11-15-2012

A Comparison of two concurrent think-aloud protocols: Categories and relevancy of utterances

Katie Greiner

A Comparison of Two Concurrent Think-Aloud Protocols: Categories and Relevancy of Utterances

Katie Greiner

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Human-Computer Interaction

Rochester Institute of Technology

B. Thomas Golisano College

of

Computing and Information Sciences

Date approved: November 15, 2012

Committee members: Dr. Evelyn Rozanski, Dr. Michael Yacci, and Dr. Cecilia Ovesdotter Alm

Table of Contents

List of Figures

List of Tables

Abstract

This paper discusses the results of an exploratory study that compared type and quality of participant verbalizations experienced from two concurrent think-aloud methods. The speech communication and traditional think-aloud methods were compared in terms of the number of participant utterances spoken and relevancy of those utterances in terms of further usability analysis. Though the speech communication method produced fewer utterances, it produced more relevant utterances than the traditional method. Participants preferred hearing the moderator's acknowledgment tokens in the speech communication condition to the moderator's silence in the traditional think-aloud method. There was a significant difference with how natural participants felt while experiencing the various protocols. These findings suggest that the moderation style has a potential impact on the type of verbalizations produced during usability sessions and on how participants feel about thinking aloud.

Introduction

It is common practice for usability practitioners to use the think-aloud protocol during studies to gain access to what participants are thinking while interacting with a particular system (Ramey, Boren, Cuddihy, Dumas, Guan, van den Haak, & de Jong, 2006). Participant verbalizations are important in think-aloud studies as they offer valuable feedback concerning the product being tested. Evaluators quickly gain information about participants' concerns, frustrations, or surprises that arise during the testing process; these verbalizations assist in identifying usability issues and areas of the product or interface to improve (Rubin & Chisnell, 2008; van den Haak, & de Jong, 2003). Using the think-aloud protocol helps to better understand users' intentions that sometimes cannot be understood through observation alone (Cook, 2010).

One of the most commonly used think-aloud protocols in the usability field is concurrent think-aloud (Bowers & Snyder, 1990; Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010a; Ramey et al., 2006). Following this method, participants are encouraged to provide a continuous commentary of their thoughts while working on a task (Rubin & Chisnell, 2008). Despite its common use in the field, the concurrent think-aloud procedures vary widely among usability professionals (Boren & Ramey, 2000). More specifically, the types and frequency of test administrators' probes and interventions differ among researchers and professionals who use the concurrent think-aloud protocol (Boren & Ramey, 2000; Ramey et al., 2006). This variety in protocol may cause validity and reliability issues in verbal data analysis, making it difficult to compare research results or replicate studies (Boren & Ramey, 2000).

To date, there have been few empirical studies on these variations of concurrent think-aloud protocols (Olmsted et al., 2010a). Our study intends to provide further understanding of two variants of the concurrent think-aloud protocol: the traditional think-aloud protocol by Ericsson and Simon (1993) and the speech communication protocol proposed by Boren and Ramey (2000). The categories of participants' verbalizations were examined to determine whether one of these concurrent think-aloud methods collected more useful utterances for usability evaluation than the other. Participants' perceptions about the two methods were additionally explored.

Literature Review

Traditional Think-Aloud

Often cited in usability textbooks and research publications is the think-aloud framework proposed by Ericsson and Simon (1993). Based on their research in cognitive psychology, Ericsson and Simon (1993) revealed three levels of verbalizations that occur during think-aloud sessions. The first two levels of verbalizations are considered reliable verbal data as the information is provided by the participant's short-term memory (STM). If the researcher prompts or questions the participant, the participant's subsequent verbalizations are considered Level 3 data because the flow of content in STM has changed during the task (Boren & Ramey, 2000). Ericsson and Simon (1993) recommend not relying on Level 3 data since the participant's verbalizations have been affected by interventions, and the utterances are therefore considered unreliable data. Examples of evaluator probes that would be categorized as Level 1, 2, or 3 verbalizations (Olmsted et al., 2010a) are as follows:

- **Level 1 and Level 2**

  Probes such as *Keep talking* or *Uh-huh?* do not divert attention away from participants' focus. Participants continue working on the task without distraction.

- **Level 3**

  Questions such as *Why did you click on the Home tab?* need additional cognitive processing in order to be answered. Participants would need to access their long-term memory to respond.

Verbalizations that follow Level 3 interventions have a higher risk of being unreliable (Ericsson & Simon, 1993; Nisbett & Wilson, 1977).

According to Ericsson and Simon's (1993) model, the experimenter should not interfere during the think-aloud process. Rather, the communication between experimenter and participant should be single-directional with the participant continuously verbalizing his thoughts and the experimenter only listening (Krahmer & Ummelen, 2004). If a participant falls silent for a predetermined amount of time (i.e., 15-60 seconds), the evaluator provides reminders to talk aloud by saying, *Keep talking*.

Think-Aloud in Practice

The unnatural environment of the think-aloud procedure may cause participants to feel uncomfortable and prevent them from speaking effortlessly about their thoughts (Rubin & Chisnell, 2008). In an attempt to ease the anxiety level and remove silence, test administrators, who often cite Ericsson and Simon's method, may probe the participants with abstract and leading questions to gain more relevant utterances for usability evaluation (Nørgaard & Hornbæk, 2006). Examples of these variations of the traditional think-aloud procedure include the coaching, relaxed, and active intervention protocols (Olmsted-Hawala et al., 2010a). Despite Ericsson and Simon's (1993) suggestion to exclude Level 3 data because of its threat to reliability, usability professionals continue to use these probing protocols to collect information they believe might not be achievable with the traditional protocol (Boren & Ramey, 2000).

Ericsson and Simon (1993) believe that evaluators' interventions and questions affect participants' future verbalizations and task performance while impacting the validity of data. Carter (2007) agrees and suggests some hypothetical user responses that might occur as a result of Level 3 interventions such as *What are you thinking?* or *What are you experiencing?* He states that the varying interventions and probes interfere with participants' cognitive processes and shifts their attention. Exploratory studies such as Hertzum, Hansen, and Andersen (2009), Olmsted et al. (2010a), and Krahmer and Ummelen (2004), discovered that these 'probing' think-aloud protocols compared to the traditional think-aloud protocol affect participants' behavior and mental workload, accuracy and user satisfaction, and task success and lostness, respectively. The studies suggest that test administrators' interventions cause validity issues and affect participants' performances. The benefits and concerns of probing questions or comments have received little research attention, thus the debate continues as to how often and in what manner to intervene in usability studies (Anderson, 2004; Tamler, 1998).

An Alternate Theory: Speech Communication Protocol

The variety of concurrent think-aloud protocols and lack of adherence to Ericsson and Simon's framework has caused professionals to question if another think-aloud protocol is more effective in usability research (Olmsted et al., 2010b). Boren and Ramey (2000) suggest that a think-aloud protocol based on speech communication theory may be more effective in usability research and provide a better framework for collecting reliable verbalizations.

Boren and Ramey (2000) attempt to reconcile theory and practice by creating a more natural interaction environment between facilitator and participant, based on how humans normally communicate. According to speech communication theory, each time words are spoken intentionally for another's benefit (Boren & Ramey, 2000), the active roles of a speaker-listener relationship exist. The speaker's role (participant) is to predominantly talk and send information while the listener's role (test administrator) is to respond as much or as little as necessary (Drummond & Hopper, 1993). This two-step information exchange establishes an interaction between speaker and listener (Clark & Schaefer, 1989; Goodwin, 1986). Speech communication theory states that listeners use various back channels in the form of facial expressions, body language, and verbal cues to notify the speaker that they are actively listening and acknowledging the conversation (Olmsted-Hawala et al., 2010a).

As previously mentioned, usability professionals use neutral probes such as *What are you thinking?* or *How close was that to what you expected?* to gather additional information from participants about their task performances and preferences about the interface (e.g., Chisnell & Rubin, 2008; Nielsen, 1993; Stone, Jarrett, Woodroffe, & Minocha, 2005). These types of questioning probes are not to be used with Ericsson and Simon's model because they disrupt the task flow (Boren & Ramey, 2000). Boren and Ramey (2000) recommend the use of acknowledgement tokens such as *ok*, *mm hmm*, or *yeah* as they can provide the expected response from an active listener and still remain nondirective.

The acknowledgement token of *mm hmm* is suggested to be the best utterance to notify the participant to continue speaking during usability studies (Drummond & Hopper, 1993). Boren and

Ramey (2000) agree, that the most relevant acknowledgement tokens are *mm hmm* or *uh-huh* followed by interrogative voice inflections.

Since the acknowledgment tokens carry almost no content, participants use little cognitive processing to receive and comprehend the tokens. In contrast to Ericsson and Simon's model, tokens should be spoken periodically, adapting to the demands of communication, not just when participants are silent (Boren & Ramey, 2000). The tokens are natural continuers and do not infringe upon the flow of communication. If the participant does fall silent, under the speech communication protocol, Boren and Ramey (2000) suggest that a practitioner use the acknowledgement token of *Mm hmm?* despite there being nothing to be acknowledged. If the participant continues to remain silent, then a neutral, content-free probe of *And now…?* may be a more obvious notification to maintain speakership (Boren & Ramey, 2000).

Studies Comparing Traditional and Speech Communication Think-Aloud Protocols

Despite Boren and Ramey's (2000) call for more research pertaining to the speech communication protocol, few researchers have contributed, as there have been only a couple studies that compared the traditional think-aloud protocol with the speech communication protocol. Krahmer and Ummelen (2004) compared success rate and a quantitative measure for lostness using various *find* tasks for an unconventional website, and they discovered that the participants in the Boren and Ramey condition were significantly more successful in completing tasks and less lost than the participants in the traditional think-aloud group. There were no significant differences in the numbers of detected usability problems or number of words uttered by participants between the two conditions. It is worth mentioning that the speech communication protocol seemed relatively similar to an active intervention protocol as approximately 22% of the experimenter's interventions consisted of clarifications and suggestions. These interventions affected the participants' success rate, implying that test administration influences user performance.

Olmsted-Hawala et al. (2010a) evaluated three variants of the concurrent think-aloud protocol (i.e., speech communication, traditional, and coaching) in terms of efficiency, accuracy, and user

satisfaction for *find* tasks on a data dissemination web page. There was not a significant difference between the traditional and speech communication protocols with regard to task success or failure. Olmsted-Hawala et al. (2010a) proposed that usability researchers use the traditional or speech communication method because the coaching protocol injected bias on their results in terms of accuracy and user satisfaction due to the active interventions.

In another empirical study on the same website, Olmsted-Hawala et al. (2010b) discovered that there were no significant differences between the traditional and speech communication think-aloud protocols in terms of counts of verbal and non-verbal frustrations and positive comments which they coded using pre-identified behaviors. Olmsted-Hawala et al. (2010b) suggested that further investigation is needed to understand the verbalized comments from participants and their role pertaining to identifying usability problems since simply counting utterances does not effectively measure differences in think-aloud protocols.

Problem Statement

The previous comparative studies did not investigate verbalizations produced by participants in the think-aloud conditions. The following research focuses on categories and relevancy of participants' verbalizations gathered from the traditional think-aloud and speech communication protocols as the nature of utterances has received little attention (Zhao & McDonald, 2010). After comparing participants' verbalizations produced in traditional and relaxed think-aloud conditions and discovering that the protocols produced similar amounts of relevant utterances (i.e., utterances useful for usability evaluation, such as expressing user frustration), Zhao and McDonald (2010) conclude that it would be beneficial to explore how usability professionals might increase the amount of useful and relevant utterances without the need for the evaluator to intervene, which may be achievable with the speech communication framework.

The goal of our study is to contribute to the literature on the speech communication protocol and compare it to Ericsson and Simon's traditional framework in terms of usefulness of participant utterances in usability research. Participants' preference in moderation style was gathered to analyze how these

protocols influence users and their comfort level. In order to identify the differences between these concurrent think-aloud protocols the following research questions were investigated:

1. Is there a difference in the conditions with respect to the number of participant utterances per predefined verbalization category? See Appendix A for the list of utterance categories established by Zhao and McDonald (2010).
2. Is there a difference in the conditions with respect to the relevancy of utterances, i.e., utterances useful for usability problem discovery and subsequent analysis (van den Haak, de Jong, & Schellens, 2006)? A relevant utterance will contain information that indicates participant difficulty or causes for difficulty (Zhao & McDonald, 2010).
3. Is there a difference in the conditions with respect to participants' perceived preference in test administration style? Is there a difference in the conditions with how participants felt thinking aloud?

Methodology

Location and Setup

The comparative study took place in the Usability Lab 70-2293 in Rochester Institute of Technology's Golisano College. Participants used a *Windows 7* PC and had a high-speed connection to *Internet Explorer 8.0. Morae Recorder 3.2* software was used to record the computer screen, and a *Logitech* microphone recorded participants' verbalizations.

Participant Recruitment

An email was sent to the RIT student community asking for volunteers to take part in a usability study with an incentive of receiving a $10 Java Wally's café card (Appendix B). A link to the recruitment survey was provided in the email (Appendix C). The email described the background of the study and linked to the recruitment survey. The prospective participants' answers were screened to fit the user profile (Table 1).

Table 1. Prospective participant characteristics

Participant Type	16 regular, 1 pilot
Usability Participation	Have never participated in a usability study
Internet Use (Excluding Email)	12 or more hours a week
Different Websites Visited	8 or more websites a week
Physical Ability	No limitations in dexterity, sight, speech, or hearing
Age	20-30 years old
Gender	Male and female

Because of the nature of the study, it was important that participants did not have prior experience with usability studies because the think-aloud method was to be used during the sessions. If some were already familiar with the think-aloud protocol, it could have confounded the results. Participants also needed to have general experience with the Internet. A variety of Internet experience could have affected the number of utterances spoken during the sessions. The students who were interested and qualified based on the user profile were contacted and invited to participate in the study.

Study Design

This study compared two concurrent think-aloud methods: the traditional think-aloud and the speech communication protocol. A repeated measures design was used in order to reduce the transfer of learning effects. Since this study investigated participant utterances, it was critical that participants experienced both protocol conditions. Verbosity levels of participants may have varied thus conditions were counterbalanced to take these differences into account. Tasks were partially counterbalanced to reduce order bias (Sauro, 2011). The study sessions were conducted one week apart to reduce likelihood of protocol order and practice effects with the variants of think-aloud methods. Appendices D and E outline the think-aloud conditions and order of task scenarios for each testing session. Participants were randomly assigned into one of the four testing groups.

The test websites included two different information sites, chosen specifically to match the user profile. 13WHAM.com, Rochester's local news site, and the Dallas Area Rapid Transit website, DART.org, were used for the study. Figures 1 and 2 are screenshots of the websites' homepages. These information websites were chosen because they are representative of the averagely designed websites that are on the Internet today.

Figure 1. Homepage of 13WHAM's website, www.13wham.com

Figure 2. Homepage of DART's website, www.dart.org

The tasks were developed to be a realistic portrayal of a few common tasks that actual website visitors would potentially perform. The task scenarios were typical *find* tasks varying in level of difficulty (Table 2). This type of task instructed participants to find a single piece of information on the test websites. There was only one acceptable answer per task. These *find* tasks helped easily determine if the participant finished the task successfully or not. There was no time limit for completing the tasks, as the author did not want to interrupt participants' task flow by prematurely ending a session.

Table 2. Task scenarios for test websites

Website 1	**http://www.13wham.com/**
Overview	For the following tasks, you will use 13WHAM's website. 13WHAM is Rochester's news, weather, sports, and events team.
Find A	You and several friends want to go canoeing on Lake Ontario but are concerned about how cold the water might be. What is the temperature of the lake today?
Find B	You recently purchased a MEGA millions lottery ticket and need to check to see if you won! What are the winning numbers for New York?
Find C	You and a friend recently discussed how gas prices have risen in the past year. What was the average price of gas per gallon a year ago in Rochester?

Website 2	**http://www.dart.org/**
Overview	For the following tasks, you will be using DART's website. Dallas Area Rapid Transit (DART) gets people around Dallas and twelve surrounding cities with modern public transit services.
Find D	You are riding a DART bus heading to Richardson, Texas. When you reach Richardson, you need to meet a friend at a local restaurant. What is the phone number you need to call in order to get picked up by the DART shuttle when you arrive?
Find E	As a university student without transportation in the Dallas area, you're interested in

	DART's services. Approximately how much money would you have to pay per semester to receive unlimited rides on various services as a student?
Find F	You're currently in Irving, Texas, and need to purchase DART tickets for your transportation needs. Where are the ticket store locations in Irving?

Test Administrator

The author administered and facilitated the usability tests. To make the participant feel more comfortable and less self-conscious about the study, the test administrator sat in the usability room with the participant. She sat a few feet behind the participants and remained in the participants' peripheral vision, so that they were aware of the test administrator's presence (Rubin & Chisnell, 2008). The facilitator was prepared and practiced how to verbally respond to participants in both think-aloud conditions.

Procedure

Prior to participants' first test session, the test administrator discussed the usability lab's setup (i.e., room configuration, recording systems). Participants were handed a usability session packet and asked to follow along as the facilitator read the content out loud (Appendix F). Participants had read and agreed to the consent form (Appendix G). After providing consent, participants completed an online questionnaire regarding their basic demographics and computer use (Appendix H). The test administrator clearly explained the think-aloud method and had participants practice using the technique by asking a generic question. The think-aloud script was the only content from the usability session packet that was read for the second sessions. The thinking-aloud instructions were read to remind participants of what was expected as they performed the tasks.

A brief summary of the website was provided to the participant before receiving the first task as listed in Table 2. Participants were given each task on a note card and asked to read the task card out loud prior to beginning the task. Participants were not informed of which think-aloud protocol they had for each test session nor were they told the differences in protocols. The test administrator led them to

believe they were participating in a study to evaluate the usability of two websites that would be completed in two sessions. For each think-aloud condition, the test administrator interacted with the participants in the following ways:

Traditional

- Strictly adhered to Ericsson and Simon's (1993) framework
- Did not probe or interfere except for *Keep talking* if the participant fell silent for 15 seconds

Speech Communication

- Followed the think-aloud technique proposed by Boren and Ramey (2000) based on speech communication theory
- Used acknowledge tokens in form of *mm hmm* and *uh-huh* with intonation
- Probed with verbal tokens of *Mm hmm?* or *And now…?* if participant fell silent for more than 15 seconds and if the former questioning tone did not work

An online post-study questionnaire was given after each test session to understand participants' thoughts regarding the think-aloud method, e.g., comfort level, naturalness, and ease of remembering (Appendix I). The same questions about think-aloud were used in both sessions' post-study questionnaire, but additional questions about the moderator's style were asked in the second session's questionnaire (Appendix J). In this questionnaire, participants were asked if they noticed a difference in how the moderator verbally responded in each session and which think-aloud method they preferred.

Data Collection and Analysis

The author transcribed the test sessions and then followed steps 2-7 of verbal data analysis established by Chi (1997) indicated in Table 3.

Table 3. Steps taken for verbal data analysis

Verbal Data Analysis
2. Segmented the sampled protocols into utterances
3. Chose a coding scheme
4. Operationalized evidence in the coded protocols that constituted a mapping to some chosen formalism
5. Depicted the mapped formalism
6. Sought patterns in the mapped formalism
7. Interpreted the patterns

The granularity (i.e., the size of the utterance segment) varied in length but contained an individual topic. The segmentation was based on "semantic features, such as ideas, argument chains, topics of discussion, or impasses while solving problems" (Chi, 1997, p. 281). The verbalizations were read for meaning to determine the utterance segments. Because of the potential differences in participants' verbosity levels, it was important to focus on the content of participants' utterances rather than how much they verbalized, i.e., the number of words spoken (Chi, 1997). Therefore, utterance length was not evaluated in this study as it was more appropriate to measure the number of independent thoughts spoken (Chi, 1997).

The coding scheme inspired by Zhao and McDonald (2010) was used since their utterance categories were developed from a verbal analysis of concurrent think-aloud protocols. The "Other" category was added to include utterances that did not fit into one of the ten categories. Utterances were placed into one category. If an utterance had more than one possible category, the dominant category was chosen. The utterance categories and definitions are displayed in Appendix A.

Yang (2003) recommends that researchers use a contextualized perspective when segmenting and encoding verbal data. Therefore, the context of a verbalization was used during the verbal data analysis.

Contextual information such as video recordings and preceding or following utterances was utilized to categorize utterances. Meaning is not a result of discrete acts of recollection. It is contextual, entering straight into the "textures and strands of actions and events over time. Not only do present actions contain and inform past meanings, but they also anticipate future actions" (Yang, 2003, p. 106).

Once the verbalizations were segmented and coded, the differences in utterance counts per think-aloud condition were calculated using a Wilcoxon Signed Ranks Test. To determine the relevancies of utterances, the utterances were reviewed to see if they contained information that indicated user difficulty or causes for difficulty (Zhao & McDonald, 2010) and were useful statements for further usability evaluation. As a manipulation check, the number of utterances by the test administrator was counted per think-aloud method to ensure that the moderator uttered more back channels in the speech communication method. Each usability session was reviewed twice to verify that the appropriate utterance segmentations and categories were given for the participants' dialogue.

After each test session, participants were asked several questions regarding their testing experience such as their level of comfort participating in the usability test and thoughts about the test administration (Appendices I and J). Participants were not informed of the distinctions in think-aloud protocols until the end of the study as it could have influenced their answers. The preference questions were inspired by the Likert scale questions of Brush, Ames, and Davis (2004) and Olmsted-Hawala et al. (2010a). The differences in conditions' questionnaire answers were examined to determine the level of impact the think-aloud protocols had on participants.

Results

Participant Demographics and Online Behaviors

Seventeen student volunteers, aged between 20 to 30 years, from Rochester Institute of Technology (RIT) were selected to participant in this comparative study. Sixteen participants, eleven females and five males completed the usability study; one student was selected to be the pilot participant. The average age of the sixteen regular participants was 23 years old. 68.8% of participants were 20-22 years old, 18.8% were 23-25 years old, and 12.5% were 29-30 years old. Figure 3 displays the distribution of ages of the regular participants who completed the study.

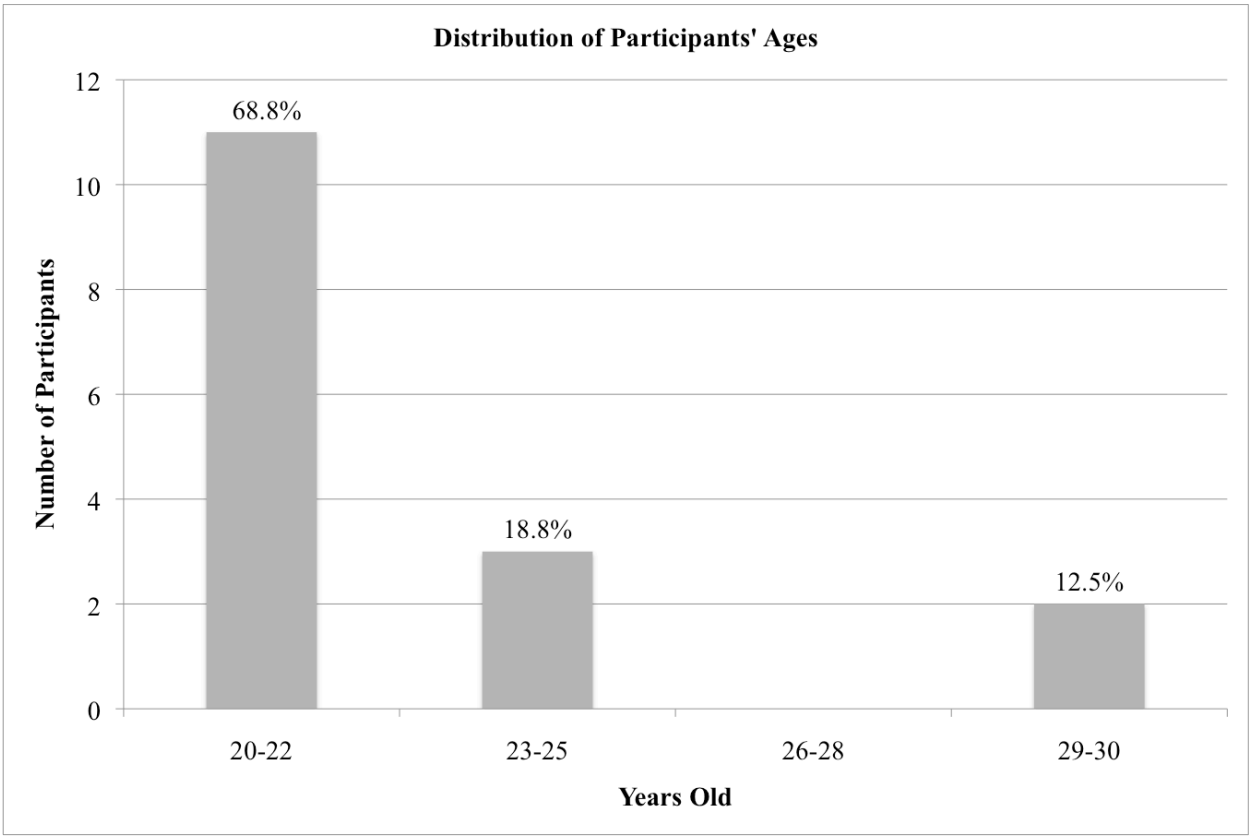

Figure 3. Distribution of participants' ages

All but one participant were native English speakers. The participant whose first language was not English rated herself to be excellent at reading and speaking English.

The participants said they typically spend 12 hours or more using the Internet, not counting e-mail each week. Of the sixteen participants, 50.0% spend 12-17 hours each week using the Internet and 50.0% spend 18 or more hours (Figure 4).

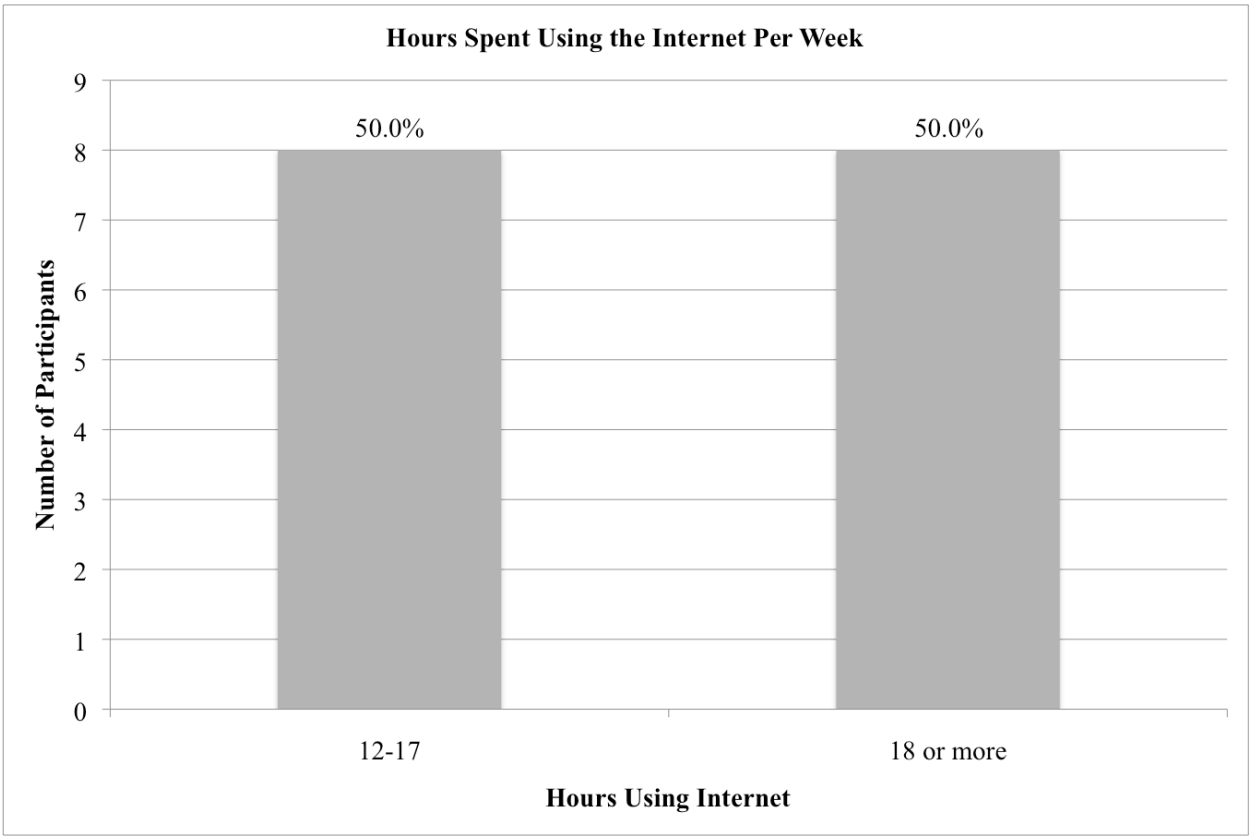

Figure 4. Distribution of hours spent using the Internet (excluding email) per week

As shown in Figure 5, the participants also go to 8 or more different websites a week. 31.3% visit 8 to 10 different websites per week while 68.8% visit 11 or more sites. Because of the purpose of this study, it was important that participants were comfortable browsing the Internet for non-email tasks on a weekly basis.

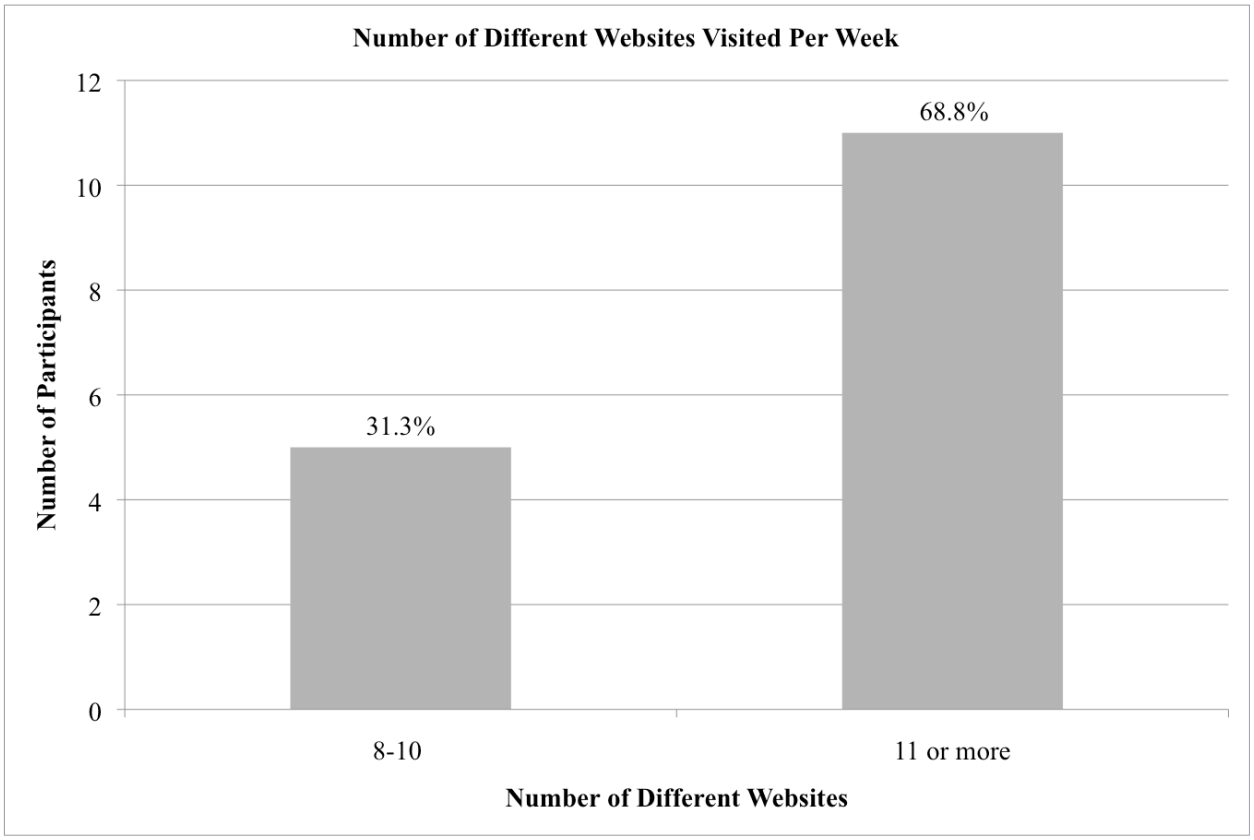

Figure 5. Distribution of the number of different websites visited per week

Pre-Study Questionnaire Analysis

There was a good distribution with how frequent participants used transportation and news websites (Figure 6). In the past six months, 31.3% of participants have been to 5 or more transportation (e.g., bus, train, place, cruise) websites, while 75.0% of participants have been to at least 5 news websites in the past six months. Since the test websites were counterbalanced during the study and each participant used both sites, this helped reduce the potential effect of participants feeling more comfortable using 13WHAM's website than DART's website given these pre-study usage ratings.

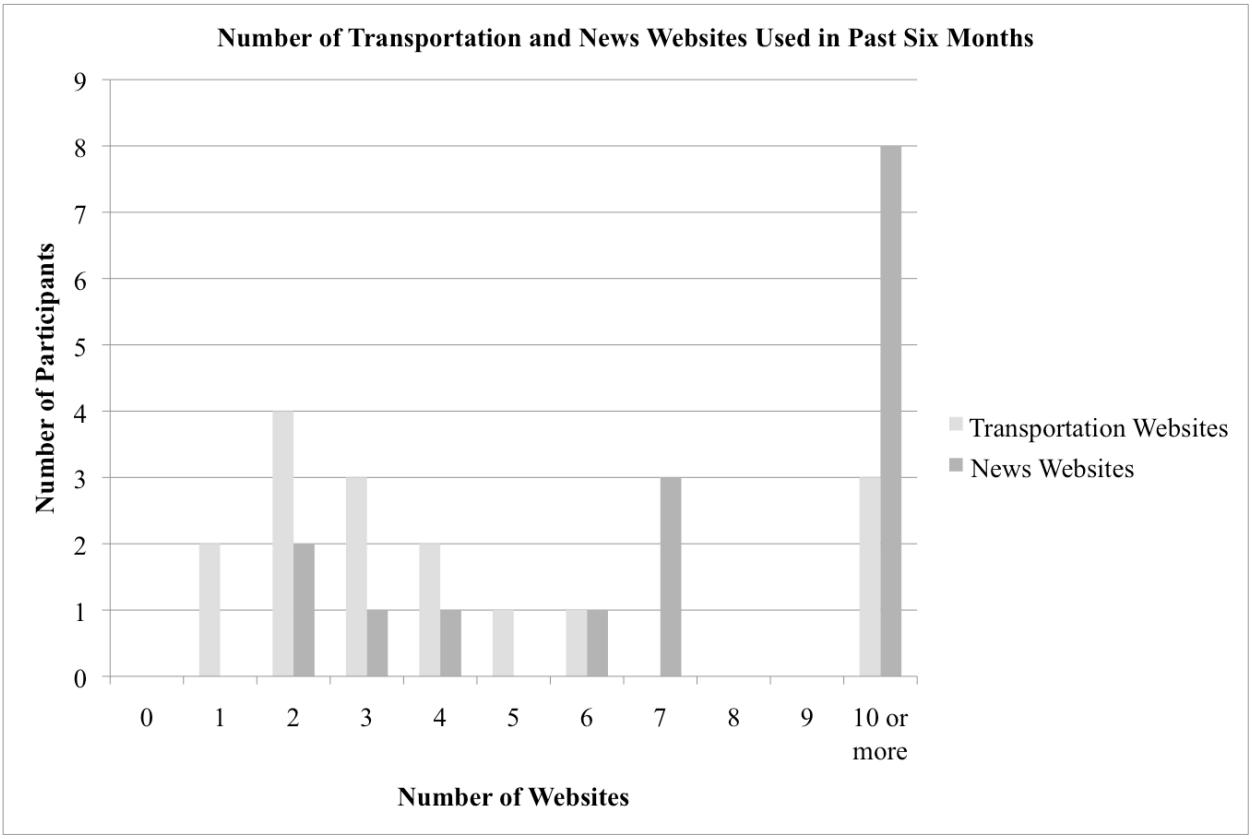

Figure 6. Distribution of transportation and news websites visited in past six months

All participants stated that they did not have prior experience in participating in a usability study nor in performing the think-aloud method.

Preference of Think-Aloud Method

All participants said that they have never participated in a study in which they were required to think aloud as they performed a set of tasks. After explaining the difference between the facilitation styles after the second sessions, only five of the sixteen participants admitted noticing the difference in how the facilitator was interacting with them as they performed the tasks in the two usability sessions. When asked if they preferred one think-aloud method to another, 43.8% of participants said they preferred the speech communication method while only 12.5% said they preferred the traditional method (Figure 7). Seven participants did not notice the difference in facilitation or have a preference in how the facilitator was interacting with them during the sessions. Since the presentation of think-aloud protocols was counterbalanced this reduced the tendency of participants to prefer the most recent think-aloud

method. There was no connection in regards to preference of think-aloud method and most recent protocol experienced.

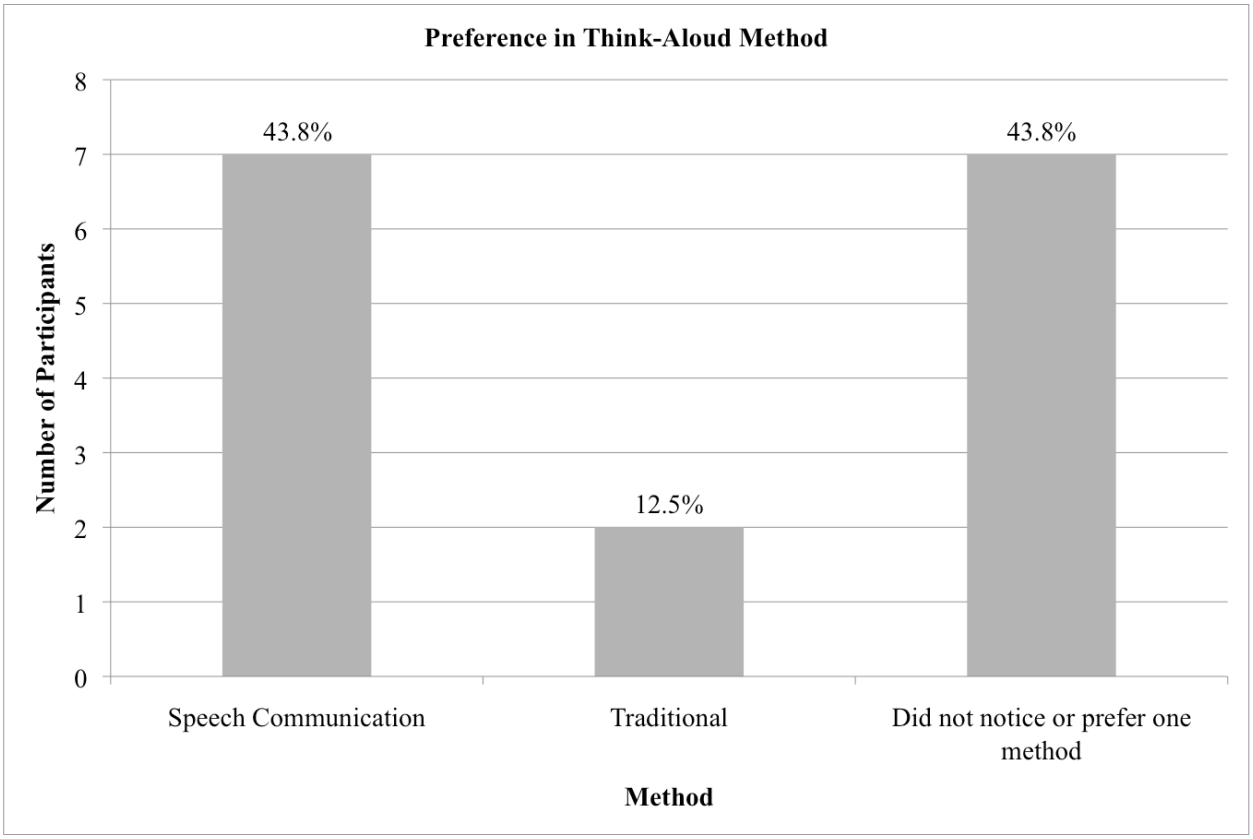

Figure 7. Participants' preference in think-aloud method

Participants who preferred the speech communication to the traditional method made comments about how they liked hearing the facilitator's back channels while performing the tasks. For example, participant nine said, "I felt like I was getting a little closer to the answer when I heard a response from the facilitator. I prefer hearing something from them; it definitely helps." Two participants preferred the traditional method because they thought the facilitator's acknowledgement tokens in the speech communication condition were somewhat bothersome. Participant six expressed this frustration with the facilitator's utterances, "It bothered me when she kept saying *mm hmm* and *uh-huh*." Appendix K lists participants' comments about their think-aloud preferences.

Post-Session Questionnaire Analysis

Overall, participants rated the speech communication method somewhat higher than the traditional method in terms of ease of remembering to think aloud, ease of concentrating on tasks, and feeling more natural thinking aloud. In contrast, participants were a bit more comfortable performing tasks and believed it was easier to articulate thoughts while thinking aloud in the traditional method. Figure 8 displays the average agreement ratings based on a Likert scale of strongly disagree (1) to strongly agree (7) for each post-session question pertaining to thinking aloud.

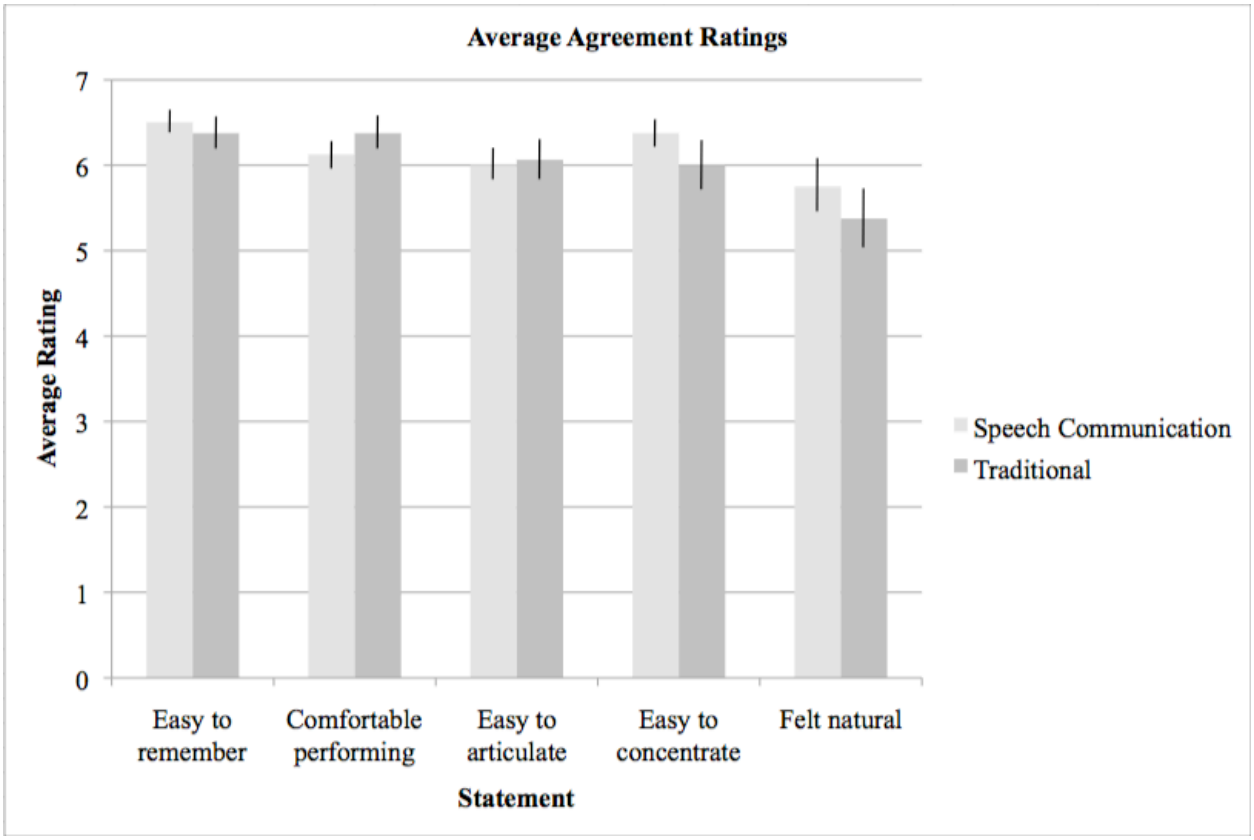

Figure 8. Average ratings of post-test questions for think-aloud methods

A paired t-test revealed that there was a significant difference ($p = 0.029$) with how natural participants felt thinking aloud while performing the tasks among the think-aloud conditions (Table 4). Participants felt more natural thinking aloud in the speech communication condition than they did when the facilitator was using the traditional think-aloud method. The facilitator's back channels of *mm hmm*

may have unconsciously made the participants feel more natural speaking their thoughts out loud compared to the facilitator's silence observed in the traditional sessions.

Table 4. Analysis of post-test questions for think-aloud methods

	Speech Communication		**Traditional**		
Statement	**Avg.**	**SD**	**Avg.**	**SD**	**P Value**
Easy to remember	6.500	.516	6.375	.719	.544
Comfortable performing	6.125	.719	6.375	.806	.216
Easy to articulate	6.000	.816	6.063	.929	.835
Easy to concentrate	6.375	.719	6.000	1.155	.164
Felt natural	5.750	1.438	5.375	1.586	.029*

*Significant difference achieved with $p < .05$

Participants were given the opportunity to provide additional comments regarding the think-aloud method at the end of each usability session. For the eight individuals who gave comments, their quotes are displayed in Appendix L. Three participants expressed frustration that they could not find an answer to the task and felt uncomfortable not finding the answer in front of the facilitator in both think-aloud conditions. For example, after experiencing the speech communication method participant nine said, "I felt a little uncomfortable because I felt kind of foolish for not being able to find certain things right away or at all." In contrast, five participants said they were comfortable speaking their thoughts out loud and it was simple to do so. Upon completing tasks from the traditional method, participant five stated, "Thinking aloud while doing the tasks was easier than I thought it would be. A very hassle-free experience."

Facilitator's Back Channels

As expected, because of the differences in protocol regarding the level of test administrator interaction, the test administrator uttered more frequently in the speech communication condition. In total, the facilitator uttered 241 back channels for the speech communication method's sessions. The facilitator did not have to say, *keep talking* during the traditional method as the participants never fell

silent for 15 seconds or more. During the speech communication method sessions, the facilitator solely used the *mm-hmm* back channel, as she felt most comfortable with this utterance compared to *uh-huh*.

The amount of facilitator's back channels fluctuated per participants' sessions because of the differences in time it took participants to complete the provided tasks. The longer it took participants to finish a task, the more back channels the facilitator uttered to keep the communication flowing. Figure 9 shows the number of *mm-hmms* spoken per participant's session during the speech communication condition and displays the wide range of back channels uttered. The average amount of acknowledgment tokens spoken during each participant's session was 15 back channels (SD = 6.30). The moderator uttered on average 5 *mm-hmms* per task (SD = 2.80).

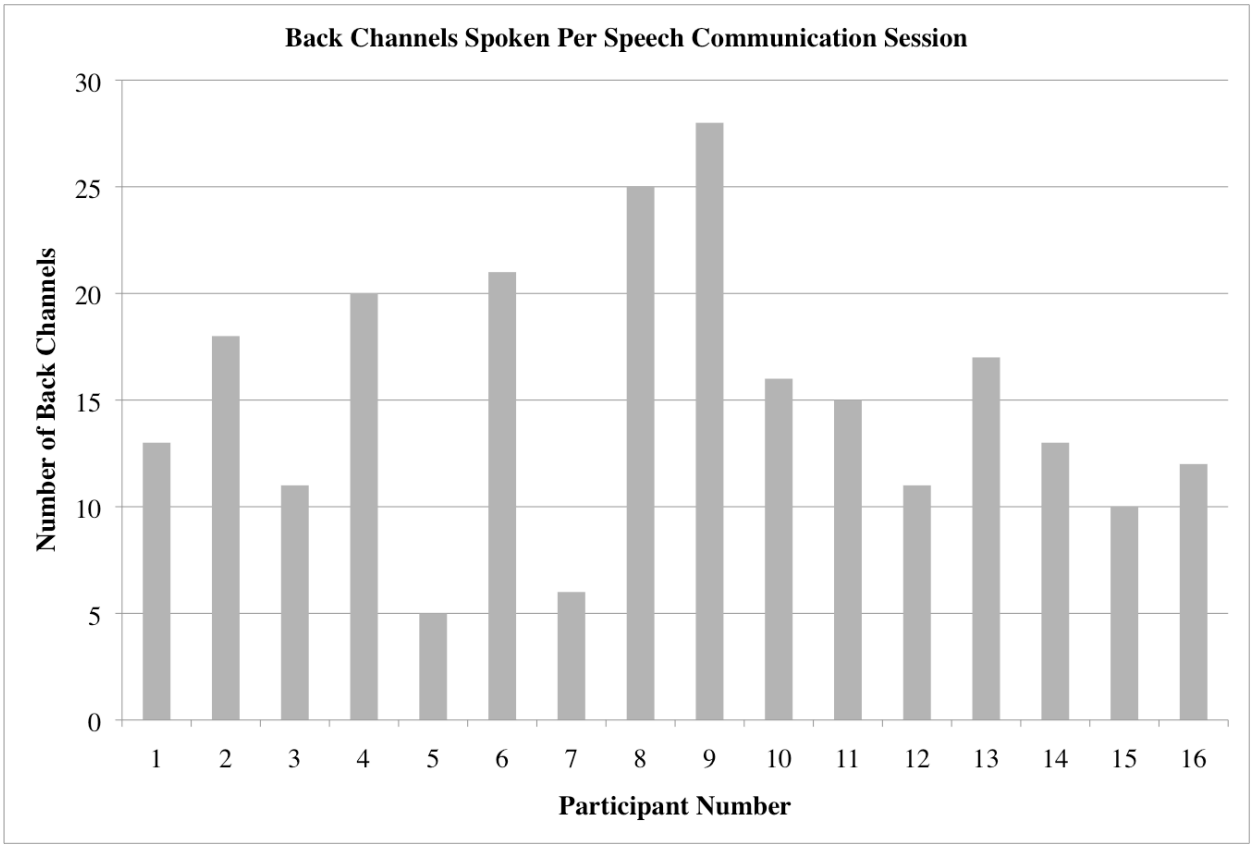

Figure 9. Number of back channels spoken per speech communication session

Utterance Analysis

The number of utterances per pre-defined utterance category was counted for each think-aloud method. As shown in Figure 10, the "Result Evaluation", "Action Description", "Action Explanation", and "Problem Formulation" categories had the largest utterance counts for both think-aloud methods. The traditional think-aloud method had more utterances in each category except for the "User Experience", "Causal Explanation", and "Recommendation" categories.

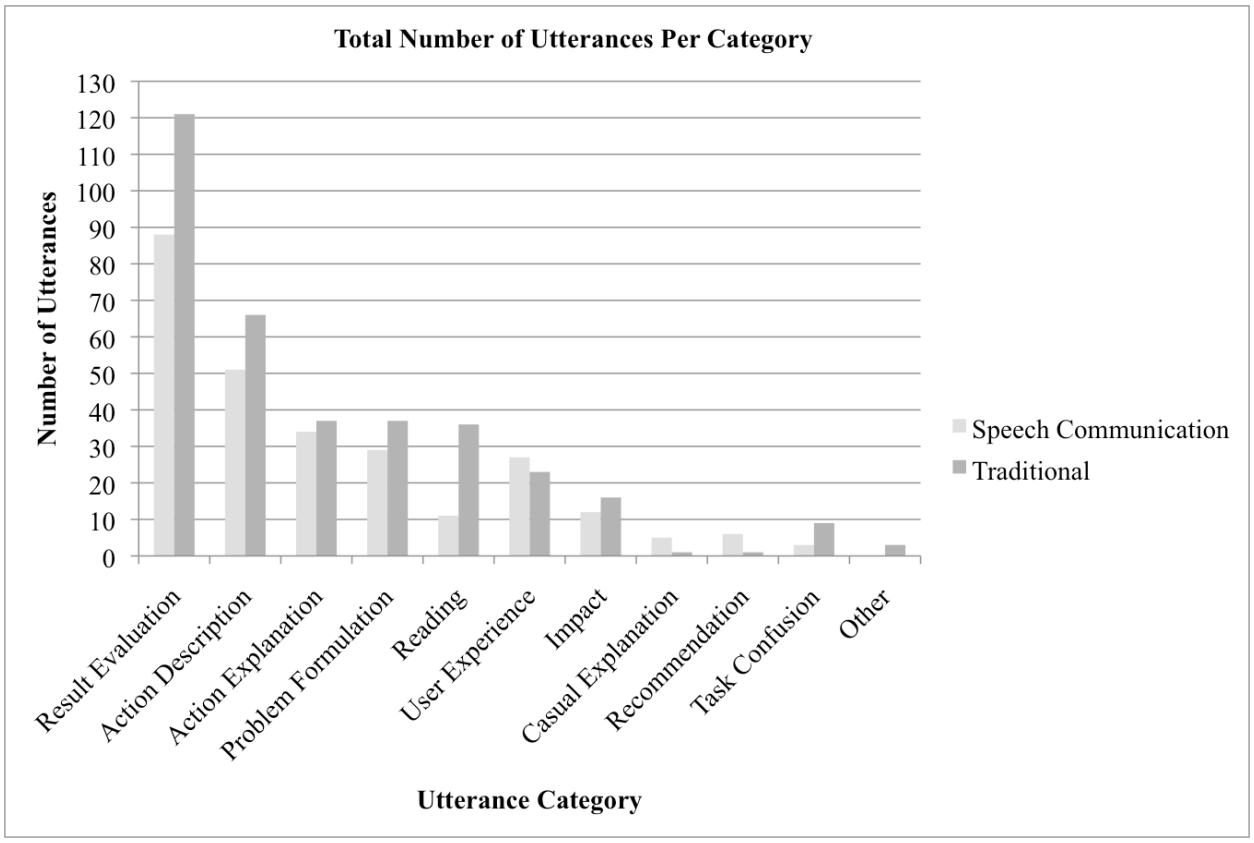

Figure 10. Total number of utterances per utterance category

Table 5 displays the ranking of utterance counts per category from highest to lowest counts for both conditions. The "Causal Explanation", "Recommendation", "Task Confusion", and "Other" categories each had less than 10 utterances for both think-aloud methods. Less than 4.5% of all utterances (speech communication: 5.3%, traditional: 4%) were in these particular utterance categories. These counts were relatively low compared to the more frequent categories. Despite how participants' statements were segmented into various utterance lengths, it is worth mentioning that the traditional

method had 84 more utterances than the speech communication method. A Wilcoxon Signed Ranks Test revealed that the total number of utterances per participant among the think-aloud methods approached a significant difference ($p = 0.077$).

Table 5. Total number of utterances per category for think-aloud methods

Speech Communication	**Counts**	**Traditional**	**Counts**
Result Evaluation	88	Result Evaluation	121
Action Description	51	Action Description	66
Action Explanation	34	Action Explanation	37
Problem Formulation	29	Problem Formulation	37
User Experience	27	Reading	36
Impact	12	User Experience	23
Reading	11	Impact	16
Recommendation	6	Task Confusion	9
Causal Explanation	5	Other	3
Task Confusion	3	Causal Explanation	1
Other	0	Recommendation	1
Total Count	266	**Total Count**	350

A Wilcoxon Signed Ranks Test was also performed to calculate the difference in the number of participant utterances per think-aloud method (Table 6). There were no significant differences per think-aloud methods for any of the utterance categories. It is worth noting that the "Result Evaluation" utterance category approached significance ($p = 0.051$) more than the other utterance categories. Unfortunately, because of the small number of utterances in some of the categories, the Z and P values were unable to be calculated.

Table 6. Categorical utterance analysis for think-aloud methods

Utterance Category	Speech Communication Avg./Participant	Traditional Avg./Participant	Z Value	P Value
Result Evaluation	5.500	7.563	1.950	0.051
Action Description	3.188	4.125	1.210	0.226
Action Explanation	2.125	2.313	0.180	0.857
Problem Formulation	1.813	2.313	0.680	0.497
Reading	0.688	2.250	-	-
User Experience	1.688	1.438	0.000	1.000
Impact	0.750	1.000	0.690	0.490
Causal Explanation	0.313	0.063	-	-
Recommendation	0.375	0.063	-	-
Task Confusion	0.188	0.563	-	-
Other	0.000	0.188	-	-

Relevant Utterance Analysis

All participants' utterances were also analyzed in terms of relevancy for usability evaluation. Relevant utterances were statements that included user frustration or difficulty (Zhao & McDonald, 2010). These types of utterances are useful for determining the problems users have with a system and encourages the discussion on how to better design for the user.

As shown in Table 7, of the total utterances in the traditional method, only 20.0% of them were relevant while 28.6% were relevant in the speech communication method. In total, only 23.7% of all utterances spoken in the speech communication and traditional methods were relevant. A Wilcoxon Signed Ranks Test determined there was no statistical difference between the think-aloud methods in terms of total relevant utterances per participant (p = .984). Over 75% of relevant utterances in the traditional method came from the "Problem Formulation" (35.7%), "User Experience" (20.0%), and

"Impact" (20.0%) categories. Similarly, over 75% of relevant utterances in the speech communication method came from the same utterance categories of "Problem Formulation" (31.6%), "User Experience" (30.3%), and "Impact" (14.5%).

In addition, Table 7 displays the relevancy of each utterance category for the think-aloud methods. The relevancy was determined by dividing the number of relevant utterances by the total number of utterances produced in that particular category. Furthermore, the higher the relevancy percentage, the more useful the category (Zhao & McDonald, 2010).

The relevancy of utterance category was somewhat comparable in both conditions, with "Causal Explanation", "Impact", "User Experience", and "Problem Formulation" being ranked as the most relevant categories. Five of six utterances in the "Recommendation" category were relevant for the speech communication while there were no relevant utterances in that particular category for the traditional method. Category relevancy for the "Result Evaluation" category was similar for both methods with 9.1% relevancy for the speech communication method and 9.9% for the traditional think-aloud condition. The "Action Explanation" and "Action Description" categories had no relevant utterances in the speech communication condition, compared to 5.4% relevancy for "Action Explanation" and 3.0% relevancy for "Action Description" in the other think-aloud method. For both methods, the "Reading", "Task Confusion", and "Other" categories did not have any relevant utterances.

Table 7. Relevant utterance analysis for think-aloud methods

	Speech Communication			**Traditional**		
Utterance Category	**Relevant**	**Total**	**Category Relevancy**	**Relevant**	**Total**	**Category Relevancy**
Result Evaluation	8	88	9.1%	12	121	9.9%
Action Description	0	51	0.0%	2	66	3.0%
Action Explanation	0	34	0.0%	2	37	5.4%
Problem Formulation	24	29	82.8%	25	37	67.6%
Reading	0	11	0.0%	0	36	0.0%
User Experience	23	27	85.2%	14	23	60.9%
Impact	11	12	91.7%	14	16	87.5%
Causal Explanation	5	5	100.0%	1	1	100.0%
Recommendation	5	6	83.3%	0	1	0.0%
Task Confusion	0	3	0.0%	0	9	0.0%
Other	0	0	-	0	3	0.0%
Total	76	266	28.6%	70	350	20.0%

Discussion

Summary of Findings

Overall, the results of this study conclude that there were no significant differences in the number of utterances per predefined utterance category between the think-aloud methods. It is noteworthy that although fewer utterances were produced in the speech communication method, it did have a higher percentage of total relevant utterances than the traditional condition. Also, even though participants produced fewer utterances per category in general than in the traditional method, the speech communication condition had larger ratios of relevant utterances in the utterance groups that would lead to further usability evaluation such as "Impact", "User Experience", and "Problem Formulation". The moderator's back channels spoken in the speech communication may have some influence on the utterances produced given these differences.

Participants' utterances primarily contained procedural descriptions such as "Result Evaluation", "Action Description", and "Action Explanation" categories in both the speech communication and traditional think-aloud method. They mainly discussed and explained the actions they took as well as verbalized their understanding of the websites' content. Of all utterances produced in the study, both think-aloud conditions produced less than 30% relevant utterances for further usability analysis.

Participants seemed to feel more natural in the speech communication condition than the traditional method with a significant difference shown. This outcome may have occurred because participants are frequently exposed to the use of back channels during regular peer-to-peer conversations. Participants may have felt that their comments were being heard and acknowledged; therefore, they spoke more about what frustrated them about the website. The moderator's silence in the traditional method perhaps caused the participants to feel more self-conscious while performing the tasks, leading to fewer relevant utterances. When asked which method was preferred, seven of the sixteen participants said that they preferred the speech communication method to the traditional method, while only two participants preferred the traditional think-aloud method. Those participants who preferred the traditional moderation were bothered by the frequent use of *mm hmm* in the speech communication condition. For example,

participant sixteen stated, "I prefer the keep talking method. The *uh-huh* or *mm hmm* method irritates me because I find it pompous and condescending."

Those who preferred the speech communication method made comments that would suggest the method could lead to different task satisfaction or task success ratings compared to the traditional think-aloud method. Several participants mentioned how they felt the back channels were acknowledgements of being on the right track when completing the tasks. Participant four stated, "I prefer the second method over the first method because it seemed like I was actually doing the right thing, rather than just blindly fumbling through the tasks." This study did not analyze post-task satisfaction ratings or task completion rates, which may have been affected based on their comments.

Comparison to Past Research

Results of this comparative study had some similarities and differences when compared to Zhao and McDonald's (2010) concurrent think-aloud study. Zhao and McDonald discovered that for seven out of ten utterance categories (excluding the "Other" category) the interactive think-aloud method produced significantly more utterances than the traditional method. The results of our study did not show any significant differences between the number of utterances spoken in each category between the speech communication and traditional methods. One explanation for why this occurred is because the speech communication method varies only slightly from the traditional method by having the moderator use back channels of *mm hmm*, compared to frequent exploratory-type interventions in the interactive think-aloud method. In the interactive condition, the moderator would interject with various intervention types including asking for explanation, suggestions, or clarifications.

Zhao and McDonald (2010) discovered that the "Causal Explanation", "Problem Formulation", "User Experience", and "Recommendation" utterance categories were the most relevant categories for problem detection when they compared the interactive think-aloud to the traditional method. Even though different think-aloud conditions were used for this study, the "Causal Explanation", "Problem Formulation", and "User Experience" categories held higher relevancy percent wise as well. In contrast to Zhao and McDonald's research, the "Impact" utterance category had high relevancy for both the

speech communication and traditional think-aloud methods. Comparable to the results of this study, Zhao and McDonald also discovered the "Action Explanation", "Action Description", "Reading", and "Task Confusion" categories had the lowest category relevancy.

For both the speech communication and traditional think-aloud method, the relevant utterances were predominantly from the "Problem Formulation", "User Experience", and "Impact" categories. Likewise, Zhao and McDonald's (2010) research revealed that the traditional think-aloud condition produced the majority of relevant utterances from the "Problem Formulation", "User Experience", and "Result Evaluation" categories.

Only about 10% of all utterances from the interactive and traditional think-aloud methods were deemed relevant in Zhao and McDonald's (2010) study, whereas the results of our study concluded that 23.7% of utterances indicated user difficulty or frustration. Both studies reveal a relatively low percentage of utterances that lead to further usability analysis; therefore, it is recommended that researchers continue investigating these various think-aloud methods to promote higher utterance relevancy.

Limitations and Suggestions

Unfortunately, because of the limited amount of resources available for the researcher of this study, there were several constraints that may have affected the results. The researcher did not have a second coder because of time constraints, and this limitation may have affected the reliability of the utterance segmentation and categorization. Having a second or third coder would have improved the validity of the results as other members could have reviewed the analysis to assure that appropriate utterance segmentation and categorization was performed. For the sole researcher to reduce subjectivity, the utterance codes were not analyzed until both methods had been transcribed, segmented, and coded.

Unfortunately, because the sample size of this study was low, it was difficult to perform statistical testing on all of the data. While our study revealed some important insights about the speech communication and traditional think-aloud protocols, our results could be strengthened with an increase in the number of participants. Another limitation was that participants were not able to meet exactly a

week apart because of conflicts with their class schedules. Scheduling participants the same time of day a week apart may have also improved the validity of the results.

Conclusion

During in-person or remote usability tests, it is critical to try and receive as much quality feedback from participants as possible given the time constraints. It would be of great value for researchers to determine which concurrent think-aloud method would lead to the most relevant utterances. The elicited relevant utterances are the key to improving the usability of the particular software or website. This research provided some new insight that moderator interactions may affect participants' verbalizations and feelings towards moderation.

Despite the fact that participants felt more natural and preferred the speech communication method, the results of the study do not necessarily promote the adoption of this method to the traditional think-aloud method. The speech communication think-aloud method did not yield significant differences in the utterance counts per utterance category and relevancy when compared to the traditional think-aloud method.

Future Research

Further research should be conducted to compare the various think-aloud methods to determine which method would gather the most useful information from participants. As previously mentioned by Boren and Ramey (2000), it would be interesting to see if the number of moderator back channels or type of back channels (i.e., *mm hmm*, *uh-huh*) affects how participants think aloud and the quality of their responses. Future research also might investigate one of the think-aloud conditions with various instructional scripts as they may have an affect on how the individuals talk aloud during their session. More specifically, does the amount of think-aloud practice affect the quality of verbalizations? Should participants be instructed to avoid discussing procedural comments while thinking aloud and trained to speak about difficulties they are experiencing? Different types of think-aloud instruction and moderator interventions should be researched to enhance the quality of participants' verbalizations.

Acknowledgments

I would like to thank several individuals for their support in receipt of my Master's Degree. To my advisors, Dr. Evelyn Rozanski, Dr. Michael Yacci, and Dr. Cecilia Ovesdotter Alm for their wisdom and guidance. To my family and friends, thank you for your encouragement and support.

References

Anderson, C. (2004). How much interaction is too much? Retrieved from http://www.stcsig.org/usability/newsletter/0404-howmuchinteraction.html

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261-278. doi: 10.1109/47.867942.

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. *Human Factors and Ergonomics Society Annual Meeting Proceedings* (Vol. 34, pp. 1270–1274). Human Factors and Ergonomics Society. Retrieved from http://www.ingentaconnect.com/content/hfes/hfproc/1990/00000034/00000017/art00020

Brush, A. J. B., Ames, M., & Davis, J. (2004). A comparison of synchronous remote and local usability studies for an expert interface. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04* (p. 1179). New York, New York, USA: ACM Press. doi: 10.1145/985921.986018.

Carter, P. (2007). Liberating usability testing. *Interactions*, *14*(2), 18-22. doi: 10.1145/1229863.12 29864.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, *6*(3), 271-315. doi: 10.1207/s15327809jls0603_1.

Clark, H., & Schaefer, E. (1989). Contributing to discourse. *Cognitive Science*, *13*, 259-294.

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions On Professional Communication*, *53*(3), 202-215.

Drummond, K., & Hopper, R. (1993). Some uses of yeah. *Research on Language and Social Interaction*, 26 (2), 203-212.

Ericsson, A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.

Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, *9*, 205-217.

Hertzum, M., Hansen, K., & Andersen, H. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*(2), 165-181. Taylor & Francis. doi: 10.1080/01449290701773842.

Krahmer, E., & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*, *47*(2), 105-117. doi: 10.1109/TPC.2004.828205.

Nielsen, J. (1993). *Usability Engineering*. Cambridge, MA: AP Professional.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-241.

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. *Proceedings of the 6th conference on Designing Interactive systems* (p. 209–218). ACM. Retrieved February 8, 2011, from http://portal.acm.org/citation.cfm?id=1142405.1142439.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010a). Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. *Proceedings of the 28th international conference on Human factors in computing systems* (p. 2381–2390). ACM. Retrieved January 15, 2011, from http://portal.acm.org/citation.cfm?id=1753326.1753685.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010b). Think-aloud protocols: Analyzing three different think-aloud protocols with counts of verbalized frustrations in a usability study of an information-rich Web site. *Professional Communication Conference (IPCC), 2010 IEEE International* (p. 60–66). IEEE. Retrieved January 26, 2011, from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5529815.

Ramey, J., Boren, T., Cuddihy, E., Dumas, J., Guan, Z., Haak, M. J. van den, et al. (2006). Does think aloud work?: How do we know? *CHI'06 extended abstracts on Human factors in computing systems*

(p. 45–48). ACM. Retrieved February 16, 2011, from http://portal.acm.org/citation.cfm?id=1125451.1125464.

Rubin, J., Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed.). Indianapolis, IN: Wiley Publishing, Inc.

Sauro, J. (2011, February 14). *10 tips for benchmark usability tests*. Retrieved from http://www.measuringusability.com/blog/benchmark-tips.com

Stone, D., Jarrett, C., Woodroffe, M., & Minocha, S. (2005). *User interface design and evaluation.* San Francisco, CA: Morgan Kaufmann.

Tamer, H. M. (1998). How (much) to intervene in a usability testing session. *Common Ground* 8(3), 11-15.

van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2006). Constructive interaction: An analysis of verbal interaction in a usability setting. *IEEE Transactions On Professional Communication*, *49*(4), 311-324.

van den Haak, M. J., & de Jong, M. D. T. (2003). Exploring two methods of usability testing: Concurrent versus retrospective think-aloud protocols. *IEEE International Professional Communication Conference, 2003. IPCC 2003. Proceedings.*, 3 pp. IEEE. doi: 10.1109/IPCC.2003.1245501.

Yang, S. C. (2003). Reconceptualizing think-aloud methodology: Refining the encoding and categorizing techniques via contextualized perspectives. *Computers in Human Behavior*, *19*, 95-115.

Zhao, T., & McDonald, S. (2010). Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10* (pp. 581-590). New York, New York, USA: ACM Press. doi: 10.1145/1868914.1868979.

Appendices

Appendix A. Utterance categories and their definitions (Zhao & McDonald, 2010, p. 584)

Categories	**Definitions**
Reading	"Read out texts and links"
Action Description	"Describe what they were going to do or just did"
Action Explanation	"Explain the reason(s) for executing or going to execute or executed certain actions"
Result Evaluation	"Summarize understanding or give evaluation of content, links or outcomes of actions"
User Experience	"Express positive or negative feelings, aesthetic preferences towards the websites and recall of past experiences"
Problem Formulation	"Verbalize difficulties, including utterances that participants indicate uncertainty; and utterances that participants not only express a negative feeling or disapproval, but also indicate that it was caused by system based issue(s)"
Causal Explanation	"Explain what had caused the difficulties"
Impact	"Indicate outcomes or impacts caused by difficulties encountered, including the repeated mention of a difficulty, and restart the task"
Recommendation	"Give recommendations on how to improve the interface or solutions to difficulties experienced"
Task Confusion	"Indicate confusion or misunderstanding about interface tasks"
Other	Utterances that do not fit into one of the categories

Appendix B. Participant recruitment email

Hello RIT students,

I am seeking individuals to participate in my thesis research regarding the usability of websites. The study will be conducted in two separate 30-minute sessions in a usability lab on campus (Golisano College 70-2293). If you are interested in participating, please fill out this 5-minute screening survey: http://edu.surveygizmo.com/s3/529225/Participant-Recruitment-Screener.

If you meet the criteria I am seeking for the purposes of this evaluation, you will be contacted by email with further information regarding the study. If selected to participate, you will receive a $10 Java Wally's gift card after the usability sessions are complete. The gift card is redeemable at any Java's location.

I appreciate your willingness to help with my research. Best of luck these last few weeks of spring quarter! If you have any questions, please feel free to contact me at klg5708@rit.edu.

Thank you,

Katie

Appendix C. Participant recruitment screener

Thank you for considering participating in a usability study. This survey will take you approximately 5 minutes to complete.

Full Name (Last, First): _____

Email: _____

1. Are you currently a student at RIT?
    a. Yes
    b. No
2. Within what college do you study?
    a. College of Applied Science and Technology
    b. E. Philip Saunders College of Business
    c. B. Thomas Golisano College of Computing and Information Sciences
    d. Kate Gleason College of Engineering
    e. College of Imaging Arts and Sciences
    f. College of Liberal Arts
    g. College of Science
    h. National Technical Institute for the Deaf
    i. Golisano Institute for Sustainability
    j. Center for Multidisciplinary Studies
    k. University Studies
3. What is your gender?
    a. Male
    b. Female
4. What age category do you fall into?
    a. 17 or younger
    b. 18-23

c. 24-29

d. 30 or older

5. Will you need assistance in using a computer mouse or keyboard?

   a. Yes

   b. No

6. Can you view a computer screen without difficulty?

   a. Yes

   b. No

7. Will you need a sign-language interpreter to facilitate communication during the study?

   a. Yes

   b. No

8. By means of computer (desktop or laptop), how many hours do you spend using the Internet, not counting e-mail each week?

   a. 5 or less

   b. 6-11

   c. 12-17

   d. 18 or more

9. By means of computer (desktop or laptop), how many different websites do you go to each week?

   a. 4 or less

   b. 5-7

   c. 8-10

   d. 11 or more

10. Have you ever participated in a usability study before?

    a. Yes

    b. No

11. Are you willing to come for two separate 30-minute sessions? The sessions will be approximately a week apart and be arranged to fit your schedule.
    a. Yes
    b. No
12. Are you willing to have your voice and computer screen be recorded for analysis purposes only? Your information will be kept confidential.
    a. Yes
    b. No
13. Will you be in or near the Rochester area this May and June? Due to scheduling difficulties of the usability lab, you may be asked to participate in the study during early summer.
    a. Yes
    b. No

Thank you for taking the time to complete this survey. If you have been selected to participate in this study, you will be contacted within a week with further information. If asked to partake in this research, you will receive a $10 Java Wally's gift card after completing both usability sessions.

Appendix D. The test protocol and task sequences for week 1

Website	**Task 1**	**Task 2**	**Task 3**
1. 13WHAM http://www.13wham.com/	Find A Lake Temperature	Find B Lottery Numbers	Find C Gas Price
2. DART http://www.dart.org/	Find D Shuttle Number	Find E College Price	Find F Irving Store Locations

Week 1

Participant	**Think-Aloud Protocol**	**Website**	**Task 1**	**Task 2**	**Task 3**
1	Traditional	1	Find A	Find B	Find C
2	Traditional	1	Find A	Find C	Find B
3	Traditional	1	Find B	Find A	Find C
4	Traditional	1	Find B	Find C	Find A
5	Speech Communication	1	Find C	Find A	Find B
6	Speech Communication	1	Find C	Find B	Find A
7	Speech Communication	1	Find A	Find B	Find C
8	Speech Communication	1	Find A	Find C	Find B
9	Traditional	2	Find D	Find E	Find F
10	Traditional	2	Find D	Find F	Find E
11	Traditional	2	Find E	Find D	Find F
12	Traditional	2	Find E	Find F	Find D
13	Speech Communication	2	Find F	Find D	Find E
14	Speech Communication	2	Find F	Find E	Find D
15	Speech Communication	2	Find D	Find E	Find F
16	Speech Communication	2	Find D	Find F	Find E

Appendix E. The test protocol and task sequences for week 2

Website	**Task 1**	**Task 2**	**Task 3**
1. 13WHAM http://www.13wham.com/	Find A Lake Temperature	Find B Lottery Numbers	Find C Gas Price
2. DART http://www.dart.org/	Find D Shuttle Number	Find E College Price	Find F Irving Store Locations

Week 2

Participant	**Think-Aloud Protocol**	**Website**	**Task 1**	**Task 2**	**Task 3**
1	Speech Communication	2	Find E	Find D	Find F
2	Speech Communication	2	Find E	Find F	Find D
3	Speech Communication	2	Find F	Find D	Find E
4	Speech Communication	2	Find F	Find E	Find D
5	Traditional	2	Find D	Find E	Find F
6	Traditional	2	Find D	Find F	Find E
7	Traditional	2	Find E	Find D	Find F
8	Traditional	2	Find E	Find F	Find D
9	Speech Communication	1	Find B	Find A	Find C
10	Speech Communication	1	Find B	Find C	Find A
11	Speech Communication	1	Find C	Find A	Find B
12	Speech Communication	1	Find C	Find B	Find A
13	Traditional	1	Find A	Find B	Find C
14	Traditional	1	Find A	Find C	Find B
15	Traditional	1	Find B	Find A	Find C
16	Traditional	1	Find B	Find C	Find A

Appendix F. Script for the first usability testing sessions

Usability Study of Websites

Welcome

Thank you for your willingness to participate in my research regarding the usability of websites. This study will be comprised of six parts:

1. Overview
2. Informed Consent
3. Background Questionnaire
4. Thinking Aloud
5. Tasks
6. Post-Study Questionnaire

Overview

This study is designed to help me understand how easy or difficult it is to navigate and find information on websites. In order for me to learn about what works and what does not work on these websites, I will provide a set of tasks for you to perform. After performing the tasks, you will be asked to fill out a questionnaire regarding your session.

I would like to stress that the goal of the study is not to assess you or your abilities but rather to evaluate the usability of the websites.

As the facilitator I'll be taking notes and will be recording your voice and computer screen to make sure I've collected accurate feedback. This data will be kept confidential and used for analysis purposes only.

Your participation is completely voluntary, and you may discontinue at any time. This session should last approximately 30 minutes.

Informed Consent

Before we begin, you'll need to read and sign this consent form. It summarizes and explains what I just discussed.

Background Questionnaire

Please fill out this background questionnaire regarding your demographic information and computer use.

Thinking Aloud (E. Olmsted-Hawala, personal communication, January 25, 2011)

In this study I'm interested in what you think about when you work at finding answers to the provided task questions. In order to do this I'm going to ask you to THINK ALOUD as you perform the tasks. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the task until you give an answer. In other words, I'd like you to tell me what you're doing, what you're expecting to see happen, what you're going to do, and why. I'd like you to give me your open impressions, both good and bad of what you see and experience on the website. It's most important that you keep talking. If you're silent for any long period of time I'll remind you to talk. I'm not permitted to answer questions during the session. Do you understand what I want you to do?

Let's take a moment to practice thinking aloud. Please think aloud as you answer the following question: How many windows are there in your place of residence?

Do you have any questions about the thinking aloud process we've just practiced?

Tasks

There are three tasks that I'd like you to perform. I'll have you begin each task by reading the task question out loud. As you work remember to think aloud. You'll do the tasks one at a time. Once you have found the information you are looking for please state your answer out loud. After you completed a question, I will give you the next task to perform. Do you have any questions before we begin?

Post-Study Questionnaire

Please fill out this online questionnaire regarding your experiences with the session.

Appendix G. Informed consent

Consent Form to Participate in Usability Study

You are invited to participate in a usability study evaluating news and transportation websites. The purpose of this consent form is to give you the information you will need to help you decide whether to participate in this study. Please read the form carefully and ask any questions before agreeing to partake in the study.

Purpose of Study

The purpose of this study is to evaluate the usability of news and transportation websites. By assessing the usability of these websites will help me better understand users' needs and expectations regarding online information. This is not a test of you or your abilities rather an evaluation of the usefulness of the websites.

About the Sessions

For this usability study, you will perform a set of tasks on various websites. The study will be conducted in two separate sessions of 30 minutes each. You will be asked to fill out a questionnaire concerning your demographic information and computer use. In addition, you will be asked to complete post-session questionnaires about your experiences.

Data Confidentiality

Data collected during the study will be kept confidential and analyzed in an anonymous manner. Your voice and computer screen will be recorded during each session for analysis purposes only. Research records will be saved in protected files, and I will be the only one with access to the records.

Risks and Benefits

There are no potential risks to you in this study. There are no benefits to you, but as a participant you will gain experience with being involved in a usability evaluation study.

Compensation

You will receive a $10 *Java Wally's* café card for your participation after the second session. The café card is redeemable at any of *Javas'* locations.

Contact Information

If you have questions later about the research, you can contact the researcher, Katie Greiner, via email at klg5708@rit.edu. If you have questions about your rights as a research participant, you can contact Heather Foci from RIT's Human Subjects Research Office by phone at 585-475-7673 or via email at hmfsrs@rit.edu.

Your participation in this study is voluntary. You may refuse to participate or may withdraw from the study at any time without penalty or loss of benefits to which you are otherwise entitled. If you have read the above information and agree to participate in the usability study, please indicate your agreement by signing below:

Printed name of subject Signature Date

Printed name of researcher Signature Date

Appendix H. Background questionnaire

1. How old are you?

_____ years old

2. Is English your native language?

a. Yes

b. No

If No, then continue to question 3. If Yes, continue to question 6.

3. How old were you when you started learning English?

_____ years old

4. Please rate your proficiency in reading English.

Very Poor Poor Fair Good Very Good Excellent

5. Please rate your proficiency in speaking English.

Very Poor Poor Fair Good Very Good Excellent

6. By means of computer (desktop or laptop), how many transportation (i.e., bus, train, plane, cruise) website(s) have you used in the past 6 months?

a. 0

b. 1

c. 2

d. 3

e. 4

f. 5

g. 6

h. 7

i. 8

j. 9

k. 10 or more

7. By means of computer (desktop or laptop), how many news website(s) have you used in the past 6 months?

a. 0

b. 1

c. 2

d. 3

e. 4

f. 5

g. 6

h. 7

i. 8

j. 9

k. 10 or more

Appendix I. Post-session 1 questionnaire

To what extent do you agree or disagree with the following statements? Please select your answer.

1. It was easy to remember to think aloud while performing the tasks.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

2. I was comfortable performing the tasks while thinking aloud my thoughts.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

3. It was easy to articulate what I was thinking while performing the tasks.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

4. It was easy to concentrate on performing the tasks while thinking aloud.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

5. It felt natural to think aloud while performing the tasks.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

6. Please share any additional comments regarding this usability session in the space below.

Appendix J. Post-session 2 questionnaire

To what extent do you agree or disagree with the following statements? Please select your answer.

1. It was easy to remember to think aloud while performing the tasks.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

2. I was comfortable performing the tasks while thinking aloud my thoughts.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

3. It was easy to articulate what I was thinking while performing the tasks.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

4. It was easy to concentrate on performing the tasks while thinking aloud.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

5. It felt natural to think aloud while performing the tasks.

   Strongly disagree 1 2 3 4 5 6 7 Strongly agree

6. Please share any additional comments regarding this usability session in the space below.

7. Prior to these usability sessions, have you ever participated in a study in which you were required to think aloud as you perform tasks?

   a. Yes

   b. No

During each usability session, the facilitator used a different method to verbally interact with you while you were performing the tasks. In one method, the facilitator did not talk except if she needed to remind you to think aloud by saying *keep talking*. While in the other method, the facilitator continuously responded to your comments with short verbalizations like *uh-huh* or *mm hmm*.

8. Did you notice this difference in how the facilitator was interacting with you as you performed the tasks in the two usability sessions?

   a. Yes

   b. No

9. Do you prefer one method to another? Please explain your answer and which method you prefer if

applicable.

10. Please share any additional comments regarding these facilitation methods in the space below.

Appendix K. Participants' preference of think-aloud method and comments about facilitation

Do you prefer one method over another? Please explain your answer and which method you prefer if applicable.	**Please share any additional comments regarding these facilitation methods in the space below.**
"I think I preferred the second method, it didn't make me feel as rushed the first session to find what I was looking for." P1	
"The slight response was probably a bit more comfortable but I'm somewhat used to speaking thoughts out loud with no verbal feedback as I do stream online as I play video games." P2	
"I'm not sure which one I would have preferred, as I don't recall being prompted to continue reading out loud, nor did I ask questions. If I had to guess, I would say I would prefer responses to questions rather than not having them answered at all." P3	"I didn't notice the change in the facilitator's method, as I kept talking continuously, and had no need to ask questions." P3
"I prefer the second method over the first method, because it seemed like I was actually doing the right thing, rather than just blinding fumbling through the tasks." P4	
"No, not really, both methods were comfortable." P5	
"It bothered me when she kept saying *mm hmm* and *uh-huh*." P6	
"Continuous responding makes it more comfortable than talking into silence." P7	
"I didn't notice enough of a difference to prefer one or the other." P8	
"I felt like I was getting a little closer to the answer when I heard a response from the facilitator. I prefer hearing something from them; it definitely helps." P9	
"I didn't notice. I was too busy focused trying to	

find the information." P10	
"I didn't even notice." P11	
"I preferred the second method because it seems as though the facilitator was helping me along and agreeing with my comments." P12	
N/A P13	
"The acceptance helped me feel more comfortable in what I was doing. Almost as if I was on the right track." P14	
"No, either is fine." P15	
"I prefer the keep talking method. The *uh-huh* or *mm hmm* method irritates me, because I find it pompous and condescending." P16	"Fortunately I didn't need to be reminded of talking because I didn't shut up." P16

Appendix L. Additional comments provided after experiencing think-aloud methods

Comments After Experiencing Speech Communication Method
"I liked this one better than the previous one and I felt that it was comfortable to think aloud, but it made me slower in completing the tasks." P4
"I had no problem thinking aloud, and it actually helped stay focused on what I was originally searching for." P5
"I'm bummed that I could not find the first answer." P6
"Very comfortable and easy process to breakdown what was being done on screen." P8
"I felt a little uncomfortable because I felt kind of foolish for not being able to find certain things right away or at all." P9
"If I were at home doing this chances are I would be talking out loud every now and then." P14

Comments After Experiencing Traditional Method
"Pretty interesting experience as I never really think out loud as I'm doing something." P3
"I liked the study, but I felt stupid when I couldn't find what I was looking for. I felt that I was taking more time because I was trying to say everything out loud while performing the tasks, but I enjoyed the study!" P4
"Thinking aloud while doing the tasks was easier than I thought it would be. A very hassle-free experience." P5
"It was easy to think when navigating, but not when trying to read down the lists of locations." P7