Think-Aloud Protocols: A Comparison of Three Think-Aloud Protocols for use in Testing Data-Dissemination Web Sites for Usability

Erica L. Olmsted-Hawala[1]
US Census Bureau
Statistical Research Division (SRD)
4600 Silver Hill Road, Washington, DC 20233
erica.l.olmsted.hawala@census.gov

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

Sam Hawala
US Census Bureau,
Data Integration Division (DID)
4600 Silver Hill Road, Washington, DC 20233
sam.hawala@census.gov

Kathleen T. Ashenfelter
US Census Bureau, SRD
4600 Silver Hill Road, Washington, DC 20233
kathleen.t.ashenfelter@census.gov

ABSTRACT

We describe an empirical, between-subjects study on the use of think-aloud protocols in usability testing of a federal data-dissemination 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. Accuracy and efficiency measures were collected, and participants rated their subjective satisfaction with the site. Results show that accuracy is significantly higher in the coaching condition than in the other conditions. The traditional protocol and the speech-communication protocol are not statistically different from each other with regard to accuracy. Participants in the coaching condition are more satisfied with the Web site than participants in the traditional or speech-communication condition. In addition, there are no significant differences with respect to efficiency (time-on-task). This paper concludes with recommendations for usability practitioners.

Author Keywords
Think aloud, user testing, verbalization, user-centered design, usability testing, experimental design

ACM Classification Keywords
H.5.2 User Interfaces, theory and method.

General Terms
Human Factors

INTRODUCTION

The think-aloud (TA) protocol is one of the primary tools used by usability professionals when conducting usability tests. Usability practitioners use the TA protocol because they cannot directly observe what a user is thinking. With the TA protocol, areas where a user is struggling and the reasons for the difficulties are verbally articulated, at least in theory. The usability practitioner uses this information along with other metrics to identify problem areas of the Web site or application being assessed and to devise suggestions for improvement. One of the most common TA protocols that usability practitioners engage in today is concurrent TA under which the participant is encouraged to "think out loud" while working on a task.

TA protocols are used widely by usability professionals, yet all too often, researchers do not describe the actual methodology they use [1,8,10,13]. This is problematic for two reasons. First, usability practitioners and other researchers have no way of knowing what type of TA was used in a particular study, which means that replicating that study is nearly impossible. Second, the TA protocol employed by a practitioner may lead to inaccurate results.

To date, there have been few empirical research studies on the TA protocol in usability testing. This lack of research-based guidance has led to the TA protocol varying from study to study, and, possibly, to the reporting of inaccurate usability results. The present study aims to provide

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

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-929-9/10/04.

practitioners with a better understanding of the strengths and weaknesses of the different variants of the TA protocol. The work described in this paper details a double-blind, empirical study of three different types of concurrent TA protocols and one silent control condition. It is the first to compare these specific protocols in an experimental context and thus adds to the research literature on TA. The study was carried out in the US Census Bureau's Usability Laboratory, using the Census Bureau's main public, data-dissemination Web site as the object of a putative usability study (http://www.Census.gov). Participants were told they were helping the Census Bureau evaluate a public Web site, but the purpose of the study was to investigate the effects of various TA protocols on user performance and satisfaction. This paper reports the results found for performance (accuracy and efficiency) as well as subjective user satisfaction. It does not identify or discuss usability issues, since they were not the focus of this research study.

Related Work

The basis for the TA protocol most cited in introductory usability texts [5,12,17], and usability/TA related articles [1,14,15,20,21,22] is the verbal protocol developed by Ericsson and Simon [6]. Yet practitioners often ignore the strict constraints these authors impose on the TA protocol [1,13].

Ericsson and Simon [6] categorize the different verbalizations that a participant makes into three levels, where the first two levels only require information processing in the participant's short-term memory, and are thus considered legitimate data. Ericsson and Simon would consider invalid any verbalizations that draw on the participants' long term memory or any verbalizations that are the result of redirection from what participants were originally doing and speaking about. Examples of probes (given by a test administrator) that would lead to a participant giving Level-1 through Level-3 verbalizations follow:

- Levels-1 and 2—"Keep talking" or "Um-humm?" (With these probes, participants are not distracted from current focus and continue with what they were doing.)
- Level-3—"Why did you click on that purple tab?" (To answer the question, participants would likely be required to access long-term memory and might get distracted from what they would ordinarily have continued doing if the question had not been asked.)

Ericsson and Simon, who originally published their manuscript in 1984, were working at a time when verbal reports were not necessarily viewed in psychology as legitimate data. From their perspective, Level-1 and 2 verbalizations are "a rich source of data, combinable with other data, that can be of the greatest value in providing an integrated and full account of cognitive processes and structures" (p. 373). Other researchers have conducted studies to further support the work of Ericsson and Simon [6]. Rhenius and Deffner [15] showed with eye tracking that concurrent verbalizations and short term memory were synchronized, that is where a participant was looking was directly related to what the participant was verbalizing. Wright and Converse [24], conducted a study which showed that Level-3 verbalizations improved user task performance during a usability study. They warned that the use of Level-3 probing in usability tests resulted in a bias toward improved user performance (p. 1223). The irony is that while Ericsson and Simon [6] argue against using Level-3 data, usability professionals often find that Level-3 data give the most useful information when determining the usability violations in software or a Web site and when formulating possible ways to resolve the problems. Boren and Ramey [1] sum it up thus: "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" (p. 265).

All usability practitioners who use the TA protocol in their studies try, in one way or another, to get participants to talk about what they are thinking about as they work on a task. The protocol that Ericsson and Simon [6] advocate for is a verbal protocol that is strictly non-intrusive, such that the only allowable verbal prompt is "keep talking." This, they assert, is the way to keep the participants' verbalizations pure, that is, free of contamination from long-term memory.

However this technique varies greatly from what is used in practice by many usability practitioners [1,8,10,13]. Often the TA probes are more intrusive or potentially leading or directive than would be advocated by Ericsson and Simon [6]. Permissible probes are intended to get at what the participants are actually thinking about as they work on a task. Yet many probes in use today could lead participants to draw information from their long-term memory, redirect them from what they would ordinarily have done or put words into their mouths (i.e., lead them in a direction not of their own choosing).

Dumas and Redish [5] refer to 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. Dumas and Redish say that the appropriate technique to use depends on what the goal is (i.e., what the researcher wants to find out), but they offer no alternatives to active intervention. Hertzum et al. [8] looked at two different TA protocols. The first was following Ericsson and Simon's [6] TA protocol, (they termed it the "classic model"). The second was what they called a "relaxed" style protocol. The "relaxed" protocol sounds similar to the active intervention of Dumas and Redish [5], which is itself similar to the coaching protocol used in this current study. Hertzum et al. [8], found that the "relaxed" style, when compared with a control group, affected user performance in many ways, whereas the classic protocol did not. While there have been repeated calls for more research into verbal protocols as they relate to usability studies [1,4,10], there has not been an overwhelming response to such calls. Thus, it is not clear which variety of TA protocol is advisable for usability practitioners to follow.

Alternative Theories

The discrepancy between the TA protocol advocated by Ericsson and Simon [6] and the practice of usability professionals has caused some researchers to question whether another type of protocol might be more effective in usability studies. Speech communication was put forward by Boren and Ramey [1] as one such alternative. Their work was based on field observation and encouraged more empirical research to be conducted on the TA techniques based on theories of speech communication. Speech-communication research suggests that for usability studies in particular, the Ericsson and Simon [6] technique of the test administrator keeping silent throughout a session with only short assertive commands to "keep talking" might be more disruptive to the participant than formerly acknowledged [1,10].

Speech-communication theory holds that the ways human beings naturally communicate within a speaker/listener relationship include a certain amount of acknowledgment and feedback or the use of "back channels" (e.g., um-hum, oh, okay) from the listener. Back channels are indicators that the listener is being an "active listener" because dialogue is more than just an information exchange. The speech communication and linguistic fields assert that in a conversation, it is important for the listener to use verbalized phrases or sounds, indicating to the speaker that the listener is paying attention and is engaged with what is being said [16,23,25]. Back channels include facial expressions and body language, although these nonverbal cues are not essential to conversations, such as in telephone conversations where the back channels are all verbal. In fact, when back channels are not used in phone conversations, the speaker is motivated to ask, "Hello, are you still there? [25]" Boren and Ramey [1] suggest that the back channel (um-hum, oh, okay) "acknowledgement token" or response token most conducive to keeping the speaker talking is the quietly affirming "um-hum" response given at the appropriate time in the "conversation" (p. 270).

Krahmer and Ummelen [10] conducted an exploratory study with 10 participants which compared the TA protocol described by Ericsson and Simon [6] with the new TA protocol proposed by Boren and Ramey [1]. The study found that in the speech-communication mode, more tasks were completed and participants appeared less "lost" on the site. The Web site tested was a highly unusual site based on work by the Dutch writer Harry Mulisch; it is an artistic attempt at a "journey through Mulisch's mind" (p. 110) [11]. The description of the speech-communication protocol used appears to reflect some amount of coaching of the participant. (For example, when stuck, a participant was encouraged to continue and was given some direction on how to do the task.)

Variety of TA Protocols in Practice

One of the primary concerns with TA in usability studies is that the wide variety of TA protocol styles introduces different reliability issues into the results. Among published usability professionals [5,12,17], there is some amount of methodological variation in TA protocol. At this point, given the available research, it is challenging to evaluate the effectiveness of the TA protocols because of the various, often loosely described techniques in use. Some of the variation includes the following dimensions:

- Instruction: No instruction to varying amounts and types of instruction with (or without) practice sessions. An example of an instruction to a participant follows: "Tell me why you clicked on a link or where you expect the link to take you. Tell me if you are looking for something and what it is and whether you can find it." Other instructions might not be as specific.
- Intervention: Different types of intervention with a variety of probes or prompts. Example probes follow: "Keep talking," or "Is that what you expected to happen?" or "What do you think of the color of the banner?" or "What are you thinking about now?" or "What do you think that question mark icon means?"
- Prompting: Varying rates of the test administrator's administering a prompt (e.g., after 10 seconds of silence, at random, after a "prolonged period" of time, etc.)

In addition, when researchers report TA protocols, many leave out details on the specifics: for example, they don't list the types of probes they used or how often probes were given, how long they waited before they probed, and so forth [9]. Many researchers say they used the TA protocol yet fail to give details on the actual TA protocol: as the work by Boren and Ramey [1] shows, when practitioners were using what they considered the "typical" TA protocol, prompts varied widely among the observed practitioners. Various practitioners used the simple "um-hum" to the extended "what do you think x button does," to questions such as "is there anything in particular you're looking for?" (p. 264). As the work by Norgaard and Hornbaek [13] describes, when watching practitioners in the field conducting real usability studies, they witnessed many instances where usability practitioners asked hypothetical or leading questions which likely would elicit what Ericsson and Simon [6] term Level-3 data.

This paper describes an experiment that looks at three different variations of the concurrent TA protocol: the traditional protocol put forward by Ericsson and Simon [6], the speech-communication-based TA protocol described by

Boren and Ramey [1], and a coaching protocol, loosely based on what Dumas and Redish [5] discuss, what Krahmer and Ummelen [10] and Hertzum et al., [8] describe, and what Norgaard and Hornbaek, [13] witness. The experiment differs from a traditional usability study in that we were not identifying usability problems although that is what we told participants we were doing. Rather, the intent of the research was to investigate the effects of variations of the TA protocol on participant success and perceived satisfaction in usability tests. Identifying usability problems would have distracted from our main focus on the effects of the TA protocol on the participants' task performance (accuracy and efficiency) and self-rated satisfaction. These relationships need to be understood before we start looking at the number of identified usability problems. In the current study, we investigated the following research questions:

- Is there a difference among the conditions (the three different approaches to TA) with respect to user task performance (accuracy and efficiency)?
- Is there a difference among the conditions with respect to self-rated user satisfaction?

EXPERIMENTAL METHOD

This study involved one independent variable (thinking-aloud) with three treatment conditions and a control. The three TA conditions were the traditional technique, the speech-communication technique, and coaching. The control condition was silence: Participants in this condition did not think aloud. The three dependent variables were accuracy, efficiency, and satisfaction. See the data analysis section below for information on measurement.

Test Administrators

Each of the four test administrators proctored only one condition. Test administrators had backgrounds in psychology, computer science and/or human computer interaction. This was a double-blind study: none of the authors of the paper conducted the sessions, and the test administrators did not know what the true purpose of the study was. The test administrators did not interact with each other nor did they know that there were different conditions. None of the test administrators were told the name of their test condition until the conclusion of the study. Test administrators were given instructions and training by the first author of the paper. At no time did the authors give names to the conditions when talking to or training the test administrators. All test administrators had a condition-specific script to read to the participants. In addition, all test administrators had condition-specific instructions on how to run and troubleshoot their sessions. As part of their training, test administrators were given a 5-minute video of a trained test administrator working with a mock participant using the particular condition that the specific test administrator was expected to use. The condition-specific video highlighted how to conduct a testing session, including examples of probes, how frequently to enunciate the probes and how to give feedback (condition-specific) to the participant. As part of the training, each test administrator conducted a “dry run” with a Census employee who was not associated with the study or the actual testing. Two of the authors of the study observed the dry runs and gave feedback on the appropriateness of the probes used during the dry-run sessions. Each test administrator ran only his or her particular condition. We did this to maintain internal consistency across the participants in each condition. Training materials are available upon request from the first author.

Procedure

We randomly pre-assigned participants to the three treatment conditions and the control condition using the SAS function Proc Plan [18]. Each participant was given instruction by the test administrator, depending on the condition, on what they were to do during the session. Participants were not informed of the true purpose of the study until their session was finished. They were not informed about the treatment conditions. They were led to believe that they were helping the Census Bureau evaluate a Web site. Each condition had a different protocol that the test administrator read aloud to the participant while sitting next to the participant. Each condition had the test administrator run a practice session while sitting next to the participant. In each condition the test administrator then left the room where the participant sat and went to the laboratory’s control room from which they recommenced communication with the participant using a microphone. The test administrator watched the participant via video tape feed as well as through a one-way mirror.

The four test administrators interacted with their participants as follows:

- Traditional: Think-aloud following the Ericsson and Simon method [6] -- (i.e., no probing words beyond “keep talking;” includes practice session before session begins; probing by test administrator after 15 seconds of silence).
- Speech Communication: Think-aloud following the speech-communication theories put forward initially by Boren and Ramey [1] -- (i.e., verbal feedback in form of “um-hum or un-hum” to keep participant talking; includes practice session before testing begins; probing by test administrator in form of feedback tokens or questioning tone picking up on last word uttered by participant after 15 seconds of silence, e.g., Participant says “that was odd…” Test administrator says after pause “Odd?”)
- Coaching: Think aloud with active intervention, or coaching of participant described by [5,8,10,13] -- (i.e., more verbal feedback and probes where test administrator asks direct questions about different

areas of Web site, such as areas where user is having difficulty/is pausing/or is describing area as confusing or frustrating; gives help or assists when participant is struggling; includes practice session before testing begins).

- Silent Control: No think aloud in the control condition -- (i.e., there is no thinking aloud; no probing or prompting by the test administrator; includes a practice session before the session begins).

Each session was video-taped. In each session, the participant signed a consent form and completed a prequestionnaire about their Internet use, computer experience and basic demographics. At the end of the session, each participant filled out a modified version of the Questionnaire for User Interaction Satisfaction (QUIS) [3]. The QUIS we used was a shortened version of the full length QUIS, and the text was tailored for the study context. The shortened and modified version that we used was sufficient for our present study, see Appendix A. This measure is the basis of the self-rated satisfaction score.

Test Participants

Eighty participants were randomly drawn from the Usability Lab's database. This database was created over a number of years and is composed of people in the metropolitan DC area who are willing to participate in a usability study and who learned of usability studies at the Census Bureau by one of the following mediums: electronic postings (Craigslist and listservs), paper flyers, and a free weekly newspaper distributed to riders of the Washington DC Metro/subway system. Each participant reported at least one year of prior experience in navigating different Web sites and did not report extensive prior experience using the Census Web site. Each participant was given a stipend of $40.00 for expenses associated with participating in the study. The experimental design, including procedures, was reviewed and approved under the Census Bureau's generic clearance for pretesting by the US Office of Management and Budget (OMB number 0607-0725).

Task Scenarios

Participants completed eight simple "find" tasks of comparable difficulty. A simple find task is defined as a task where the user is asked to find a single number or piece of information on the data-rich Census Bureau Web site. Task order was randomized to control for learning. The task scenarios come primarily from the most common tasks that users perform at the Census Bureau Web site and thus were considered to be representative of what typical users come to the site to do. When deciding on the tasks to use in the study, the authors reviewed the user queries that actual site users had typed into the search tools (both Google and the FAQs). The authors reviewed the weekly American FactFinder (a data-dissemination site off of Census.gov's main Web site) email queries, as well as the Web usage statistics to determine areas of the site that caused real users to have questions and/or that such users found confusing. An example of one of the find tasks is "You know that there are many people in the US but would like to know the actual number. What is the US population?" The complete set of tasks is available by request from the first author.

Data Coding Videotapes

We had two different independent coders code the taped sessions for task outcome (e.g., ss = task success, ff = task failure). The coding software also put time stamps on each code so we had efficiency data. Due to the double blind nature of the study, the data coders did not know of the hypotheses or goals of the study. The goal of the experiment and of coding the tapes in particular, was to be as objective as possible. Thus, each session was coded at least twice by different independent coders. When a discrepancy occurred on a particular code between the first two coders, a third coder, who was a highly experienced usability practitioner, reviewed the issue. This process was followed as a way to reduce measurement error. To maintain neutrality, the data coders were external contractors that had been hired specifically to code the video tapes.

A Priori Power Evaluation

We performed a balanced one-way analysis of variance (ANOVA) with one user group, four conditions, and three dependent variables. For our *a priori* evaluation, we adhered to the minimum recommended sample size for an ANOVA of 20 participants per cell [7,11,19] in order to detect a moderate effect size for the significance criterion of *alpha* =0.05. Another rule of sample size is that the sample in each cell should be greater than the number of dependent variables [19], which is also true for this experiment. Using an adequate sample sizes increases the robustness of the analysis. Therefore, we recruited 80 participants (20 per condition).

Data Analysis

Data analysis consisted of summarizing the accuracy and efficiency scores by condition. Accuracy was measured in terms of success or failure on the tasks. For each task there was only one acceptable answer. Efficiency was measured as the time it took in seconds to complete each task. We started the time measure at the moment after the participant finished reading out the task and went until the time when the participant either found their answer or said they were ready to move on to the next task. Data analysis also consisted of summarizing the subjective satisfaction scores by condition.

In this putative usability study we did not identify a list of usability problems by condition. Although usability problem identification is typically the goal of usability studies, in this study we were interested in identifying whether the type of TA protocol used had an effect on user performance and satisfaction. Thus we report the accuracy, efficiency and subjective satisfaction by condition.

RESULTS

Accuracy

The one-way analysis of variance (ANOVA) with alpha=0.05 shows that condition has a significant effect on Accuracy ($F_{3,636}$ =13.48, $p < 0.0001$). To understand the result of the study, we determined which condition had the biggest effect on Accuracy, and whether all the conditions were significantly different from the control condition. The first planned comparison compared the control with all the other conditions. The next comparison was the coaching condition against the first two conditions, traditional and speech communication, ignoring the control in this contrast. A third and final contrast, orthogonal to the first two contrasts, pitted the traditional condition against the speech-communication condition.

This analysis shows that the contrasts of the control with the other three conditions ($F_{1,636}$ =7.95, p = 0.010) and the contrast of coaching against the first two conditions traditional and speech communication ($F_{1,636}$ =28.52, p <0.0001) are significant. The third contrast of the traditional condition against the speech-communication condition is not significant ($F_{1,636}$ =3.96, p=0.094).

Accuracy results appear in Table 1. When comparing coaching to all other conditions, there is a significant difference for accuracy: participants were more successful on all tasks if they were in the coaching condition.

Condition	Percent Correct	St.Dev. of the mean
Traditional condition	40%	3.9
Speech-communication condition	30%	3.6
Coaching condition	60%	3.9
Silent control	31%	3.7

Table 1. Summary accuracy (percent correct) results for TA conditions: Participants in the coaching condition were more successful than participants in any other condition.

Efficiency

The analysis of Efficiency shows that condition has no significant effect on task-completion time ($F_{3,636}$ =1.01, p = 0.770).

As reported in seconds in Table 2, the efficiency results show that there was no significant efficiency difference among the four conditions on this variable. It is interesting to note that the requirement to think aloud during the three treatment conditions did not significantly differ from the fourth silent condition with respect to task-completion time.

Condition	Mean	St.Dev. of the mean
Traditional condition	243	12
Speech-communication condition	270	13
Coaching condition	268	13
Silent control	252	12

Table 2. Summary efficiency results for TA conditions, in seconds: No significant differences between conditions.

Satisfaction

The satisfaction score was calculated by summing sixteen scores from the modified version of the QUIS. Each score was on a Likert scale from 1 to 7; so summed together the total for a participant was from 16 to 112. The higher the score, the more satisfied the user reported being with the site.

The overall results for satisfaction indicated ($F_{3,76}$ = 2.39, p = 0.151) that the satisfaction score means do not differ between conditions. However, when we contrasted the coaching condition with the traditional condition and the speech-communication condition (individually or together), we found ($F_{1,56}$ = 7.14, p = 0.018) a significant difference, suggesting that when participants receive the coaching condition they are significantly more satisfied with the Web site than when they receive the traditional, or the speech-communication condition.

The satisfaction means by condition are shown in Table 3 below.

Condition	Mean[1]	St.Dev. of the mean
Traditional condition	73	4
Speech-communication condition	73	4
Coaching condition	85	3
Silent control	78	4

Table 3. Summary satisfaction results for TA conditions: on a totaled scale of 16 to 112, where 16 is unsatisfied and 112 is highly satisfied. Participants in the coaching condition were more satisfied with the Web site than participants were in the traditional or speech-communication condition.

Summary of Results

To summarize, we found that participant in the coaching condition were more successful than participants in any other condition. We also found that participants in the coaching condition were more satisfied with the Web site

[1] There was only ONE satisfaction score per user (totaling 80 scores, 20 per condition), whereas there were 640 accuracy measurements (one per user per task).

than participants in the traditional or the speech-communication condition. Finally, there was no difference among the conditions in terms of efficiency. That is, whether or not a person was asked to think aloud during a usability session did not have an effect on the amount of time it took them to complete their tasks.

Limitations

Limitations in the study include the following:

- There were only three TA conditions. These are three common types used by practitioners; however, there are others that we did not study.
- Due to time constraints, we used only eight simple find tasks on a single data-dissemination Web site. More complex tasks and other user interfaces (e.g., commercial Web sites) could be used with the three TA protocols that we assessed.
- There was only one administrator per condition, to maintain internal consistency across participants; however, the design may have allowed the personalities of the test administrators to influence the outcomes in unknown ways. We think the extent of such influence is limited, however, because of the rigor and extensive amount of condition-specific training that the test administrators underwent. If we had trained all of the test administrators in all of the conditions we would have lost the double-blind aspect of the study.
- We did not analyze differences by participant characteristics, such as age, sex, or education level.
- The characteristics of the usability database from which we recruited the participants is also a limitation All participants were from the Metropolitan DC area, and all had indicated at some point that they would be interested in participating in a usability study. Thus, ours was not a random sample of the entire US population, but we feel it is a good mix of people who may have reason to use the Census.gov Web site. We did assign the participants at random to condition. According to guidance we received from OMB, random assignment is sufficient to meet the assumptions of statistical hypothesis testing.
- Many studies on TA protocols give their results in terms of the number of usability problems identified. While it is true that identifying usability problems is the main purpose of usability studies, the focus of this study was not on usability problems. Therefore, we cannot draw meaningful conclusions about them. Since this was experimental research and not a usability study, we were focused on understanding the effects of the different TA protocols on the participants' performance and satisfaction.

Discussion

In looking at other studies that have been conducted on TA protocols, our findings support the previous work by Hertzum et al. [8], Krammer and Umullen [10], Wright and Converse [24] and Rhenius and Deffner [15] wherein the TA protocol that uses some amount of coaching leads to higher accuracy rates compared to a control condition and/or higher accuracy rates compared to a traditional condition modeled after the protocol put forward by Ericsson and Simon [6]. We also found that the speech-communication protocol (like the traditional protocol) did not have an effect on accuracy; however, no other studies have used such a speech-communication protocol as was used in our study.[2]

Our findings contradict some of the current literature with respect to our efficiency measure. Although none of our TA treatment conditions were statistically significantly different from the silent control in terms of time-on-task (e.g., efficiency), both Hertzum et al. [8] and Rhenius and Deffner [15] found that tasks took longer to complete in their TA condition than in their control. In contrast, Wright and Converse [24] found that participants in their verbal protocol were more efficient than were those in the control. Any expectation of participants' being more or less efficient depending on condition was not borne out, as we found no significant differences in efficiency between the control and any of the conditions. This finding supports the work of Bowers and Snyder [2] who likewise found no efficiency differences between their two conditions. Any differences or lack of differences in efficiency are of interest to practitioners who would like to know whether or not it is good practice to collect time-on-task measures during usability studies. Our study suggests it is legitimate to do so. Since it is one of the few studies to find no differences in efficiency between TA conditions, the issue of TA effects on efficiency might benefit from additional research.

Implications for Usability Practitioners

Earlier studies have shown that usability practitioners use a variety of different TA protocols [1,13]. However when tested in this experimental study, the different TA protocols commonly in use have effects on user performance. Specifically the coaching protocol, in which the test administrator asked more probing or leading questions or gave assistance to a participant struggling with a task, led to

[2] It is true that Krammer & Umullen [10] based one of their protocols on what Boren & Ramey [1] referred to as a speech-communication protocol; however, since the test administrator in the Krammer & Umullen [10] study offered assistance and encouragement to the test subject during the session, we think their speech-communication protocol is more akin to the coaching condition in our study.

a significantly higher success (accuracy) rate and significantly higher satisfaction ratings. This is detrimental in a typical usability study because the coaching injects bias into the results: the results are skewed toward better performance than the participant would have achieved without help. This study also highlights that practitioners have a choice between using the traditional TA mode put forth by Ericsson and Simon [6] or the newer mode suggested by Boren and Ramey [1], as these two conditions do not show any statistically significant differences in accuracy or satisfaction ratings. Finally the study shows that practitioners can collect time-on-task or efficiency measures during usability studies as there were no statistically significant differences in time between the control condition and any of the other conditions (which had participants using a verbal TA protocol).

In addition, it is recommended that rather than writing a vague statement such as "we had participants think aloud," practitioners need to document their type of TA protocol more completely, including the kind and frequency of probing. More complete documentation of TA methods will support valid comparisons across studies. Researchers interested in replicating studies need to be able to classify what kind of probing and what type of TA protocol was employed in a particular study. The methods used and procedures followed must be described in more detail. As the practice of describing more fully the type of TA used in a particular study becomes common practice, it can be expected that usability practitioners will benefit and begin to follow a more standard set of procedures.

CONCLUSION

Usability practitioners currently use variations of TA as the primary way to identify usability problems. The practice of TA varies greatly, and there are few research articles on which protocol is most effective. This double-blind research study used 80 participants in four different conditions: traditional TA, speech-communication TA, coaching TA, and a silent control condition to compare user performance with respect to accuracy, efficiency and subjective satisfaction in a putative usability study of a federal data-dissemination Web site. The results of this study indicate that of the three TA techniques we compared, the one that involves coaching improves the users' performance and increases the users' satisfaction over the others. So, if usability researchers want to portray user performance as it might occur "in the field" (without a coach), they should not use the coaching method. By using one of the other techniques, traditional or speech communication, practitioners will obtain user performance outcomes that more closely resemble how users would do without help. Our study shows that practitioners who want measures that reflect unaided user behavior can choose between two methods—traditional or speech communication—that do not differ significantly from each other in terms of their effects on user performance.

REFERENCES

1. Boren, T., and Ramey, J. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication 43*, 3 (2000), 261-278.
2. Bowers, V., and Snyder, H. Concurrent versus retrospective verbal protocol for comparing window usability. In *Proc Human Factors Society 34th Annual Meeting*, (1990), 1270-1274.
3. Chin, J.P., Diehl, V., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proc CHI 88,* ACM Press (1988), 213-218.
4. Deffner, G. Verbal protocols as a research tool in human factors. In *Proc Human Factors Society 34th Annual Meeting*, (1990), 1263-1264.
5. Dumas, J and Redish, J. *A Practical Guide to Usability Testing*. Intellect Press, Portland, OR, USA, 1999.
6. Ericsson, K.A. and Simon, H.A. *Protocol Analysis: Verbal Reports As Data*. (Revised ed.) MIT Press, Cambridge, MA, USA, 1996.
7. Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. *Multivariate data analysis*. (5th ed.). Prentice Hall, Englewood Cliffs, NJ, USA, 1998.
8. Hertzum, M, Hansen, K., and Anderson, H. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology 28,* 2 (2009), 165-181.
9. Hughes, J., and Parkes, S. Trends in the use of verbal protocol analysis in software engineering research. *Behaviour & Information Technology 22,* 2 (2003), 127-140.
10. Krahmer, E., and Ummelen, N. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication 47*, 2 (2004), 105-117.
11. Khasawneh, S. A. *Construct validation of an Arabic version of the learning transfer system of the learning transfer system for use in Jordan.* (2004), Dissertation: Louisiana State University.
12. Nielsen, J. *Usability Engineering*. Morgan Kaufmann, New York, NY, USA, 1993.
13. Norgaard, M., and Hornbaek, K., What do usability evaluators do in Practice? An explorative study of think-aloud testing. In *DIS 2006,* ACM Press (2006), *209-219.*
14. Novotny, E., and Cahoy, E. If we teach, do they learn?: The impact of instruction on online catalog search strategies. *Portal: Libraries and the Academy*, *6*, 2 (2006), 155-167.
15. Rhenius, D., and Deffner, G. Evaluation of concurrent thinking aloud using eye-tracking data. In *Proc Human Factors Society 34th Annual Meeting*, (1990), 1265-1269.

16. Romaine, S. *Communicating Gender*. Erlbaum, Mahwah, NJ, USA, 1999.

17. Rubin, J. *Handbook of usability testing: How to plan, design, and conduct effective tests*. Wiley, New York: NY, USA, 1994.

18. SAS Institute, Inc. *Statistical Analysis Software (SAS 8.02)*, 2001, Software package.

19. Tabachnick, B.G and Fidell, L.S. *Using Multivariate Statistics* (3rd Ed.). HarperCollins, New York, NY, USA, 1996.

20. Van Den Haak, M., De Jong, M., and Schellens, P. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology 22*, 5 (2003), 339-351.

21. Van Den Haak, M., De Jong, M., and Schellens, P. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers 16*, 6 (2004), 1153-1170.

22. Van Den Haak, M., De Jong, M., and Schellens, P. Evaluation of an informational Web site: Three variants of the think-aloud method compared. *Technical Communication 54*, 1 (2007), 58-71.

23. Ward, N and Tsukahara, W. Prosodic features which cue back-channel feedback in English and Japanese. *Journal of Pragmatics 32*, (2000), 1177-1207.

24. Wright, R., and Converse, S. Method bias and concurrent verbal protocol in software usability testing. *Proc Human Factors Society 36th Annual Meeting*, (1992), 1220-1224.

25. Yngve, V. On getting a word in edgewise. *Papers from the sixth regional meeting [of the] Chicago Linguistic Society*, (1970), 567-577.

Appendix A

The shortened and modified version of the Questionnaire for User Interaction Satisfaction (QUIS) [3] used in the current think-aloud study.

Please circle the numbers that most appropriately reflect your impressions about using the Census Web site. Please do this quietly. Do not think aloud.

1. **Tasks can be performed in a straightforward manner:**
Never Always
1 2 3 4 5 6 7

2. **Organization of information on the site:**
Confusing Very clear
1 2 3 4 5 6 7

3. **Use of terminology throughout the site:**
Inconsistent Consistent
1 2 3 4 5 6 7

4. **During the session, the test administer appeared to be**
Unfriendly Friendly
1 2 3 4 5 6 7

5. **Information displayed on the screens:**
Inadequate Adequate
1 2 3 4 5 6 7

6. **Census Bureau-specific terminology:**
Too frequent Appropriate
1 2 3 4 5 6 7

7. **Characters on the computer screen:**
Hard to read Easy to read
1 2 3 4 5 6 7

8. **Learning the site:**

Difficult Easy

1 2 3 4 5 6 7

9. **Experienced and inexperienced user’s needs are taken into consideration:**

Never Always

1 2 3 4 5 6 7

10. **Finding what you were looking for:**

Difficult Easy

1 2 3 4 5 6 7

11. **During the session, the test administer acted in the following way**

Unhelpful Helpful

1 2 3 4 5 6 7

12. **Forward navigation:**

Impossible Easy

1 2 3 4 5 6 7

13. **Backwards navigation:**

Impossible Easy

1 2 3 4 5 6 7

14. **Overall reactions to the site:**

Terrible Wonderful

1 2 3 4 5 6 7

Frustrating Satisfying

1 2 3 4 5 6 7

Difficult Easy

1 2 3 4 5 6 7

15. **Please add any additional comments:**