

# USABILITY TESTING OF AN INTERNET FORM FOR THE 2004 OVERSEAS ENUMERATION TEST: ITERATIVE TESTING USING THINK-ALoud AND RETROSPECTIVE REPORT METHODS

Kent L. Norman

Laboratory for Automation Psychology and Decision Processes  
Department of Psychology, University of Maryland  
College Park, Maryland 20742-4411

Elizabeth D. Murphy

U. S. Census Bureau, Statistical Research Division  
Washington, DC 20233

An Internet form for the U. S. Census Bureau's 2004 Overseas Enumeration Test was evaluated in two rounds of usability testing. Participants were assigned to one of two conditions: Think-Aloud, in which they talked about what they were doing; or Retrospective-Report, in which they completed the form and then talked about their experience while viewing a recording. Participants also completed follow-up tasks. Sessions were video taped and logged. Round 1 testing identified 28 usability issues. Round 2 testing found that 13 of the issues had been resolved following design changes made to the interface. Round 2 testing identified 21 new and continuing usability issues. Results suggest that changes made to the interface increased the likelihood that respondents would be able to successfully complete the form. Task completion times in the think-aloud condition were only slightly longer than they were in the retrospective condition, while retrospective reports required a substantial amount of added time.

## INTRODUCTION

Usability testing is an important part of software development and user acceptance (Neilson & Mark, 1994; Shneiderman & Plaisant, 2003). According to John and Marks (1997), the two central questions are "How effective is usability testing in identifying and remedying interface problems and how do different methods of testing compare in terms of their efficiency and effectiveness?"

A usability study was conducted that employed both the think-aloud and retrospective report methods in two rounds of user testing. This design allowed for a comparison of the two testing methods and provided a test of the effectiveness of changes made by the developer to the interface following the first round. Prior to testing, a set of qualitative and quantitative usability goals was agreed upon. These goals pertained both to the overall system and to specific screens and data entry. We evaluated the site against the criteria of an 80 percent achievement of the goals during both rounds of testing.

Sessions were video taped and logged for subsequent analysis. In addition, a number of follow-up scripted tasks were added to evaluate less frequent actions and the types of errors caused by unusual data entry. Finally, participants rated their subjective satisfaction with the interface using the Questionnaire for Interaction Satisfaction (QUIS) (Norman, Shneiderman, Harper, & Slaughter, 1998).

## Internet Form

In preparation for the 2010 Census, the U. S. Census Bureau conducted a test of the feasibility of counting American citizens living abroad using an Internet form. The production version of this form was launched for overseas testing in February 2004. The on-line form was similar to the paper form, but added navigation options and validation of data entry. Figure 1 shows one of the screens. Based on the number of individuals that the respondent enters for the household count, the system displays person tabs in the left hand margin. Respondents could page from screen to screen using the "next" and "previous" buttons at the bottom of the screen. After a screen had been visited via the "next" button, the respondent could jump to that screen using the tabs in the top margin.

Each time the respondent moved to another screen, the server checked the input for blank fields and valid input (e.g., month from 1 to 12). If a blank field or an error was encountered, the system gave an edit message as illustrated in Figure 2. Edit messages were only displayed on the first attempt to leave a screen and were not triggered on the second attempt so that users could leave information blank and would not be trapped by errors they could not correct.

There were 11 unique screens: a welcome screen, a household count and address screen, a screen for the respondent (Person 1), seven screens per person, a screen for adding additional people, and a review screen.

The 2004 Overseas Enumeration Test  
Easy, Important, Confidential.  
USCENSUS BUREAU

1. Household    2. Persons    3. Review/Submit    FAQ's | HELP

Name	Citizenship	Identification	U.S. Address	Sex/Age	Origin	Race
------	-------------	----------------	--------------	---------	--------	------

### Persons - Person 1 - Name

Please answer the following questions for each person living in this household. Start with the name of any adult living here. We will refer to this person as Person 1.

**What is Person 1's name? Enter name below.**

First Name      Middle Initial      Last Name

          

To navigate within this form, use the Previous and Next buttons, do not use your browser controls for Back and Forward.

    

---

Respondents are not required to respond to any information collection unless it displays a valid approval number from the Office of Management and Budget (OMB). This test has been approved and given the following OMB number: 0607-0903.

USCENSUS BUREAU

Figure 1. Example screen for entering the name of Person 1 in the household.

The 2004 Overseas Enumeration Test  
Easy, Important, Confidential.  
USCENSUS BUREAU

1. Household    2. Persons    3. Review/Submit    FAQ's | HELP

Name	Citizenship	Identification	U.S. Address	Sex/Age	Origin	Race
------	-------------	----------------	--------------	---------	--------	------

### Persons - John - Sex/Age

**What is John's Sex?**

Male  
 Female

**What are John's age and date of birth?**

Age on April 1, 2004

February only has 29 days. Please verify the date of birth you entered. If you have provided your best answer, or if you don't know this person's date of birth, please continue to the next question.

*Enter numbers in boxes.*

Month    Day    Year of birth

      

    

---

Respondents are not required to respond to any information collection unless it displays a valid approval number from the Office of Management and Budget (OMB). This test has been approved and given the following OMB number: 0607-0903.

USCENSUS BUREAU

Figure 2. Screen for entering sex, age, and date of birth showing an example of an edit message.

## Think-Aloud vs. Retrospective Report

Two different procedures were used in this test: think-aloud and retrospective report. The purpose of using both methods was to offset the disadvantages of one method with the advantages of the other.

The main strengths of the think-aloud procedure are that participants can give their immediate reactions and comments on the user interface and describe any difficulties while experiencing them and the test administrator can immediately probe to clarify the cause of a problem. The main weakness of the think-aloud procedure is that people sometimes have trouble saying what they are thinking while focusing on what they are doing.

The main strengths of the retrospective procedure are that the test participant can complete the task without the distraction of commenting simultaneously and the time to complete the task is not confounded with time to think aloud. The disadvantages are that people are likely to forget what they were thinking by the time they go back and view the tape and the reporting session adds considerable time to user testing.

## Effectiveness of Changes

In general, usability testing helps to reveal aspects of a site that could be simplified or improved and make the task easier for users. Observation of respondent behaviors 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. The schedule for development of the Internet form included two rounds of user testing. After the first round of testing, the contractor made changes to correct a number of usability issues. The user interface was then re-tested to assess the effectiveness of the changes. Further changes were made based on the findings in the second round, and other recommendations were taken under advisement for future iterations.

## METHOD

### Participants

Two females and five males participated in Round 1, and three females and three males participated in Round 2. Seven participants were in the 25-45-age range, and six were in the 46-65 range. Five participants were Black or African American; one was Asian American; six were non-Hispanic Caucasian; and one white participant was from a Hispanic background. All reported that they used computers and the Internet.

### Facilities

Usability testing took place in the U. S. Census Bureau's Usability Laboratory. Participants sat in one of the testing rooms facing a one-way glass and a wall-mounted camera, under a ceiling-mounted camera, and in front of an LCD

monitor placed on a table at standard desktop height. Computers in the test rooms are equipped with session-recording software, which we used to support retrospective debriefings.

## Procedures

Each participant was asked to complete a questionnaire on computer and Internet usage. The test administrator then read aloud the introduction to testing and obtained the participant's consent to participate and be videotaped. After answering any questions, the test administrator gave the participant a card with a foreign address and telephone number and a list of passport and Social Security numbers.

The browser (Internet Explorer) was set at the welcome screen for the 2004 Overseas Enumeration Test. Participants were asked to complete the form using their own household data for up to three people, and were instructed to stop before submitting their data. If a participant preferred to change the names of household members or alter other information for privacy reasons, we allowed this.

If a participant was in the think-aloud condition, the test administrator provided a brief demonstration. The respondent was then given an opportunity to practice thinking aloud at the U. S. Census Bureau's Home Page ([www.census.gov](http://www.census.gov)). They were asked to provide a running commentary and to explain what was happening if they had difficulty completing the task. If a participant seemed to be pausing for an unusual length of time, he or she was asked to describe the situation.

In the retrospective condition, participants were instructed to complete the form without verbalizing except to say if they did not know how to proceed. They were told to complete the form using their own household data (up to three persons) and to stop before submitting their data. During a retrospective debriefing, the participant was shown the recording of the session. Participants were asked to describe what they were doing, what they expected in response to their actions, and whether they were ever surprised.

Following the think-aloud sessions and retrospective debriefings, the participants were asked to add a person to their household. Participants were then given a number of follow-up tasks that required them to go back and change or correct previously entered data. Finally, they were asked to complete the QUIS designed to measure their satisfaction with the user-interface and interaction design.

## RESULTS

Performance measures included accuracy of data entry and completion time. We expected completion times to be longer in the think-aloud condition, but we thought that they might not be significantly longer than those in the retrospective condition. Some researchers believe that concurrent thinking aloud adds significantly to the participant's task-completion time (e.g., Rhenius & Deffner, 1990). Although the average time per screen was longer for the think aloud (58 sec) than for the retrospective report (39 sec), the difference was not statistically significant. In addition, we found that verbalizations in the think-aloud

condition tended to be primarily running commentaries of actions whereas in the retrospective condition, they tended to be comments on user mistakes and points of confusion.

We evaluated the user interface against the qualitative usability goals by examining the frequency of participant confusion, negative comments, and errors. We obtained the means and standard deviations of the immediate satisfaction ratings by task.

Table 1. Performance and Ratings on Follow-Up Tasks. (Frequencies represent observed participants who performed the task.)

Task	Performance Efficiency (Both Rounds)			Ratings (Means and Standard Deviations)	
	Most	Less	Least	Round 1	Round 2
Add A Person	5	1	5	4.40 (2.41)	7.50 (1.29)
Jump Ahead to Sex/Age	Not possible			8.50 (0.58)	7.00 (2.28)
Go to Person 1 Race	4	4	1	8.80 (0.45)	7.67 (1.86)
Go to Person 2 Race	8	1	1	9.00 (0.00)	7.50 (2.81)
Go to Person 2 Passport	9	1		9.00 (0.00)	8.40 (0.89)
Go to Person 1 Passport	8	1		9.00 (0.00)	8.67 (0.82)
Go to Rev/Sub Screen	9			9.00 (0.00)	9.00 (0.00)
Edit Household Count	5	2	2	7.50 (1.91)	7.67 (3.27)
Edit SSN	7	2	1	9.00 (0.00)	9.00 (0.00)
Edit DOB	8		1	8.75 (0.50)	9.00 (0.00)
Edit Error in DOB	Comprehension Only			3.75 (2.99)	6.60 (2.51)
Leave Origin and Race Blank	5		5	6.25 (3.59)	5.17 (2.86)
Help on Navigation	4	5	1	8.00 (2.00)	6.00 (3.10)
Find Information	7	1	2	8.67 (0.58)	7.17 (3.13)
Submit Survey	5	5		6.00 (3.56)	8.67 (0.52)

Ease of navigation and data correction were the main issues in the follow-up tasks. Participants were sorted according to their judged efficiency of performance. Performance was judged as least efficient if the participant used the Next button repeatedly to get to a destination screen, instead of using the tabs and sub-tabs. For example, five participants used this method in follow-up Task 1, which

involved adding a person to the household, while another five used the most efficient method of clicking on the household count link in the summary table or clicking on the household tab to take them back to that screen. Table 1 shows these data and descriptive statistics on ratings of the perceived difficulty performing the tasks.

Finally, we analyzed the QUIS ratings by individual item and overall means. We examined the raw ratings for instances of participant dissatisfaction (any ratings below 7, i.e., below an 80 percent level of satisfaction on a 9-point scale).

In Round 1, only two of the general usability goals were met: (1) 88 percent of participants were able to complete the form successfully; and (2) the form supported efficient and effective navigation by 89 percent of the participants. However, the 80 percent goals were not met for (1) efficient, accurate data entry or (2) subjective satisfaction with filling out the form. In Round 2, all of the goals were met.

Changes recommended and implemented between Round 1 and Round 2 resolved thirteen of the usability issues (about 46 percent), including a high-priority issue and two moderately high issues. The usability issues observed in Round 2 were primarily continuing issues that had been identified in the Round 1 testing but not resolved, for whatever reason. Only two new issues surfaced in Round 2 testing.

## DISCUSSION

By comparing user performance against a set of both general and specific usability goals, we were able to identify a number of usability problems. The goals helped to set criteria that could be used in the evaluation of the interface in both rounds of testing. We recommend this procedure since it helps to inform the testing personnel about the critical issues to be aware of during user observation.

After the primary task of completing the survey online, a set of follow-up tasks were used to identify problems that would not normally arise in the primary task but that could cause problems to some users. Interface problems observed on these tasks were either fixed in the subsequent version of the interface or deemed by the design team to be infrequent or inconsequential enough to not be of concern. A thoughtful choice of follow-up tasks is recommended in user testing to anticipate interface problems that might not be observed in small samples. These tasks are usually designed to exercise (a) areas and options in the interface that might be less frequently accessed but that could cause substantial problems or (b) tasks users might attempt but that are not directly supported by the interface and that could lead to errors or termination of the survey.

The iterative method of user testing and redesign was successful. A number of problems identified in Round 1 were addressed in a redesign of the interface. These changes were shown to be successful in Round 2. We highly recommend the use of iterative testing and rapid redesign. It is particularly effective when the design team is responsive to the results of user testing and able to make changes in a timely manner.

Finally, a comparison of the think-aloud and retrospective report methods indicated longer, but not significant, task completion times for the think-aloud method. However, in

terms of overall time, the retrospective report added an average of 17 minutes to the testing time. Differences emerged in the type of verbalizations made. In the think-aloud method, users tended to read text on the screen and recited more of what they were doing rather than what they were thinking. In the retrospective method, as users viewed the recording, they tended to be silent if there were no problems and to explain errors and hesitations in their actions only when they occurred. In fact, the task administrators encouraged this by fast forwarding and pausing the recording as needed. Overall, however, we did not observe any substantial differences in the number or type of interface problems reported between the two methods. Consequently, we have no recommendation for one method over the other except to note that the retrospective method requires more time and technology than the think-aloud method does. In tasks that are cognitively demanding and time critical, the retrospective method would be preferred since it does not interrupt the flow of the user's information processing.

### REFERENCES

- John, B. E., & Marks, S. J. (1997). Tracking the effectiveness of usability evaluation methods. *Behaviour and Information Technology*, 16, 4, 188-202.
- Nielsen, J. & Mack, R. L. (eds.), (1994). *Usability Inspection Methods*, New York: John Wiley.
- Norman, K. L., Shneiderman, B., Harper, B., and Slaughter, L. (1998). *Questionnaire for User-Interface Satisfaction*.

College Park, MD: University of Maryland, Human-Computer Interaction Laboratory.

- Rhenius, D., and Deffner, G. (1990). Evaluation of concurrent thinking aloud using eye-tracking data. *Proceedings of the Human Factors Society 34<sup>th</sup> Annual Meeting* (pp. 1265-1269). Santa Monica, CA: Human Factors Society.
- Shneiderman, B. & Plaisant, C. (2003). *Designing the user interface: Strategies for effective human-computer interaction (4th Ed)*. Reading, MA: Addison-Wesley.

### ACKNOWLEDGEMENTS

This work was conducted by the U. S. Census Bureau , Decennial Management Division, with support from the Statistical Research Division. Appreciation is given to Census Bureau personnel (Kent Marquis, Suzanne Fratino, Jennifer Lins, Juan-Pablo Hourcade, Idabelle Hovland, and Patricia Montgomery) for their collaborative support and helpful reviews.

This report is released to inform interested parties of research and to encourage discussion. The views expressed on methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.