messages are sent.

The TDE method minimizes labor-intensive activities for mail-out, mail-back, and data entry. However, it does not directly addresse another expensive activity: data editing and reconciliation. Our current labor-intensive editing and reconciliation operations can be directly handled under Web collection. The respondent will address all edit failure questions through on-line edits generated immediately after data entry. This change will allow the elimination of the large semi-clerical operations of staff poring over reams of computer rejected data and attempting to "correct it" or label it as unusable.

We can implement both longitudinal and data integrity edits in the Web environment. Integrity edits are based exclusively on rules, while longitudinal edits require immediate access to several months of previously reported data. Security considerations become important if historical data are located behind a firewall. Persistent client state cookies can be used for implementing comparisons with previous month data if the length of the cookie is kept under 4 kb. The cookie is a text file that can be used for storage and retrieval of data from the client. We have developed prototype integrity edits and are developing cookie based longitudinal edits. Under Web methodology, most survey data collection operations can be fully automated and the overall process simplified for both the survey agency and the respondent.

Total Design Method On-line: The eventual replacement of traditional methods with the Web will require a careful review of all mail-based research. The results serve as reasonable starting points for Web methodology. Under TDE, high response rates have been attained using a combination of advance notices, easy to use data entry interfaces, and carefully-timed nonresponse prompts. Will Web methodology work the same? Also, the Total Design Method (TDM) offers a rigorous approach to maximizing response rates (Dillman, 1978). Under the TDM, each survey feature (prenotification message, the survey instrument, reminders and the timing of each) carries potential for improving response rates. Will Web collection behave similarly to mail with regard to these? Will the response rate increases seen be commensurate under Web? How does forms design research carry over into research on screen design and human-computer interface? These, among other questions will be evaluated in the CES Web pilot tests. Table 1 outlines the salient characteristics of automated data collection techniques. Each of these features can be manipulated in order to maximize response rates.

Table 1. Features of Automated Collection Methods

Feature	TDE	CATI and CAPI	CSAQ	Web
Advance Notice Message	Yes	Varies	Varies	Yes
On-line Entry	Yes	Yes	Yes	Yes
Questionnaire Branching	Yes	Yes	Yes	Yes
On-line Editing	No	Yes	Yes	Yes
Visual Interface	No	Yes	Yes	Yes
On-line Help	No	Yes	Yes	Yes
On-line Access to Survey Data Products	No	No	No	Yes
Non-Response Prompting	Yes	Yes	Yes	Yes

Web Versatility: Unlike telephone collection methods, Web collection can accommodate a wide range of surveys and survey operations. The use of telephone collection procedures is often limited by the length and complexity of the questionnaire, the frequency of the collection cycle, and the need to immediately respond to an interview question. Like telephone collection, Web collection can easily accommodate changes in questionnaire content.

Questionnaire Length--CATI is limited to surveys which could be conducted within a 20-minute session and is problematic if respondents need to refer to their records. TDE is limited to the number of items for which a respondent is willing to push buttons. The Web, however, has the ability to accommodate structured questionnaires of any form or length including "form-layout" designs or traditional "question-by-question" designs. The respondent has the ability to refer to records as frequently as needed or to partially complete the questionnaire and return to it at a later time with no noticable effect on costs.

Survey Frequency--Ongoing Web surveys are easy to maintain if a file of contact information, including the E-mail address is used. One-time multi-mode surveys are more difficult to implement as a complete file of E-

mail, mail, and phone addresses may be difficult and expensive to obtain.

Altering Questionnaire Content--Web collection has more flexibility than mail in accommodating content changes such as new data items or survey supplements. The Web system can be modified and loaded at a single point. Once loaded, all respondents have immediate access to the modified software. The telephone collection operations of CATI and TDE are more limited since they require an immediate answer for the new data item during the interview and this may not be possible if the respondent needs to refer to his/her records. Web questionnaires may offer calculation worksheets, whereby the respondent can enter portions of answers which would be automatically calculated into the final response. Mail surveys require a staff for printing, forms distribution, and mailing.

Costs: Over the decades we have invested large sums of money to develop and refine the labor-intensive operations which help ensure the quality of our estimates. These operations include: collection and collection control, multiple modes of nonresponse follow-up, key entry with verification, and editing with reconciliation of all failures. However, under Web reporting, all collection activities can be fully automated and centralized using a dedicated LAN system. Messages are electronically sent at predetermined dates and information checked-in on a flow basis. On-line edits are implemented as part of the Web data collection session.

The cost-effectiveness of Web collection is difficult to fully measure at this time; however, enough is understood about the economics of software to come to some general conclusions. Most analysts point to the fact that software has a high fixed cost for development and very low marginal costs for adding a new user (Anderson, 1996). In contrast, conventional production assumes that producers face decreasing returns to scale. That is, it costs money to add new users of a product, and that these costs will increase to the point where it is no longer is profitable to produce output. Software's increasing returns to scale can be applied to applications using the World Wide Web.

For organizations purchasing unlimited Web access, the average cost of a session should approach zero, as the constraint on Web usage would be the capacity of a telephone line, renting for a fixed charge. Telephone line rentals could be spread over a large number of users, minimizing data transmission unit costs. Under other collection methods, efforts are always made to keep respondents' costs to a minimum by providing pre-paid postage, or toll-free telephone service. Using a TDE system, the respondents call a toll-free number to gain access to the system. The technology also exists for providing "800" number access to the Internet. Servers can be equipped with dialup access utilizing toll-free numbers. This approach could be used if the pricing paradigm for Internet access ever changed.

CES unit costs of data transmission are shown in the following table. The table illustrates the dramatic impact on costs obtained from moving to TDE and Internet collection (Clayton and Harrell, 1989).

Table 2. CES Unit Costs of Data Transmission

Function	Mail	TDE/FAX	Internet
Out	$.32	$.08	$.00
In	$.34	$.16	$.00
Nonresponse Prompting	$.10	$.04	$.00
Total	$.76	$.28	$.00

Product and Customer Service Improvements: The improvements offered by automation and electronic communication will ultimately lead to simplified respondent reporting, more accurate microdata, more timely responses, and improved customer access to our survey products.

Simplified Reporting--Web collection offers significant opportunity for simplified reporting. A well designed, intuitive interface can minimize or eliminate the need for reading documentation. Any interaction needed because of edit failures can be resolved during the data collection transaction, minimizing interruptions to the respondent's schedule.

Accuracy--Web collection can offer improved accuracy compared to telephone methods. The respondent is able to see all data displayed prior to submission, providing an additional opportunity to review the data for any errors. Response rates should also increase since nonresponse prompting can be handled on a far more timely and controlled basis, making the process less vulnerable to publication cut off dates. The unit-cost per schedule will be significantly reduced by the elimination of postage and mail processing, and by significantly reducing telephone charges for edit, prompting, and collection calls. This reduction in unit cost can then be redirected towards increased sample size to reduce the level of sampling error for the survey or directed towards other quality-enhancing activities.

Timeliness--Our customers will benefit from more timely data. For some surveys this will mean "final"

estimates will be quickly available and thus will eliminate the need for "preliminary" estimate surveys, or for others, a reduction in the size of revisions between preliminary and final estimates. Some surveys may be able to publish their data with only a very limited time-lag making the data more relevant to current economic conditions, while others may be able to increase their publication frequency from annual to quarterly or quarterly to monthly.

Customer Service--There will also be many benefits in terms of information dissemination. Our respondents will benefit since we will be able to provide to them, also electronically, a profile of their firm's information against national (or State) industry averages derived from the survey results; examples include employment trend, earnings data, work week hours, and overtime. This will allow the participating firms to directly follow the performance of their firm against current trends in their industry, thus adding an extra benefit for participating in the survey.

In addition to providing products back to our respondents, electronic communication will provide all users with quick, easy, and cost-effective access to our survey products. Instead of waiting long periods for press releases or periodicals, calling or writing for specific tables, or purchasing specialized diskettes, users will have Web access to our large longitudinal public-access databases. This will significantly reduce the labor-intensive overhead associated with our information dissemination activities while providing improved services to the users.

TDE methodology provides the respondent with a form for record keeping and additional documentation on how to use TDE. For respondents who wish to continue using paper documentation, printable copies could be provided on the Web server.

CES Model for Web Collection: The entire Web environment is rapidly changing. Features which were not available even a year ago are entering the marketplace on a daily basis. Each new advance in hardware, software, and communications represents new opportunities and challenges.

In 1995, the CES program developed a "proof of concept" model of a Web data collection system. This prototype was implemented on a Sun Sparc 10 workstation, using Solaris 2.4 UNIX. The server software we chose was the National Center for Supercomputing Applications HTTPD Version 1.4. The site was developed for Mosaic browsers.

In 1996, we moved to a new configuration. The CES Web prototype system uses a Windows NT server, Netscape Secure Commerce Server, the Netscape browser, and a digital ID from Verisign, Inc. The respondent needs to have a browser that supports Hypertext Markup Language (HTML) tables and the Secure Sockets Layer (SSL) protocol. Browsers, such as Netscape's, can be obtained free. HTML tables are required for forms-based data entry while SSL provides security.

Security: Perhaps the single most critical feature of the Internet infrastructure is the security of the transmitted information. This limitation is repeated by every student of the Web and is drawing the attention of much of the computer community. The CES has met its goal of a C2 level of security. C2 refers to a security standard developed by the National Security Agency. Some characteristics of the Web security profile are:

- Authentication of the respondent
- Protection against snooping during transmission (Packet data security).
- Protection of the session (hijacking).
- Protection of confidential data once it has arrived at the server.
- Prevent non-BLS user access to the BLS LAN

Authentication: The Netscape commerce server also provides utilities for a relatively high degree of user authentication. To connect to our Web site, the user must have established a user name and password with our system administrator. After entering the site's address, the user enters their name and password. Upon verification, the user is given access to the login screen of the data collection system. The user again enters their user name, the name is checked against a database, and the appropriate collection screen is supplied. The double level authentication can be compared to a policy of accepting whatever is submitted in mail collection.

Snooping and Hijacking: Our prototype relies on a Windows NT server and the Netscape Secure Commerce Server software. Netscape is preferred because of its widespread acceptance in the corporate world, and its support for the SSL protocol. The SSL protocol relies on public key encryption to ensure the integrity of data transmissions. The user is assured that their Web session is protected against eavesdropping and tampering, and that they are actually interacting with the survey agency's Web site. NT is preferred over UNIX due to its capability to tightly control access to directories on the server.

Protection of Data on the Server: We intend to further enhance site security through use of added layers of encryption. SSL encrypted data are received and momentarily stored in an encrypted format on the server. An automated polling agent brings the data inside the firewall, where it is decrypted and processed.

Protection of BLS LAN: BLS relies on firewall software for protection of its LAN.

CES Pilot: In April 1996, we began collecting CES data from 7 firms. These respondents were selected from firms reporting by TDE and in Services SICs. By September, the sample had expanded to 38 firms. Respondents were contacted to determine whether they have access to the Web, their E-mail address and their willingness to participate. Eligible units received a specially developed package describing Web reporting. They were asked to try the system within the next 2 days and CES interviewers would call them back. They are expected to continue to report using the Web. Periodic follow-up will assess their reactions to this method, a source of feedback for ongoing improvements in the user interface. For subsequent months, messages will be sent providing an advance notice reminder that it is time to report. Nonresponse prompting messages will also be sent to respondents that are late in reporting their data.

Results: We are steadily improving our rate of conversion to Web reporting. As of September, we were able to convert 12.1% of the units we contacted. We believe it is too early to provide any estimate of an upper bound to the number of Web eligible respondents. Respondent comments have been universally favorable. A typical response has been: "Very fast and easy to use. Better than I expected." Table 3 summarizes our results as of September 1996. SIC 737, Computer and Data Processing Services was selected first for solicitation, as we believed that respondents would be more familiar with the Internet. The other services industries were then contacted to see if characteristics were similar. Finally, State and Local Government was studied as the Bureau of the Census has found interest in reporting through the Internet among government reporters (Sweet and Russell, 1996).

Our results lead to some interesting observations. While many companies have Web access, not all staff have access to the Web. Our solicitation criteria for Web reporting are relatively strict. We only enroll respondents who have Web access from their desktop PC and have at least Netscape 2.0 browsers. We could obtain an additional 3% increase to the conversion rate by allowing participants to use the Web from another PC. Respondents who preferred another reporting mode did not want to convert to Web because they had recently converted to TDE.

In 1995, BLS asked a non-scientific panel of firms a series of questions regarding E-mail availability and use. Comparing these results to our current results shows that there has been an increase in the availability of E-mail access. In 1995, 7% of firms could send E-mail outside of their company (Clayton and Werking, 1995). In 1996, the number increased to 36%.

Table 3. Web Sample Solicitation Results as of September 1996

	SIC 737 n=313	Other Service SICs n=121	State and Local Govern -ment n=264
E-mail only	5%	0%	2%
E-mail and Web, not on desktop	19%	9%	6%
Compatible browser, E-mail/Web on desktop	12%	0%	6%
Prefer other mode	3%	0%	1%
Out of business, no E-mail/Web	45%	62%	47%
Unable to reach	16%	29%	38%
Total	**100%**	**100%**	**100%**

Plans: Over the next year, we plan to implement many system enhancements to our Web site. We will expand the industry coverage to allow all industries to use Web reporting. Client side editing will be introduced. We currently have prototype Perl script files which implement data integrity edits. Our preference is to have a Java applet that would do client side editing. In addition, we would like to integrate automated generation of nonresponse prompting and advance notice messages with the collection system. An automated link to our production database is also needed. We are researching approaches that would allow us to conduct longitudinal editing of data. Enhancements to the user interface are also needed. We would like to implement cursor control, a feature that is not currently available with HTML. Regarding security, we will be researching ways of implementing layers of encryption to better enhance communications security.

We are also planning some methodological improvements. Nonresponse prompting messages will have an icon attached which will allow users to link to

our Web site from within the E-mail message. This feature requires an E-mail system that will process mail file attachments using methods compatible with the BLS E-mail system. Not all Web reporters will be able to receive the icon.

The use of video offers a broad area for research, drawing on knowledge about respondent-interviewer interaction. A quality enhancing activity that we are studying is the use of streaming video for special surveys and general information. Streaming video technology allows video to be transmitted across the Internet at relatively slow speeds, such 28.8kb per second. If the respondent has a sound card in their PC, they can also receive audio. Special surveys could be introduced and explained with a video clip. Other items such as videos of the BLS Commissioner's monthly testimony to the Joint Economic Committee of Congress could also be included. We would need a suitable video server and the respondent would need a browser that could support video.

Video conferencing technology could also be tested for respondent help applications. Internet based video conferencing can be introduced at a low cost per work station. Currently, the picture is not the same quality as commercial television, but the software is continually improving.

EDI and the Web: The CES is using Electronic Data Interchange (EDI) to collect data from respondents with large numbers of establishments. Approximately 14,000 establishments are collected from 28 respondents. The standard approach to EDI collection is to transmit the data using a private network. However, private networks charge a fee for their service while the Internet is relatively free. The economic benefit of using the Internet will eventually cause EDI and Web reporting techniques to merge. We envision an EDI respondent linking to a Web server and securely transmitting their datafiles. The data would be encrypted and a polling agent would move the data inside the BLS firewall.

Multi-Mode Integration: The essential production activities supporting Web, TDE, and Voice Recognition will be integrated into a single system. A single sample control file will record the type of messaging required for each respondent. A standard data record will be produced and uploaded to the estimation system.

Conclusions: We have just begun our tests and have developed some preliminary observations. Web access is not universal for our respondents. In fact, it is very limited. For those with access, they can receive e-mail, respond to e-mail advance notice prompts, and respond to non-response prompting. E-mail non-response prompting is not as effective as TDE non-response prompting. Research will look at the impact of graphics and text in non-response messages.

The combination of TDE self reporting with a graphical interface offers a powerful, promising tool for high quality, low cost data collection.

Acknowledgements: The authors would like to thank Chris Manning, of BLS, for his help in the preparation of this paper.

REFERENCES:

Statistical Policy Working Paper 19 (1990); *Computer Assisted Survey Information Collection*, Office of Management and Budget.

Werking G.S., (1994) "Establishment Surveys: Designing the Survey Operations of the Future", *Proceedings of the Section on Survey Research Methods,* Invited Panel on the Future of Establishment Surveys, American Statistical Association, pp. 163-169.

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

Anderson, Christopher. "A World Gone Soft: A Survey of the Software Industry", The Economist, May 25, 1996. http://www.economist.com/surveys/software

Dillman, D. A. (1978) Mail and Telephone Surveys: The Total Design Method, New York: Wiley-Interscience.

Clayton, Richard L., Werking, George S. (1995) "Using E-Mail/World Wide Web For Establishment Survey Data Collection", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 527-532.

Clayton, Richard and Harrell, Louis (1989), "Developing a Cost Model of Alternative Data Collection Methods: Mail, CATI and TDE," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 264-269.

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