
Erica L. Olmsted-Hawala
US Census Bureau
erica.l.olmsted.hawala@census.gov

Elizabeth D. Murphy
US Census Bureau, (Retired)
bettymurphy495@yahoo.com

Sam Hawala
US Census Bureau
sam.hawala@census.gov

Kathleen T. Ashenfelter
US Census Bureau
kathleen.t.ashenfelter@census.gov

Abstract

We describe an empirical, between-subjects study on the use of think-aloud protocols in usability testing of an information-rich Web site. This double-blind study used three different types of think-aloud protocols: a traditional protocol, a speech-communication protocol, and a coaching protocol. A silent condition served as the control. Eighty participants were recruited and randomly pre-assigned to one of four conditions. With the goal of keeping unintended bias to a minimum, data analysis did not count the number of identified usability problems by condition, which was considered too subjective. Rather, the study collected the number of verbalized and non-verbalized counts of frustration by condition that users experienced. The study also did a count of the number of verbalized and non-verbalized instances of positive comments by condition that users expressed. Results show that there were no statistical differences in the number of counts by condition with respect to the traditional, speech communication, or coaching condition. The study concludes that simply counting the verbalizations of users by condition is not an effective measure when comparing protocols. Keywords: Think aloud, user testing, verbalization, usability testing.

Introduction

Web sites have fast become one of the primary tools that both private industries and public agencies have of disseminating information. As such, in a self-service society it is crucial that Web sites be usable. One primary way that developers can make their Web sites more usable is by incorporating user-centered design (UCD) strategies during the development cycle. A major facet of UCD is usability testing. There are many ways to conduct usability tests, and the methodology used in a study directly affects the results of the study. Rigorously researching the methods used in usability testing makes for better usability test results and when applied to Web sites, can also lead to improved communication in the self-service society in which we live. In this paper, we describe an empirical research study on the think-aloud (TA) protocol in usability testing of an information-rich Web site.

The TA protocol is one of the main techniques used by usability professionals in usability testing. In theory, when using a TA protocol during a usability session, users articulate where they are struggling and the potential reasons for the struggles. Usability practitioners use the TA protocol because they don’t know what a user is thinking. The usability practitioner uses the participants’ verbal data with other measures to identify problem areas of the Web site and to formulate recommendations. One of the more widespread TA protocols used by usability practitioners is concurrent TA, in which the participant is encouraged to “think out loud” while working on tasks.

The present study aims to provide practitioners with a better understanding of the strengths and weaknesses of variations of the TA protocol. We conducted a double-blind, empirical study of three different types of
concurrent TA protocols and one silent control condition. The study was carried out in the U.S. Census Bureau’s Usability Laboratory, using the Census Bureau’s public data-dissemination Web site (http://www.Census.gov). Participants were led to believe that they were participating in a usability study geared toward improving the Census Bureau’s Web site, but the actual purpose of the study was to investigate the effects of various TA protocols on user performance and satisfaction. We report the number of verbalized and non-verbalized counts of frustration by condition and the verbalized and non-verbalized counts of positive comments that users experienced.

**Think-aloud protocols**

Two of the most common TA protocols that usability practitioners engage in are:

- Concurrent TA where the participant is encouraged to “think out loud” while working on a task.
- Retrospective TA where the participant talks only after the session is completed, typically while watching a video of his or her session.

Within concurrent TA there are many variations, such as whether participants have a practice session, whether test administrators sit next to the participant, and whether test administrators use minimal or active verbal probing. This study examined concurrent TA and focused on the differences in user behavior when different amounts and types of verbal probing were used by the test administrator.

TA protocols are used widely by usability professionals; however, researchers often use different or non-standard methodology [1,2,3]. In addition, when researchers report their use of the TA protocol, many leave out details on the specifics. For example, they may not list the types of probes they used, how often probes were given, how long they waited before they probed, and so forth [4]. The basis for the protocol most cited in introductory usability texts [5,6,7] is the verbal protocol put forward by Ericsson and Simon [8]. Practitioners often ignore the strict constraints [1,2,3] required of the TA protocol. For example, Boren and Ramey [1] showed that when practitioners used what they believed was a “typical” TA protocol, the prompts differed greatly among the observed practitioners. Some practitioners used the simple “um-hum,” others used the extended “what do you think x button does,” and others used questions such as “is there anything in particular you’re looking for?” (p. 264).

User verbalizations can be categorized into three levels [8]. The first two levels require information processing in the participant’s short-term memory. The verbalizations that occur in the third level necessitate additional cognitive processing (such as explanations of thought processes, interpretations of information one is working with, and retrieving additional information from long term memory). In addition, third level verbalizations can also include outside influence, such as comments and probes from the test administrator which may distract the participant from what they ordinarily would have done. This type of interaction may redirect attention, interrupt the task flow or require participants to analyze what they are thinking and doing. Third-level verbalizations often require participants to access their long-term memory. Ericsson and Simon consider first and second level verbalizations as valid data, and third-level verbalizations as invalid to use in analysis.

In an informative review of the different studies that use verbal reports as data, Nielsen et al. [9] contend that the rigid restrictions that Ericsson and Simon describe for what are acceptable data are actually akin to working with an information processing system rather than a psychological being with emotions and senses, and who works within a context, not in isolation. This rigidity in accepting only “pure” verbal reports came about at a time (1984) when verbal reports were not seen as relevant to scientific research. Nielsen et al. offer alternative solutions such as practicing introspection or using a cultural-historical Activity Theory [9].

While Ericsson and Simon [8] argue against using third-level verbalizations, usability practitioners often find that these data give more useful information when determining a Web site’s usability problems and recommending solutions. Dumas and Loring give a set of rules on how to moderate usability tests. Inherent in some of the instruction is the opportunity to elicit third-level verbalizations [10]. Boren and Ramey write, “Many [published researchers] prefer explanations, coherency of verbalization, and participants’ design or revision ideas over strictly procedural information—suggesting that Level-3 data is of greater importance than Level-1 and 2 data” [1, p. 265]. Thus it is important to evaluate if third-level data are compromised and/or if there are alternatives to the traditional [8] protocol that would work for practitioners, while assuring that data are accurate.

**Alternatives to the traditional think-aloud protocol**

The difference between traditional TA [8] and the practice of usability professionals has caused some researchers to question whether another approach to thinking-aloud might be more effective in usability studies. Nielsen, Clemmensen and Yssing [9] have suggested introspection while Boren and Ramey [1] advocate speech communication. Speech-communication research suggests that for usability studies, the traditional TA protocol where the test administrator remains silent throughout a session providing short assertive commands
to “keep talking” might be more disruptive to the participant than formerly acknowledged because humans communicate within a speaker/listener relationship [1, 11]. In both the speech communication and linguistic fields the understanding is that during a conversation, it is essential for the listener to use verbalized sounds or phrases which indicate to the speaker that the listener is paying attention and is absorbed in the communication act (e.g., OK, um-hum, ohhh) [12,13,14]. Nielsen et al. [9] contend that the constraints of the Ericsson and Simon [8] protocol do not have a holistic understanding of humans and the vast ways of thinking combined with emotions and feelings that go on during any interaction.

Dumas and Redish [5] mention a technique called “active intervention” (p. 31) in which the test administrator actively probes to get at the participant’s mental model, or the participant’s thought process for how something works. They say that the appropriate technique to use depends on what the goal is (i.e., what you want to find out). Dumas and Loring instruct that how much or how little to intervene depends on the type of test (e.g., formative or summative) and the probes and questions put to the participant should “further the objectives” of the test [10, p 72]. They go on to give examples of various probes which include direct questions about specific areas of the interface being tested, as well as probes about different navigational paths or options that a participant chose (pg. 74) among other probes. With such a proliferation of different strategies and techniques at eliciting participant verbalizations during TA usability studies, the need is clear for research into effects of different protocols on user verbalizations during usability testing.

**Experimental Methods**

This paper is an extension of an earlier study [15] that involved one independent variable (thinking aloud) with four treatment conditions: the traditional protocol, the speech-communication protocol, the coaching protocol and a silent control condition. In the control condition participants worked in silence and were not asked to think out loud. The focus of this paper is on the dependent variable as the number of verbalized and non-verbalized counts of frustration and counts of positive comments.

**Participants**

Participants were recruited from the Census Bureau Usability Lab database. The database was created over a number of years and is composed of people in the metropolitan DC area willing to participate in a usability study. The database is filled with people who learned about usability studies at the Census Bureau by electronic postings (Craigslist and listservs), paper flyers, or a free weekly newspaper ad. Each participant had at least one year of prior experience in navigating different Web sites and did not have extensive prior experience using the Census Web site. Participants were reimbursed $40.00 for expenses associated with participating in the study.

**Materials**

Participants worked on eight simple find tasks of comparable difficulty. A simple find task is one in which the participant is asked to find a single piece of information. Task order was randomized to control for learning. See Appendix A for a list of the tasks. The Web site used was the main Census Web site [http://www.Census.gov](http://www.Census.gov).

**Data coding video tapes**

Two different independent coders coded the various taped sessions using pre-identified behaviors (e.g., vf = verbal frustration, nf = non-verbal frustration, vp = verbal positive, np = non-verbal positive). To keep the data coders as objective as possible, they were not told the hypotheses or goals of the study. Additionally, to obtain accurate results, each session was coded at least twice by independent loggers. The person who coded it the first time was not the person who coded it the second time. When discrepancy on a particular code occurred, a third logger, who was a senior usability specialist, reviewed the issue, thus reducing measurement error.

**Design**

This was a between-subjects design where each test administrator proctored only one condition, and each participant experienced only one condition. We randomly assigned participants to a condition. None of the test administrators knew of the true purpose of the study, or that there were different conditions, nor did they interact (with respect to the study) with the other test administrators during the duration of the testing. The test administrators were given one-on-one instruction by the principle investigator, as well as a brief condition-specific training video which was created by two of the authors. Each test administrator was given a set of test procedures including a condition-specific script to read to the participants, a list of acceptable probes to use with participants and instruction on how often to use the probes. The test administrators had backgrounds in either psychology or computer science. Three test administrators were interns in graduate level programs. The fourth test administrator (for the control condition) was a full time employee of the Census Bureau Usability Lab.
The “think-aloud” condition was the independent variable. Each test administrator interacted with their participant as follows:

1. Traditional TA [8] -- (i.e., no probing words beyond “keep talking,” includes practice session; probing by test administrator getting first and second-level verbalizations from participant).
2. Speech-Communication TA [1] -- (i.e., verbal feedback in form of “um-hum or un-hum” to keep participant talking, includes practice session; probing by test administrator getting first and second-level verbalizations from participant).
3. Coaching TA [5,10] -- (i.e., more verbal feedback and probes where TA asks direct questions about different areas of Web site, asks direct questions about areas where user is having difficulty/is pausing/or is verbalizing area is confusing or frustrating, gives help or assists when participant is struggling, includes practice session; probing by test administrator getting third-level verbalizations from participant).
4. Control condition -- (i.e., there is no thinking aloud, no probing or prompting, includes practice session).

Procedure

Eighty participants were randomly pre-assigned to condition (20 per condition). This number was determined by an \textit{a priori} power analysis detecting a moderate effect size at alpha=0.05 [15]. Participants arrived at the Census Bureau and were met by a member of the Usability Lab (not their test administrator) and directed to the room where the usability session was going to take place. The test administrator entered the room and began the session by reading aloud the condition-specific information and instructions for the participant. In each session, the participant listened to the testing procedure, practiced using it during a pre-task, and signed a consent form agreeing to be video-taped. While the participant filled out a pre-questionnaire about his/her demographics and Internet experience, the test administrator left the room and re-commenced communication with the participant via microphone and speakers from the control room. The test administrator watched the participant through a video-tape feed as well as through a one-way mirror. The participant was directed by the test administrator to work on the tasks and, if appropriate to the condition, probed to think-aloud during the session. At the conclusion of the session, participants filled out a satisfaction questionnaire and were given a stipend of $40.00.

Results

We ran the non-parametric Kruskal-Wallis one-way analysis of variance test to determine if there were differences in the number of verbalized frustrations between the three conditions. (We did not analyze whether there was a difference between the control condition because the control was a silent condition so participants were not asked to think aloud.) Rather than using the raw number of counts we used the ranks of the counts. If the conditions are not different from each other then the ranking will be random.

There were no statistically significant differences on the number of verbalized and non-verbalized counts of frustration or the number of verbal and non-verbal positive counts among the three treatment conditions. In Table 1 below, we can see that there were differences in the total counts but these differences are not statistically significant.

The test results for verbal frustration are $F_{2,477} = 1.5199$, $p = 0.2198$. The test results for non-verbal frustration are $F_{2,477} = 0.5204$, $p = 0.5946$. The test results for verbal positive are $F_{2,477} = 1.3772$, $p = 0.2533$. The test results for non-verbal positive are $F_{2,477} = 0.0702$, $p = 0.9322$. The rank sums of the Wilcoxon scores did not differ from what they are expected to be if the ranks were random.
Table 1. Total counts and mean ranks for verbal/non-verbal frustration and verbal/non-verbal positive classified by condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>Total verbal frustration count</th>
<th>Mean rank</th>
<th>Total non-verbal frustration count</th>
<th>Mean rank</th>
<th>Total verbal positive</th>
<th>Mean rank</th>
<th>Total non-verbal positive</th>
<th>Mean rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>349</td>
<td>236</td>
<td>329</td>
<td>236</td>
<td>58</td>
<td>242</td>
<td>17</td>
<td>244</td>
</tr>
<tr>
<td>Speech-communication</td>
<td>356</td>
<td>256</td>
<td>343</td>
<td>244</td>
<td>61</td>
<td>237</td>
<td>9</td>
<td>239</td>
</tr>
<tr>
<td>Coaching</td>
<td>283</td>
<td>229</td>
<td>374</td>
<td>242</td>
<td>63</td>
<td>242</td>
<td>9</td>
<td>238</td>
</tr>
</tbody>
</table>

Summary

In summary there were no statistically significant differences in the number of verbal and non-verbal counts of frustration or positive counts among the three treatment conditions.

Discussion

In a typical usability study, the goal is to identify usability problems and suggest solutions to the problems. However, as this was an experiment and not a usability study, we did not identify usability problems nor did we categorize or prioritize the types or quality of the usability problems. The goal of the experiment and in coding the tapes was to obtain accurate results by being as objective as possible. We felt that identifying and counting the number of usability problems would allow for too much subjectivity and interpretation on the part of the coder/analyst which could lead to unintended bias [16, 17]. Thus we only counted verbal and non-verbal frustrations, and verbal and non-verbal positive comments. We felt that these counts would be the most objective data elements that we could identify from the sessions. The basic assumption was that any instance of annoyance, confusion or frustration with the site could be an indicator of a usability problem with the application.

We hypothesized that the speech-communication condition would elicit more verbalized instances of frustration with the site. We hypothesized this because, as the test administrator was engaging in a communicative style of interaction, the participant would be more likely to respond to the communicative style probes and enunciate more instances of confusion or frustration. However, this did not happen. This could indicate that all three conditions would equally elicit comments that could lead a test administrator to identify usability problems with the application being tested. It could also indicate that simply counting frustrations is not enough of an indicator to lead to insight on the conditions, with respect to usability problems, like we had expected. Further research on the data could include parsing the verbalized comments into usability problems rather than just counting them. The challenge in identifying usability problems is remaining objective and not allowing unintended bias to creep into the analysis. A first step in this direction could include looking at the video tapes for usability problems only after a careful definition of the possible set of problems with the site is made.

Olmsted-Hawala et al. [15] found that participants were significantly more successful in the coaching condition than participants in all other conditions. They also found that participants in the coaching condition were more satisfied with the Web site than participants in the traditional and speech-communication conditions. As there was a significant difference among task accuracy and user satisfaction with respect to the coaching condition, there is evidence that the think-aloud protocol used in usability testing has an effect on the participants. These are additional indicators that although there were no statistically significant differences in terms of the counts, further investigation of the verbalized comments of participants by condition is warranted.

Conclusion

In a double-blind empirical study on the difference of think aloud protocols in usability testing, we analyzed the differences in the traditional [8], the speech communication [1], the coaching protocol [5,10], and silent control conditions. This study looked at the number of verbalized and non-verbalized counts of frustration as well as the number of verbalized and non-verbalized counts of positive comments by condition with the Web site. There were no differences in the number of counts by condition with respect to the traditional, speech communication, or coaching conditions. We didn’t analyze the control condition as this was the silent control
and participants were not asked to think aloud. Previously we demonstrated that with respect to accuracy, there were significant differences in the coaching condition such that participants were more successful in task accuracy than in any other condition [15]. Based on this inflated accuracy score for the coaching condition, we expected that there would be more positive counts and fewer frustration counts with the coaching condition. However, in this study we found that there were no differences by condition. Simply counting the verbalizations of users by condition did not confirm the problems with the coaching protocol that were highlighted in the accuracy score [15]. Thus, counting the number of verbalizations by protocol might not be an effective measure to use when comparing think-aloud protocols. Although initially we felt that identifying and counting the number of usability problems would allow for too much subjectivity and interpretation on the part of the coder/analyst, which could lead to unintended bias [16, 17], based on the null results found when simply counting the verbalized and non-verbal instances of frustration and positive comments, future analysis of the data might suggest that we revisit the idea of identifying usability problems by condition.

Appendix A: List of Tasks

Practice Task on Canadian Census Web site: Imagine you are going on a trip to visit your friend who lives in Nova Scotia, Canada. You don’t speak French and want to know how many people speak English and how many people speak French in that area of the country.

1. You know that there are many people in the US but would like to know the actual number. What is the US population?

2. You live in Maryland and would like to know what the population of Maryland is?

3. You just heard a news report about income in America. You are wondering what the average household income of Americans is

4. You are writing a paper on education in America and would like to get some details. Specifically, how many people in the US have graduated from high school?

5. Your sister is thinking of starting her own business. You were wondering how many women-owned businesses are there in Virginia?

6. You are considering whether to retire to Florida in 7 years. You are wondering what the population will be for Florida in the year 2015?

7. You have heard recently discussions about those with and without health insurance in America. You’d like to know what percent of the U.S. population was without health insurance in 2006?

8. You are writing a historical overview on the US economy and wanted to check what the unemployment rate for the US was in 2006?

References


**Acknowledgements**

We would like to thank the test administrators, data coders, and recruiters who assisted in this project: Alex Trofimovsky, Allison Morgan, Jen Romano, Larry Malakhoff, Jenna Beck, Temika Holland and Andy Su, from the U.S. Census Bureau, as well as Dana Douglas and Mark Becker, from UserWorks, Inc.

**About the Authors**

Erica Olmsted-Hawala has been involved in usability since 2000 and is currently a member of a usability team at the U.S. Census Bureau. She has conducted, reported and presented on numerous usability studies focusing on the usability of data dissemination Web sites.

Elizabeth Murphy recently retired as a research psychologist at the U.S. Census Bureau. She has led teams conducting human factors and usability analyses since the 1980s.

Sam Hawala is a statistician at the U.S. Census Bureau. His focus is on Small Area Data modeling, model selection, and synthetic data generation.

Kathleen Ashenfelter is a research psychologist at the U.S. Census Bureau.