UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/2005/14
6 April 2005

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-third plenary session
(Geneva, 13-15 June 2005)

USABILITY AND ACCESSIBILITY TESTING IN SUPPORT OF INTERNET REPORTING

Invited paper submitted by the U.S. Census Bureau*

INTRODUCTION

1. Developing questionnaires for Internet reporting requires a broad understanding of potential respondents as well as an understanding of the criteria for successful interaction with online forms. Potential respondents to an online Census form bring with them a wide range of human capabilities and limitations. To make the online form as usable and accessible within program constraints, the U. S. Census Bureau conducts usability and accessibility testing and redesigns the data-collection form, as necessary. Working with test participants, the professional staff of the Census Bureau's Usability Laboratory continually refines a set of procedures and techniques based on best practices documented in the literature (e.g., Dumas & Redish, 1999). This paper focuses on our testing procedures, with examples and lessons learned from our experiences in developing and testing a series of prototype Internet forms for the decennial Census.

2. In the context of responding to an online questionnaire, we define usability as the extent to which the software user interface supports the respondent's effective, efficient, and satisfying completion of an electronic form (cf. ISO, 1998). Usability is, thus, a multidimensional construct. We consider accessibility to be a facet of usability.

* This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author (Elizabeth D. Murphy) and not necessarily those of the U.S. Census Bureau.

3. Accessibility means that persons with and without disabilities are able to gain access to and use online applications, Web sites, and other products of electronic information technology to a comparable extent. In the U.S., we have Federal regulations governing what it means for software to be accessible (U. S. Access Board, 2000). A sticking point, however, is that technically accessible software may not support a successful and satisfying experience for the respondent. Internet respondents need an online form to be both usable and accessible, no matter what their abilities or disabilities.

ITERATIVE USABILITY AND ACCESSIBILITY TESTING

4. Usability testing helps to reveal aspects of an Internet site that can be simplified or otherwise improved to make the form-filling task easier and more satisfying for respondents. Observation of respondent behaviors during laboratory or field testing can reveal usability issues, such as excessive scrolling, convoluted navigation paths, unexpected system responses, confusing and/or inconsistent conventions, and other hard-to-predict effects.

5. Testing with representative respondents is necessary because actual respondents will invariably try to make the software behave in ways that it was not designed to behave. It is not possible for designers, developers, or even usability specialists to act as surrogate respondents because they know too much; they are too familiar with the design of the user interface and too knowledgeable about the designer's concept of progression through the form. Test participants must be people who have had absolutely no prior involvement in pre-design or design and development activities.

6. Accessibility testing using automated tools identifies violations of legal requirements, e.g., in the U. S., Section 508 of the Rehabilitation Act of 1973, as amended (29 U.S.Code §794d). The programmers' code for the site is examined by the automated tool, which diagnoses violations and suggests revisions. Some of the potential violations identified need to be checked manually by an analyst, for example, to evaluate readability in gray scale, readability without frames, and the level of screen flicker for Javascript elements.

7. Accessibility testing with participants who have disabilities is necessary because meeting the legal requirements does not guarantee that the software is truly usable by persons with disabilities. As we have found in our own testing, software can be technically accessible, but not usable for such participants, cf., Theofanos & Redish, 2003, p. 38: "Meeting the required accessibility standards does not…necessarily mean that a Web site is usable for people with disabilities. And if a Web site is not usable, it is not really accessible, even if it has all the elements required by law."

8. Testing should be iterative for at least three reasons: 1) to verify that changes made have actually resolved the issues they were designed to resolve; 2) to determine whether the changes have raised any new issues, and 3) to evaluate new functionality that may have been added since the last round of testing. We try to conduct at least two iterations of testing over the course of user-interface development. More would be better. Frequent usability and accessibility testing throughout the design and development cycle can eliminate the need for expensive retrofitting in late phases of the cycle.

U.S. CENSUS BUREAU'S USABILITY LABORATORY

9. After the development of a detailed test plan, usability testing generally takes place in the Census Bureau's Usability Laboratory. The lab includes three testing rooms and three observation and recording consoles. These consoles are located in an observation room adjacent to the testing rooms (Exhibit 1). The test participant sits in one of the three testing rooms, facing a one-way glass and a wall-mounted camera, under a ceiling-mounted camera. Two microphones pick up sound in each testing room.

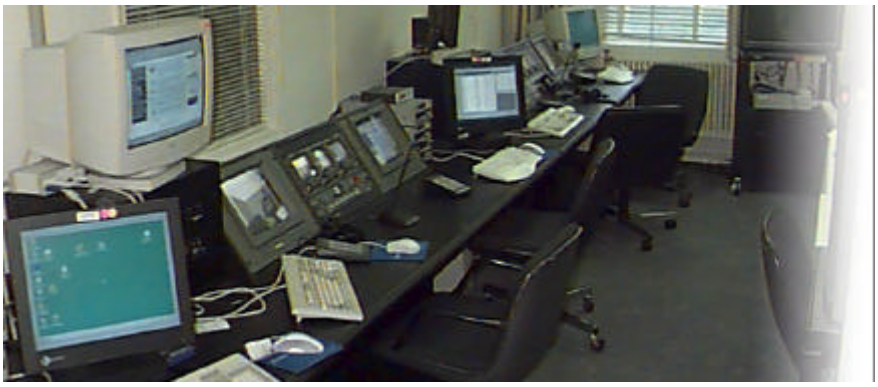

Exhibit 1. View of observation and recording area, Census Bureau Usability Laboratory

(a) The lab has microphones and headsets for the test personnel as well as logging equipment and software, video cameras, scan converters, and digital recording equipment. The observation room has monitors at each console, which allow observers a clear view of the participant's screen. Headsets are available for up to four observers.

(b) We sometimes have the test administrator sit with the test participant in the testing room. This is a more natural situation if the test participant is thinking aloud.

(c) The Usability Lab includes an area for running automated accessibility tests using the Insight/InFocus software (SSB Technologies, 2004) and a screen reader, Job Access With Speech (JAWS) by Freedom Scientific (2004). The Insight software diagnoses violations of the Federal regulations, and the InFocus software recommends coding changes to correct the violations. JAWS allows the accessibility tester to determine, for example, whether the screen will be read in an order that will make sense to a blind respondent.

(d) JAWS is also installed on the computers in the Usability and Accessibility Lab. When participants with visual impairments are helping us evaluate a user interface, they can be accommodated in one of our regular testing rooms. We equipped the computers with magnification software and other customizing software, based on the preferences of visually impaired participants in recent evaluations of the 2005 Census Internet form.

RECRUITING TEST RESPONDENTS

10. We attempt to recruit test participants who resemble the actual respondents in their demographic characteristics and their Internet usage. We aim for diversity in age, education, race, and ethnicity. We try to recruit equal numbers of males and females for a particular study to control for any gender-based differences in responding. In testing a Census Internet form, we require that test participants report at least one year of prior experience using the Internet because actual respondents are expected to be frequent users of the Internet.

(a) For testing in our laboratory, we recruit participants from the Metropolitan Washington, District of Columbia (D.C.) area (including the city of Washington, D.C. and nearby counties in Virginia and Maryland). We have been successful in advertising for test participants in a daily newspaper that is passed out to local subway riders, "The Washington Post Express." This advertising has generated hundreds of calls to our recruiter, who administers a screening questionnaire to prospective test participants.

(b) We have agreements with two remote testing sites, one in Texas and one in California. We provide recruiting requirements to the site managers, who conduct recruiting in their local areas. We have not used our remote facilities in testing the Census Internet forms because of concerns about the security of the transmitted data.[1] However, the site at the University of Texas-Pan American has been certified as secure for collection and transmission of protected data. We plan to use this site in future tests of Census Internet prototypes to expand the diversity of test participants' demographics.

(c) Since it is important that test participants have no prior exposure to the software being tested, we try not to use Census employees as participants. Occasionally, as in dry runs, however, we do use individuals from our clerical or technical support staff as participants. Dry runs help us evaluate and hone the test procedures and materials. Data generated by dry-run participants are kept separate from the data provided by actual test participants.

(d) Some test participants with visual impairments come from within the ranks of Census Bureau employees and contractors who have had no prior connection with the project. Others are recruited through contacts with organizations for the blind and visually impaired. We plan to expand our recruiting to include persons with motor/manual disabilities and cognitive disabilities, all of which are covered by the Federal regulations on accessibility of electronic information technologies (U.S. Access Board, 2000).

(e) We generally aim for 10 to 12 participants for each version of a Web site or application to be tested. We find that after the eighth or ninth participant we rarely identify "new" issues.

MATERIALS AND PROCEDURES FOR TESTING

11. Prior to a testing session, we have tried providing test participants with a brochure about the experience they will have in the lab, e.g., what they can expect to happen. We have found, however, that it is difficult to get the brochures to the participants in a timely manner. We have tried e-mailing an electronic version of the brochure to scheduled participants a day or two in advance of their session. This has worked well with Census employees and contractors, for whom we have e-mail addresses, but not so well for external participants. Obtaining active e-mail addresses is a challenge. In any case, we always orient test participants to what they can

expect to happen during a test session when they arrive at the lab.

12. Materials needed for testing typically include a consent form, a general introduction to the test's purpose and procedures, a questionnaire on computer and Internet experience, tasks for participants to perform, a satisfaction questionnaire, and a set of debriefing questions. Our typical procedure follows the same order in which the materials are described in the following paragraphs. We videotape all sessions unless the participant prefers not to be taped.

(a) The consent form has several purposes: 1) to inform the participant about the circumstances of the test, especially any hazards associated with it; 2) to obtain consent for videotaping and limited use of the videotapes; and 3) to assure the participant that the testing has been approved at higher levels of the U. S. government (the Office of Management and Budget, an agency of the executive branch).

(b) The general introduction provides further background and details on the purpose of the testing. A major objective here is to assure the participant that we are not testing his or her skills or abilities. Rather, the participant is helping us evaluate the software. We want the participants to relax and not to feel that they are being judged.

(c) The questionnaire on computer and Internet experience allows us to ask more detailed questions about these topics than it is possible for the recruiter to ask during a screening interview. For example, we ask about the participants' access to computers at home and at work; we ask how many hours they typically spend on the Internet each day; and we ask how often they fill out Internet-based surveys. Self-reported ratings of experience with computers and the Internet are given on a nine-point scale, where 1 equals "no experience" and 9 equals "very experienced." These data help us extend our knowledge of each participant.

(d) Once a test participant has completed the initial paperwork, we typically give the person practice in thinking aloud, since they will be asked to provide a running commentary during their session. Thinking aloud provides access to the participant's expectations, their general thought process as they perform tasks, and their rationales for decisions they make about navigating through the online form. A typical practice task is to find the local weather forecast on the Internet. It can be informative to see how people go about performing this task. For example, if the participant has no clear strategy for finding the weather forecast, it is reasonable to assume that the person is a relatively novice Internet user. The test administrator encourages the person to keep thinking aloud if they pause. This practice session is not scored.

(e) With the preliminaries completed, the test participant is ready to begin the first task. In the case of the Census Internet form, the tasks are to gain access to the form itself (i.e., to complete the authentication process); to complete the questions; to resolve any edit messages; to review and correct data entries, if necessary; and to submit the completed form.

(f) Throughout a testing session, a note taker uses logging software to develop a textual record of the participant's interaction with the software. Using a pre-defined set of behavior codes, the note taker enters appropriate codes, e.g., for navigational actions, changes to answers, confusion, and participant comments. Descriptive statistics (e.g., frequencies, means, standard deviations) are generated for each code, for each participant and across participants.

(g) Immediately following submission of the form, we ask the participant to complete a short version of the Questionnaire for User Interaction Satisfaction (Norman, Shneiderman, Harper, & Slaughter, 1998). This is a rating scale where 1 equals extremely low and 9 equals extremely high on the attribute being rated. The following are a few examples of attributes and anchor terms:

- overall reaction to the electronic questionnaire (Terrible/Wonderful);
- information displayed on the screens (Inadequate/Adequate);
- going back to previous questions (Impossible/Easy).

(h) Next we conduct a debriefing session using pre-formulated questions as well as questions that have been triggered by the particular participant's testing experience. The prepared questions allow us to collect data from all participants on the same issues. During the debriefing session, we often go back through the completed form with the participant and ask about interactions we observed during the session. Thus, we use both think-aloud and retrospective methods in working with test participants.

TESTING THE CENSUS INTERNET FORM FOR USABILITY AND ACCESSIBILITY: LESSONS LEARNED

13. The primary purpose of usability testing is to identify aspects of the user-interface design that lead to less than successful and satisfying completion of the Internet form. Success and satisfaction are usually defined by the project team and documented in a set of usability goals. Identified issues provide a basis for design recommendations. Accessibility testing has a dual purpose: a) to verify that the software complies with the Federal regulations, and b) to evaluate the extent to which persons with disabilities are able to complete the Internet form. In testing the Census Internet form, we have focused on "showstoppers," i.e., design factors that make it impossible for the participant to complete the online form. Constraints on time and resources limit the design team's chances of resolving less critical issues.

(a) The first Internet form used by the Census Bureau was developed in-house for the 2000 Census. Software development for the Census 2000 Internet form focused on providing functionality and supporting the electronic receipt of data from up to 8,000,000 expected respondents. The development team followed a Rapid Application Development (RAD) approach (e.g., Maner, 1997), which produced minimal documentation (i.e., no documented requirements or usability goals). Two rounds of usability testing were conducted in the Census Bureau's Usability Laboratory, and some recommendations on the user-interface design were implemented (Murphy, Marquis, Nichols, Kennedy, & Mingay, 2001). Change control consisted of an informal, undocumented process by which the developer assessed the impacts of proposed changes and implemented the changes that he deemed cost effective. In

general, changes that were implemented addressed functional showstoppers. From this experience, we learned that it is preferable to have a formal change control process in place.

(b) Use of the Census 2000 Internet form to submit household data required successful navigation of an up-front authorization process for the user to gain access to the form itself. A major step in this process was the user's entry of a 22-digit housing unit identification number, referred to as the Census Identification Number (CIN) (Figure 1). We attempted to make usability testing as realistic as possible by preparing mailing packages for our test respondents. One of the test respondent's first tasks was to locate the Census ID number and enter it into the seven parsed fields, as shown in Figure 1. By observing the test respondents, we learned they had little trouble entering their Census ID numbers; but they did have trouble finding the numbers to begin with. The numbers were printed on the back of the paper form underneath a bar code. To help sighted respondents find their numbers, the developer designed a graphic (Figure 2). For Census 2000, this graphic resided in the help files. A customer satisfaction survey run during the actual Census data-collection period in 2000 indicated that very few actual respondents accessed the graphic (Murphy & Stapleton, 2002).

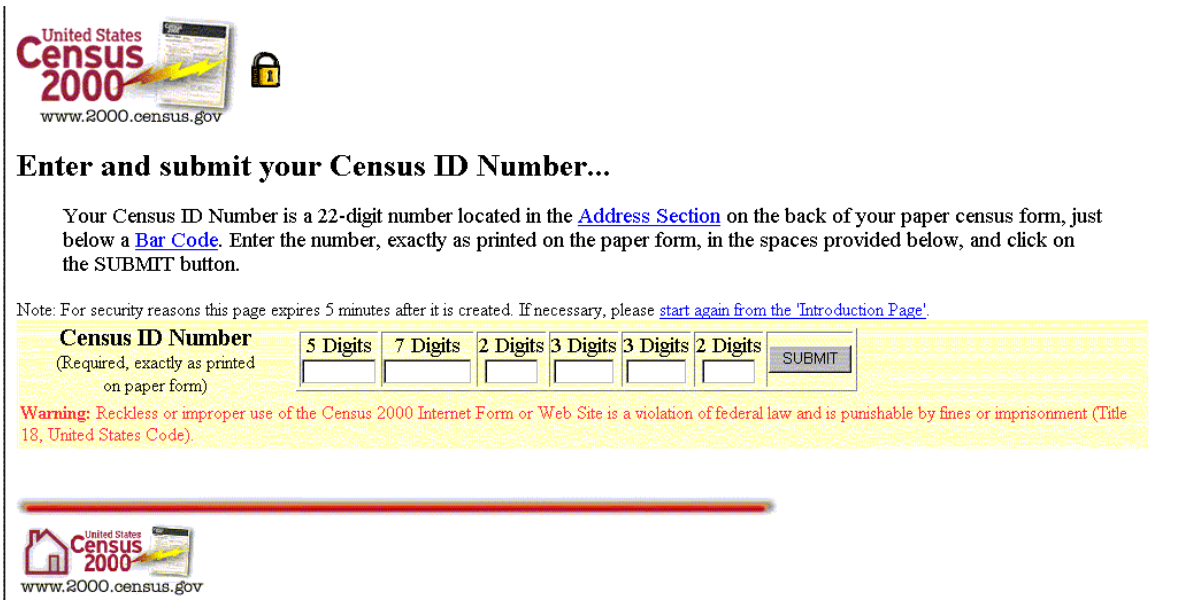

Figure 1. Census 2000 Internet form: Census ID number (CIN) required

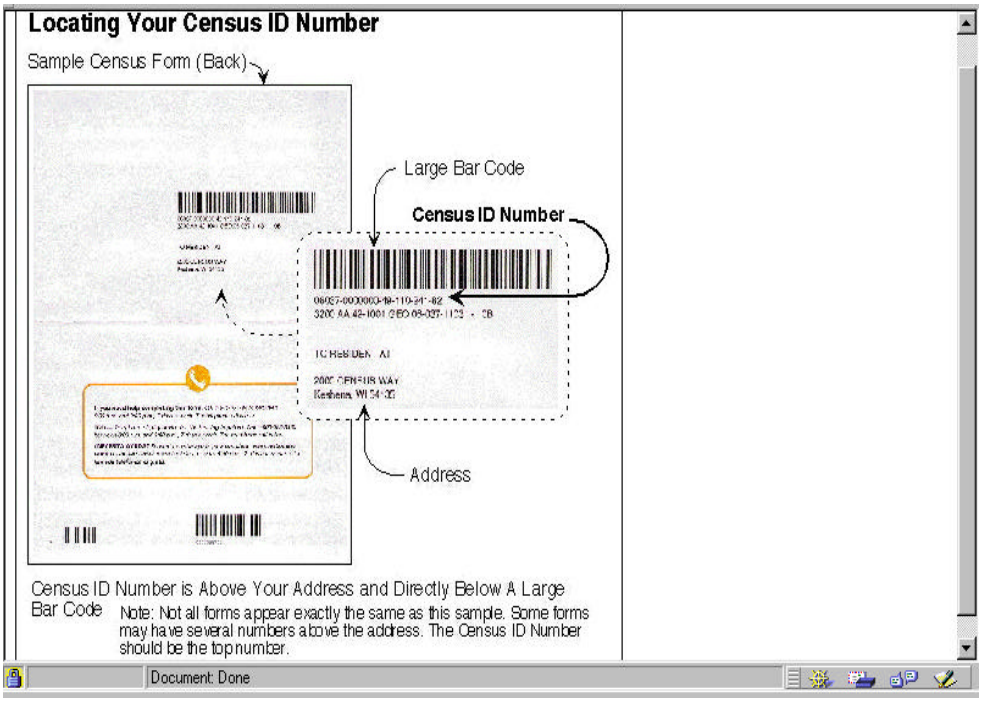

Figure 2. Graphic provided in Census 2000 to help respondents find their Census Identification Numbers on the back of the paper form

(c) In a later prototype of the Census Internet form, developed for the 2003 National Census Test, a simplified version of the graphic was displayed to the right of the data-entry fields where the respondent was asked to enter the CIN. The number of digits in the eye-readable portion of the CIN was reduced to 14, although the complete CIN remained at 22 digits. This page layout has been largely retained for the 2005 National Census Test, as shown in Figure 3; but the full number of digits has been decreased to 18 (all eye-readable) and parsed into four fields. In the 2003 version, the 14 digits were parsed into three fields, and several test participants thought they were being asked for their Social Security Number, which is parsed into three fields (e.g., 111-00 -1111). This was a showstopper for some participants and a potential showstopper for actual respondents. As expected, the recent testing of the 2005 version did not find any participants attempting to enter their Social Security Number. The lesson here is that participants are likely to jump to unwarranted conclusion based on the visual design of the user interface as they interpret it in the context of their prior experience.

14. In Census 2000, accessibility testing of the Census Internet form took place during a team member's visit to the National Federation for the Blind (NFB) in Baltimore, Maryland, with the Census 2000 Internet form on floppy disks. The NFB had legacy versions and the newest versions of all the major screen readers, which were used to test the form's accessibility. Even though the code had been written to the latest standard of HTML Transitional 4.0, the screen readers at NFB could not easily interpret the Census Internet form. At that time, the usability lab was not equipped for accessibility testing. We soon acquired the necessary software and sent a staff member for training in its use.

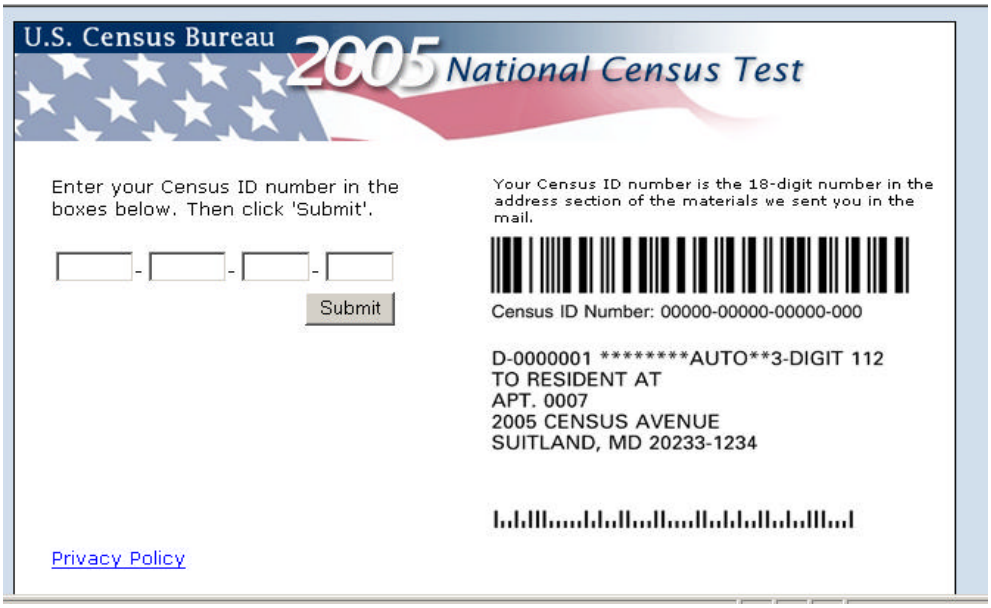

Figure 3. Prototype screen for entry of CIN in the 2005 National Census Test: 18 digits parsed into four chunks

15. In later versions of the Census Internet form, we evaluated the user interface in our own laboratory with JAWS, Insight/InFocus, and with participants who were blind or visually impaired. In testing the online form for the 2004 Overseas Enumeration Test, we learned that none of the visually impaired participants was able to complete the form, enter data efficiently and accurately, or navigate efficiently or effectively, even though the site was in compliance with the Section-508 regulations. However, these usability goals were met for sighted users, who also reported quite high levels of satisfaction with the user interface in both rounds of testing.[2] We have found that users who complete the form tend to report high satisfaction even if they have experienced navigational or other problems during the test session. This is consistent with findings reported in the literature (e.g., Andre & Wickens, 1995). A lesson learned is that evaluating user satisfaction is necessary, but not sufficient in evaluating usability.

16. We have learned from working with visually impaired users that only users with no vision at all rely on screen readers. Users with varieties of low vision prefer to use magnification techniques and customized foreground/background combinations. In recent accessibility testing of the 2005 Census Internet form, we asked test participants, in advance of testing, how they preferred to customize the display system, and we provided their preferred magnification or other software for use during testing.

A BRIEF CASE STUDY: PARTICIPANT REACTIONS TO DYNAMIC EDITS

17. Among the requirements that emerged after Census 2000 was a requirement for dynamic edit checks of data entered by participants. For example, after the participant entered an age and date of birth for a household member and pressed the Next button, the computer checked the age against the date of birth. If a discrepancy was found, the computer issued an edit message to the participant.

(a) In usability testing, we found that participants would not continue to the next question

unless they could resolve the edit messages, even though they were perfectly free to go on. In debriefings, participants reported their belief that they could not proceed unless they resolved the discrepancy pointed out by the edit message. To resolve this showstopper, we recommended changing the color of the messages from red to blue. Since red is a Western-hemisphere stereotype associated with danger or an order to stop, it needed to be softened. Blue was chosen for its lack of association with danger or traffic signals. After discussion, the team also decided to add text encouraging respondents to continue to the next question if they did not know the answer or had given their best answer to the current question. This design strategy seems to have served the purpose of giving respondents permission to proceed even if an edit message remains unresolved. We do not want actual respondents to break off their sessions because of a mistaken perception that they must resolve every edit message.

(b) In general, we have learned that people expect the computer to be more than a passive receiver of data. Based on their experience with commercial software and online surveys, they have come to expect the computer to perform calculations for them and to help them keep their data consistent by running edit checks for valid data. In recent usability testing, we have tried having the computer calculate age given date of birth, something that many participants in prior tests have suggested.

REPORTING RESULTS

18. We provide an initial report to the test sponsor within a few days of the end of testing. This report contains a list of prioritized findings (usability and accessibility issues) as well as our recommendations for resolving the issues. Priorities are based on the impact of the issues on respondents' successful task performance (High, moderate, low). Later, we provide a full report detailing our testing methods and including tables of descriptive statistics on the measures collected (e.g., task completion time, satisfaction ratings).

PROSPECTS FOR THE FUTURE

19. In industry, models of the software development lifecycle are moving toward the early and iterative integration of usability-and-accessibility engineering and testing (e.g., Addelston & O'Connell, in press). At the Census Bureau, usability is establishing itself as part of the culture, although it is still often thought of as something to be done near the end of development. Accessibility has gained attention since the passage of the Federal regulations in 2001. The ideal is usability and accessibility for all users and respondents, no matter what their abilities or disabilities. In the context of the 2010 U. S. Census, the objective of a design-and-development process into which both usability and accessibility are fully integrated from the beginning is to provide a functional, usable, and accessible Internet form for the collection of census data.

REFERENCES

Addelston, J. D., & O'Connell, T. A. (in press). Integrating accessibility into the spiral model of the software development life cycle. To appear in *Proceedings of the International Conference on Universal Access in Human-Computer Interaction (UAHCI)*, Las Vegas, Nevada, July 2005.

Andre, A. D., & Wickens, C. D. (1995). When users want what's NOT best for them. *Ergonomics in Design*, 4, 10-14.

Dumas, J. S., & Redish, J. C. (1999). *A Practical guide to usability testing*. Portland, OR: intellect.

Freedom Scientific. (2004). *JAWS*® 5.0. St. Petersburg, Florida: Freedom Scientific Blind/Low Vision Group (http://www.freedomscientific.com).

ISO (International Organization for Standardization). (1998). *Ergonomic requirements for office work with visual display terminals (VDTs)* – Part 11: Guidance on usability (ISO 9241-11). Geneva, Switzerland: International Organization for Standardization.

Maner, W. (1997). *Rapid application development*. Retrieved January 15, 2005, from http://csweb.cs.bgsu.edu/maner/domains/RAD.htm

Murphy, E., Marquis, K., Nichols, E., Kennedy, K., & Mingay, D. (2001). Refining electronic data-collection instruments and data-collection tools through usability testing. *Research in Official Statistics, 4* (2), 23-33.

Murphy, E. D., & Stapleton, C. N. (2002). Evaluation of users' experience of the Census 2000 form.. Presented at the *International Conference on Questionnaire Development, Evaluation, and Testing Methods* (QDET), Charleston, South Carolina.

Norman, K. L., Shneiderman, B., Harper, B., and Slaughter, L. (1998). *Questionnaire for User Interaction Satisfaction*. College Park, Maryland: University of Maryland, Human-Computer Interaction Laboratory.

SSB Technologies. (2004). *InSight 4.0 /InFocus 4.0*. San Francisco, California: SSB Technologies, Inc.

Theofanos, M. F., & Redish, J. (2003). Guidelines for accessible – *and usable* – Web sites: Observing users who work with screenreaders. *interactions*, 10 (6), 38-51.

U. S. Access Board (Architectural and Transportation Barriers Compliance Board). (2000). *Electronic and information technology accessibility standards*. Washington, DC: Federal Register (Dec. 21). Retrieved January 15, 2005, from http://www.access.board.gov/secton508/508standards.htm

[1] We have used these sites for testing the Census Bureau's home page (www.census.gov) and various data-dissemination tools (e.g., the American FactFinder).

[2] First round (N = 7: range = 2 – 9, mean = 8.14, standard deviation = 1.21; Second round (N = 6): range = 4 – 9, mean = 7.83, standard deviation = 2.04); scale anchors: 1 = very low, 9 = very high.

* * * * *