Journal of Usability Studies
Vol. 12, Issue 3, May 2017 pp. 95–110

# Eye Tracking in Retrospective Think-Aloud Usability Testing: Is There Added Value?

**Fatma Elbabour**
Assistant Lecturer
Faculty of Information Technology
University of Benghazi
Benghazi
Libya
fatma.elbabour@uob.edu.ly

**Obead Alhadreti**
Assistant Professor
College of Computer
Umm Al-Qura University
Al-Qunfoudhah
Saudi Arabia
oghadreti@uqu.edu.sa

**Pam Mayhew**
Senior Lecturer
School of Computing Sciences
University of East Anglia
Norwich
UK
p.mayhew@uea.ac.uk

## Abstract

Eye tracking is the process of recording users' eye movements while they are looking at the location of an object. In usability testing, this technique is commonly used in combination with think-aloud protocols. This paper presents an experimental study involving 24 participants; with the aim of comparing two variants of retrospective think-aloud (RTA) methods, that is, video-cued RTA method and gaze-cued RTA method, to address the value of having an extra eye-cue in retrospective think-aloud usability testing. Results suggest that both RTA variants are effective in detecting major usability problems. Moreover, the combination of eye tracking techniques and think-aloud protocols can further help evaluators to detect more usability problems, especially minor navigational and comprehension problems. It also helps participants to remember their behavior details, such as what they were looking at on a web page, as mouse movement alone might not be representative of their actual thoughts. Nevertheless, we found that participants might become distracted while seeing their eye movement, which can affect their verbalization performance and, hence, they might experience longer silence periods.

## Keywords

usability testing, think-aloud, retrospective think-aloud, eye tracking, verbalization

## Introduction

One of the most popular usability testing methods is known as the think-aloud protocol, which as the name implies refers to the act of verbalizing thoughts of a certain cognitive process. This method was first introduced in the field of psychology to study the cognitive processes in humans, and more recently it became widely used in the field of computing, particularly in usability testing, and it has been the dominant usability testing method for decades (Nielsen, 1993 and 2012). Think-aloud protocols have been traditionally classified into two types: concurrent think-aloud (CTA) and retrospective think-aloud (RTA). In the former, participants verbalize their thoughts while they are performing a task, whereas the latter refers to participants verbalizing their thoughts about a task they performed earlier (Boren & Ramey, 2000; Ericsson & Simon, 1980). The retrospective think-aloud method has an advantage over the concurrent think-aloud method as researchers found the latter to be more prone to *reactivity* (Russo, Johnson, & Stephens, 1989). In comparison with the RTA method, the reactivity in CTA usability testing conditions can have a negative effect on the task performance of participants (van den Haak, De Jong, & Schellens, 2003). However, since the initiation of these two think-aloud methods into usability testing there has been a growing interest in testing their validity and reliability (Cooke, 2010; Ramey et al., 2006), as well as comparing between the variants of these protocols (e.g., Goh et al., 2013; van den Haak et al., 2003).

A more recent trend in usability research is to combine think-aloud methods with eye tracking techniques. It has been proven that eye tracking can help usability researchers to understand how displayed information and visuals can affect the usability of a system (Bojko, 2006; Poole & Ball, 2005). Accordingly, many recent studies have focused on using eye tracking to overcome the drawbacks of think-aloud methods (Ball, Eger, Stevens, & Dodd, 2006; Cooke & Cuddihy, 2005). To date, there is little research on the validity of think-aloud protocols and the advantages of combining them with eye tracking techniques. For instance, some believe in the benefits of the extra eye-cue to the retrospective think-aloud method (Ball et al., 2006; Eger, Ball, Stevens, & Dodd, 2007; Hyrskykari, Ovaska, Majaranta, Rih, & Lehtinen, 2008; Olsen, Smolentzov, & Strandvall, 2010). Others argue that there is no added value to the retrospective think-aloud method when combined with eye tracking methods (Elling, Lentz, & De Jong, 2011), and the verbalizations of the traditional retrospective think-aloud method are valid and reliable when compared with captured eye-movement records (Guan, Lee, Cuddihy, & Ramey, 2006).

Our study aims to provide usability practitioners with a better understating of the impact of eye tracking techniques on the retrospective think-aloud protocol. This study contributes to theory by showing new understandings of the types and severity levels of usability problems detected by the gaze-cued RTA method. Moreover, the results of our study contribute in developing a better understanding of how the extra eye-cues affect the experience of participants, highlighting major topics such as distraction and silence periods.

## Related Work

Researchers are keen on combining eye tracking techniques with think-aloud protocol. Some studies used eye tracking technology to address the limitations and the validity of think-aloud methods (Cooke, 2010; Cooke & Cuddihy, 2005; Elling, Lentz, & De Jong, 2012; Guan et al., 2006). Others were interested in comparing gaze-cued think-aloud methods with the traditional methods in different test settings (Eger et al., 2007; Elling et al., 2011; Olsen et al., 2010; Hyrskykari et al., 2008).

### *Addressing the Validity of Think-Aloud Methods*

Cooke and Cuddihy (2005) investigated the validity of the verbalizations of participants who participated in a concurrent think-aloud website usability study. They found that the extra eye-cue data can reveal participants' behavior which is not verbalized during the think-aloud session. In a different study, Cooke (2010) also investigated the validly of the concurrent think-aloud method by combining it with eye tracking techniques. His work suggests that the value of the think-aloud verbalizations compared with what can be observed from eye movement records alone is limited. He argued that most of the information obtained from the participant's verbalizations can be easily obtained from their eye movements alone.

Elling et al. (2012) replicated the work of Cooke (2010) but, contrary to Cooke, they concluded that eye tracking techniques can benefit concurrent think-aloud usability testing methods by having a better understanding of what users are doing, especially when being silent. They reported different verbalization types other than the ones mentioned by Cooke (2010) as well as a higher percentage of silence periods during which interesting observations were made from the recorded eye movements.

Guan et al. (2006) applied eye-tracking techniques to test the validity and the reliability of the retrospective think-aloud method, in their experiment, they compared the verbalizations of the participants with a captured record of their eye movements during the session. They concluded that the retrospective think-aloud method provided valid verbalizations of the actual actions and behaviors of the participants during the test, and there was a low risk of fabrication during the verbalizations.

### *Comparing Between Think-Aloud Variants*

Hyrskykari et al. (2008) compared the traditional concurrent think-aloud method with the gaze-cued retrospective think-aloud method. Their findings yielded more informative verbal data in the gaze-cued retrospective think-aloud session compared with the results of the concurrent think-aloud sessions. Eger et al. (2007) also compared three variants of the think-aloud protocols (i.e., CTA, RTA, and RTA with eye tracking), and they concluded that gaze-cued retrospective think-aloud method detected more usability problems than the other methods. They also found that the detection of some types of problems were associated with having the additional eye tracking data.

However, Elling et al. (2011) who also compared a video-cued retrospective think-aloud method to the gaze-cued retrospective think-aloud method, concluded that there was no difference between the two think-aloud variants in terms of the number of discovered usability problems, the type of problems, and the way the problems were detected. Olsen et al. (2010) compared four types of retrospective think-aloud methods (i.e., no cue RTA, video cued RTA, gaze-plot cued RTA [eye movement on still image], and gaze-cued RTA), and found that the gaze-cued retrospective think-aloud method was more effective than the other variants. They highlighted that it helped participants to verbalize almost double the amount of words, while the number of the detected usability problems were nearly the same between the video-cued and the gaze-cued RTA method.

The previous studies were conducted in different settings in terms of the number and the profile of participants and the followed think-aloud protocol (i.e., strict or relaxed protocol). The latter is an important factor in usability studies; the work of Olmsted-Hawala, Murphy, Hawala, and Ashnefelter (2010) highlighted the importance of reporting the kind of probing used in a usability experiment. They argued that adopting a relaxed approach in a usability session—described as coaching, which is a think-aloud method where the evaluator asks questions about different areas of the website—can improve participants' performance and satisfaction.

### *The Present Study*

The focus of our study is to assess the impact of combining eye-tracking techniques with the traditional RTA method in website usability testing. We considered the differences between the results of the detected usability problems in the RTA variants. Furthermore, we considered the differences between participants' verbalizations, including measures such as the length of the verbalizations, the word count, silence period, and the types of the comments. We also considered the opinions of the participants, taking advantage of the within-subject design of the study whereas most similar studies followed a between-subject approach. Accordingly, we investigated the following research questions:

- To what extent do usability problems detected by each RTA method differ in terms of numbers, types of problems, levels of severity, and ways of detection?
- To what extent do verbalizations produced by each RTA method differ in terms of verbalization time, word count, silence periods, and types of comments?
- To what extent do participants' opinions about each RTA method differ in terms of ease of verbalization, remembering thoughts, enjoyment of seeing recordings, and method of preference?

The rest of the paper is structured as follows: A description of research methods, followed by data analysis, and finally a discussion of the main findings, recommendations, and conclusions.

## Methods

The following sections present the methods used in this study for designing the experiment and choosing the test objects and tasks. Participant data, equipment specifications, experimental procedures, and data analysis information are also included in this section.

### *Experimental Design*

To achieve the objectives of our study and in order to compare the results of the two variants of retrospective thinking aloud usability testing, we conducted a 2*2 mixed factorial design. The RTA method was either video-cued RTA or gaze-cued RTA while the website was either Website 1 or Website 2. There were four conditions labelled according to the RTA method type and the selected website. To have a within-subject design, we classified participants into two groups. Each group experienced the two types of RTA methods with a different website each time, amounting to four tasks per participant. The order of displaying the websites and tasks was randomized in each group of participants.

### *Test Objects and Tasks*

Two charity related websites were used as test objects in this study. The first website belongs to the British Heart Foundation (BHF) charity (www.bhf.org.uk), which provides heart health information to the public and offers other online services such as a gift shop and a donation facility. The name of the second website will not be disclosed in this report upon the request of the website owners, and it will be referred to as the "CCU website." This website offers services and information about charities and charity events in the UK.

We designed one scenario and two tasks for each website. The BHF website scenario read, "You are thinking of buying something from a charity shop. A friend has recommended the British Heart Foundation shop, and you have decided to visit their website." Two tasks followed: (1) "Buy two or more items for yourself and/or friends. Spend less than 15 GBP"[1] and (2) "Find the address of the nearest British Heart Foundation clothing shop to you." The scenario for the CCU website was, "You are willing to participate in a charity related activity, but you have not decided which charity to choose. A friend has recommended the CCU website, which provides information about charities in the UK. You have decided to visit this website." This scenario was also followed by two tasks: (1) "Find the contact details of a charity of your choice" and (2) "Find an upcoming event that interests you that you would like to attend."

### *Participants*

In total, the data of 24 participants were included in the analysis stage of this study, excluding others with missing eye tracking data samples. The majority of the 24 participants were women (75%), while men represented only 25% of the sample. The age distribution was as follows: 45.83% of the participants were between 25 and 40 years old, 37.5% were between 41 and 64 years old, and only 16.66% of the participants were aged between 18 and 24 years old. All participants were native English speakers and residents of the UK. In terms of education levels, 70.83% of the participants were educated to higher education levels (i.e., Bachelors, Masters, and PhDs) in various disciplines, and the educational levels of the rest ranged between secondary and vocational education. All the participants had been involved in at least one charity related event or activity, and most of them indicated that they had visited a charity related website at least once; however, they were not familiar with the two selected websites used in this study. Only two participants had participated in usability studies before, while none of the 24 participants had ever participated in think-aloud experiments or similar eye tracking studies in which participants see their eye movements. Participants were allocated equally between the design groups according to their characteristics: age, gender, and level of education.

[1] The target was to find the items, add them to the shopping cart, and reach the checkout page, but no actual buying was involved.

### *Questionnaires*

Participants' demographic information (i.e., age, gender, educational level, nationality, and first language) and relative experience details (i.e., internet usage, recent online activities, participation in similar studies, and charity experience) were collected as part of the screener questionnaire. Details about their charity interests were collected on the day of the experiment using an interest questionnaire. We used a post-test questionnaire, a 5-point Likert scale with 5 as the most positive score, to evaluate the participants' opinions on each variant of the retrospective think-aloud tests (adopted from Elling et al., 2011). The participants evaluated the following questionnaire statements: (a) I found it easy to verbalize my thoughts, (b) Seeing the playback video supported me in remembering what I thought, and (c) I liked seeing the playback video while verbalizing my thoughts.

### *Equipment*

The experiment was conducted in an eye tracking laboratory, a small air conditioned room, located at the University of East Anglia (UEA). The adopted eye tracking system was a Tobii TX300 Eye Tracker that runs on 120Hz with Tobii Studio 3.0.3 software. We used the software to capture and display video recordings, and due to the limited features of this version of the software, audio data was recorded using a digital audio recorder. The web browser used in the study was Internet Explorer version 11 installed on a 64-bit (RAM 8 GB) PC with a Windows 7 Professional operating system.

### *Experimental Procedure*[2]

On the experiment day, participants were first welcomed and thanked for their interest in the study. They read and signed a consent form, and filled out an interest questionnaire to collect information about their charity interests. To warm-up and minimize stress, the evaluator introduced the participants to the lab and had an informal conversation with them. Subsequently, the evaluator instructed them to look at the screen in order to start the calibration process and to adjust their seating if deemed necessary. The calibration process was repeated in case it did not work properly the first time.

After that, the evaluator instructed the participants, using a session script, to perform the two sets of tasks on the two websites, and encouraged them to ask questions before starting to perform the tasks.

While they were performing the tasks, the evaluator was observing and taking notes of their task performance using the evaluator's computer screen. If participants forgot the task, they asked the evaluator to read it for them. Once they completed the first task, participants notified the evaluator in order to display the second task. Again, once the participant completed the second task, the evaluator ended the first part of the test and instructed the participant to follow a similar procedure with the second website.

After a short break, the evaluator read the instructions for the RTA session and allowed the participants to ask questions, if any. In the case of the eye movement playback video, a brief explanation and demonstration of the gaze paths and fixations were given. After that, the evaluator asked the participants to verbalize their thoughts once, while they were watching a playback of their eye movement, and then once again while they were watching their task performance only without eye movement[3]. If needed, they were allowed to pause the video in both conditions and increase the speed of the eye movement[4]. Meanwhile, the evaluator was recording the session using a digital recorder as well as taking notes. Also, the evaluator was listening carefully to the participants to communicate with them using acknowledgement tokens, such as "aha," "yeah," "I see," and "ok," as well as reminding participants to think aloud, if needed, using phrases such as "keep talking please" and "please tell me what were you thinking of at this stage."

[2] Ethics approval was obtained from UEA.

[3] Each type of playback video was displayed according to the actual order of the websites on the previous phase in which participants performed their tasks. It also depended on the group a participant belonged to, whether the eye movements were recorded for Website 1 or Website 2.

[4] Eye movement playback videos were first displayed at the slowest pace (speed 1). The playback speed bar of Tobii studio has 7 paces, where pace number 1 is the slowest, pace number 4 is equal to real time speed, and the speed number 7 is the fastest.

Following each playback video, the evaluator asked the participants to fill out the post-test questionnaire. At the end of the session, the evaluator asked the participants to answer one final question regarding their method of preference.

### *Data Analysis*

Research variables were composed of usability problems, RTA verbalizations, and participants' experiences.

*Usability problems* were measured quantitatively to determine the number of discovered problems, their types, levels of severity, and their method of discovery. In the case of observation, usability problems were either identified from the evaluator's notes during the session or during the analysis of the video recordings. The latter was based on the Target Search Analysis Framework, which is a four-step framework used in formative user experience (UX) research (Bojko, 2013). Although, this framework was meant for observing videos with eye movement data, it was applicable for both observation conditions in this study. Figure 1 shows the adopted framework with modifications so as to be suitable for both conditions.

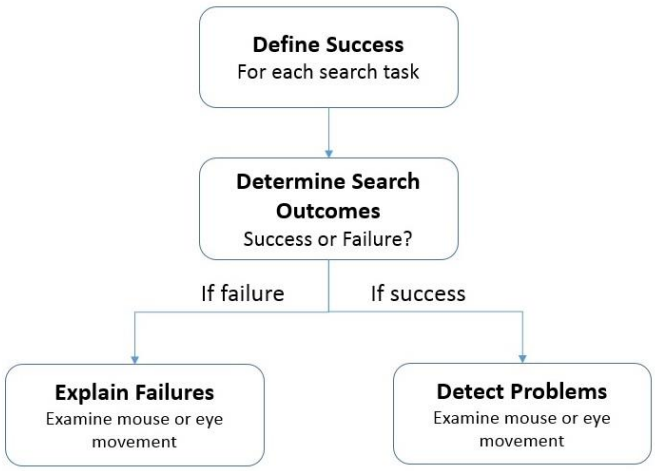



**Figure 1.** Usability problems detection method adapted from the Target Search Analysis Framework. Adapted from *Eye Tracking the User Experience: A Practical Guide to Research*, by A. Bojko, 2013, Brooklyn, NY: Rosenfeld Media. Copyright 2013 by Aga Bojko. Adapted with permission from Creative Commons license.

*RTA verbalizations* were documented, including notes of the start and end time of the verbalization session, and any prompts from the evaluator were noted along with the time. The frequency of silence periods longer than five seconds for each verbalization was calculated using open source software called Audacity, excluding the time when the evaluator spoke.

We tokenized transcriptions into usability problems and the following operational comments (Hansen, 1991): manipulative, visual, and cognitive. We classified the types of the problem according to a usability problems scheme used in previous research (Eger et al., 2007; Olsen et al., 2010): layout, navigation, feedback, comprehension, terminology, and data entry. Usability problems discovered through either observation and verbalization analysis stages were classified according to three categories: observation only, verbalization only, or both (Elling et al., 2011; van de Haak et al, 2003). Detected usability problems were classified into four levels of severity, adopted from Dumas and Redish, (1999): Level 1—prevents task completion, Level 2—creates significant delay and frustration, Level 3—problems have a minor effect on usability, and Level 4—subtle and possible enhancements or suggestions.

*Participants' experience* analysis: For quantitative data, a 5-point Likert scale analysis was performed by using the post-test questionnaire to compare between the responses of the participants in each RTA condition. For qualitative data, we followed a thematic coding approach (Robson, 2011) to analyze participants written comments.

## Results

We conducted a preliminary analysis to measure task performance (i.e., task completion time and task success rate) using a paired samples *t*-test to ensure that the comparison between the two RTA conditions was possible. For variables that follow Poisson distribution (i.e., number of usability problems, methods of detection, types of usability problems, levels of severity, frequency of silence periods, and operational comments), we used generalized estimating equations (GEE) models to analyze the significance. For normally distributed variables (i.e., verbalization time and number of words produced per minute), we used two-way mixed ANOVA tests. Due to the ordinal nature of the Likert scale data, we used GEE analysis with the ordinal logistic model to analyze the participants' responses to the post-test questionnaire.

### *Usability Problems*

The following sections present the usability problems we discovered during our study. The usability problems were measured by the following means: the total and unique number of problems revealed by each RTA condition, the levels of severity, the ways of detection, and the types of the revealed usability problems.

#### *Number of problems*

In total, 78 problems were found in the BHF website, 28% of the problems were commonly discovered in both RTA conditions, while the percentages of unique problems discovered in the video-cued RTA condition and the gaze-cued RTA condition were 27% and 45%, respectively. For the CCU website, 69 problems were detected, of which, 26% were commonly discovered in both RTA conditions, while the percentages of unique problems discovered in the video-cued RTA condition and the gaze-cued RTA condition were 25% and 49%, respectively. The difference between the total number of usability problems identified by each of the two methods was statistically significant ($p$-value = 0.002 ($p < 0.01$)).

#### *Levels of severity*

Of the usability problems, 62.5% (n = 92) were rated as minor and subtle (Level 3 and Level 4), of which 52% (n = 48) were identified only by the gaze-cued RTA method and 24% (n = 22) only by the video-cued RTA method. From the 55 critical and major (Level 1 and Level 2) usability problems, the number of problems of which 38% (n = 21) were identified by the gaze-cued RTA method and 29% (n = 16) only by the video-cued RTA method. Our results indicated a significant difference between minor usability problems ($p$-value = 0.004) and subtle and possible enhancements or suggestions ($p$-value = 0.003).

#### *Method of detection*

Analysis of the method of detection showed that the amount of usability problems identified by the gaze-cued RTA method through verbalization only (n = 88) was higher than that identified by the video-cued RTA method (n = 47) with a significant difference of $p$-value = 0.002. Although the total number of problems detected through observation by the gaze-cued RTA method (n = 77) was slightly higher than those identified by the video-cued RTA method (n = 59), we found no significant difference between problems detected through observation only ($p$-value = 0.088). Table 1 presents the average number of usability problems identified by the two RTA methods, both at their severity level and the method of detection.

**Table 1.** Average Usability Problems Identified by the Video-Cued RTA and the Gaze-Cued RTA Methods in Terms of Numbers, Severity Levels, and Method of Detection

| | **Video-cued RTA** | | **Gaze-cued RTA** | |
|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** |
| ***Usability problems**** | 5 | 3.1 | 7.4 | 3.6 |
| ***Severity level*** | | | | |
| Level 1 (*Critical problems*) | 0.83 | 0.91 | 1.17 | 1.63 |
| Level 2 (*Major problems*) | 1.5 | 1.44 | 2.04 | 1.73 |
| Level 3 (*Minor problems*)* | 1.29 | 1.12 | 2.25 | 1.25 |
| Level 4 (*Suggestions*)* | 1.16 | 0.91 | 1.92 | 1.31 |
| *Average severity of total problems* | *2.56* | *1.05* | *2.79* | *1.06* |
| ***Method of detection*** | | | | |
| Observed only | 2.45 | 2.1 | 3.2 | 1.7 |
| Verbalized only* | 1.95 | 1.6 | 3.66 | 2.2 |
| Combination of both methods | 0.58 | 0.8 | 0.54 | 1.1 |

* $p < 0.01$

*Types of usability problems*

Table 2 shows the average number of usability problems detected by the two RTA methods categorized by their types. For navigation and comprehension problems, we found a significant difference between the two RTA conditions: *p*-value = 0.010 and *p*-value = 0.41, respectively. Although the number of layout problems detected by the gaze-cued RTA method was slightly higher than the ones detected by the video-cued RTA method, we found no significant difference between the two conditions (*p*-value = 0.165). We also found no significant difference between terminology problems (*p*-value = 0.933) and feedback problems (*p*-value = 0.116). No data entry problems were detected in this study.

**Table 2**. Types of Usability Problems

| **Type of problem** | **Video-cued RTA** | | **Gaze-cued RTA** | |
|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** |
| Layout problems<br>(Difficulty to spot an item on a webpage) | 2.7 | 2.03 | 3.38 | 1.86 |
| Navigation problems*<br>(Difficulty to navigate around the website | 0.67 | 0.87 | 1.42 | 1.28 |
| Terminology problems<br>(Difficulty to understand a terminology) | 0.17 | 0.38 | 0.17 | 0.38 |
| Feedback problems<br>(Unexpected feedback from the website) | 0.66 | 1.09 | 1.25 | 1.25 |
| Comprehension problems*<br>(Difficulty to understand instructions) | 0.58 | 0.8 | 1.21 | 1.4 |

* $p < 0.05$

Figures 2 and 3 represent the numbers of unique usability problems detected by each method categorized in terms of problem types and its severity level.

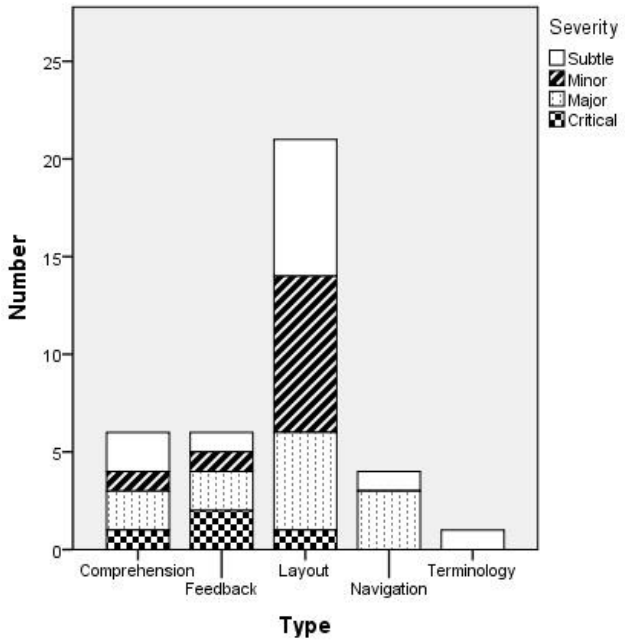



**Figure 2.** Usability problems detected by video-cued RTA method only.

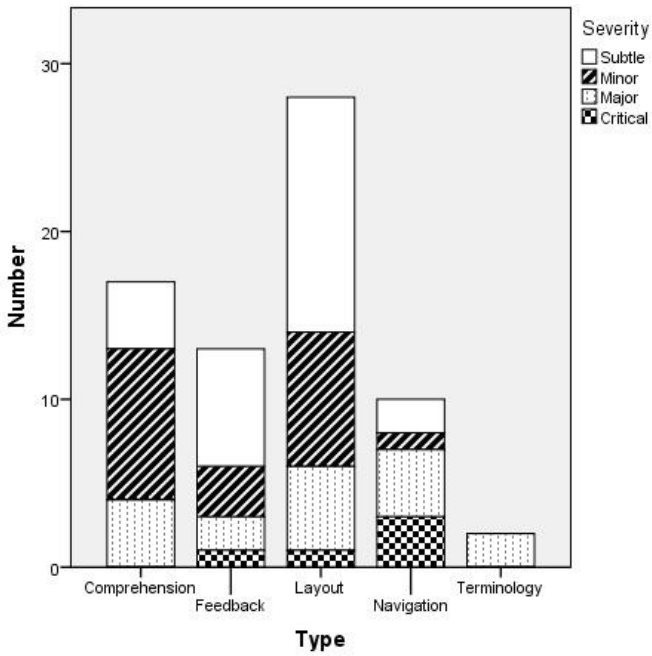



**Figure 3.** Usability problems detected by gaze-cued RTA method only.

### *RTA Verbalization*

The following sections present the results for the participants' verbalizations, which were analyzed against verbalization time, frequency of silence periods, word count, and types of operational comments.

*Verbalization time*

On average, the length of participants' verbalizations in the video-cued RTA condition was 195 seconds (i.e., 00:3:15 minutes) while the verbalizations in the gaze-cued RTA condition required an average of 353 seconds (i.e., 00:5:53 minutes). The two-way mixed ANOVA test shows that there is a significant difference between the scores in the two RTA conditions, $F(1,22) = 21.49$, $p$-value = 0.000 (< 0.05).

*Frequency of silence periods*

Our results suggest that participants experienced longer silence periods in the gaze-cued RTA condition (Mean = 4.63, SD = 3.9) than in the video-cued RTA condition (Mean = 1.46, SD = 1.7). Accordingly, we found a significant difference between the number of silence periods longer than five seconds in the two RTA conditions, $p$-value = 0.00 (< 0.05). Although participants experienced longer silence periods in the gaze-cued RTA condition, we found no significant difference between the average number of prompts required per minute to remind participants to think aloud in both RTA conditions, $p$-value = 0.14.

*Number of words*

On average, the number of produced words in the video-cued RTA condition was 132 words per minute, while the number of words produced in the gaze-cued RTA condition averaged 115 words per minute. The results of the two-way mixed ANOVA test yielded a significant difference between the number of words produced per minute in the two RTA conditions, $F(1,22) = 5,121$, $p$-value = 0.034 (< 0.05).

*Operational comments*

In general, our results suggest that the percentage of visual comments increased from 25% in the video-cued RTA condition to 30% in the gaze-cued RTA condition. This increase was associated with a slight decrease in the manipulative comments from 35% to 30%, respectively, while the percentage of the cognitive comments was similar (40% and 39%). We found there was significant differences in the scores of manipulative comments ($p$-value = 0.004), visual comments ($p$-value = 0.000), and cognitive comments ($p$-value = 0.000). Table 3 summarizes these results.

**Table 3.** Operational Comments

| **Operational comment** | **Video-cued RTA** | | **Gaze-cued RTA** | |
|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** |
| Manipulative comments*<br>(Describing a behavior, e.g., "I clicked on the charity link.") | 9.75 | 5.96 | 13.21 | 8.51 |
| Visual comments*<br>(Commenting upon their perception, e.g., "I had a little look around to see what is for sale.") | 6.75 | 3.77 | 13.75 | 7.02 |
| Cognitive comments*<br>(Interpretations, evaluations, expectations, or specification of an action, e.g., "Then I realized there was a calendar at the bottom.") | 11.08 | 6.02 | 17.16 | 7.72 |

***** $p < 0.01$

### *Participants' Experience*

The overall experience for participants indicated no significant differences between how participants rated their experience in both RTA conditions. Table 4 presents the average scores on a Likert scale. The higher the score, the more satisfied the participant.

However, answers to the last question regarding participants' preferred method of verbalization revealed different results, where 11 participants reported that they preferred the gaze-cued RTA method, nine participants preferred the video-cued RTA method. Only four participants said they equally preferred both methods. Table 5 summarizes participants' comments regarding their preferred condition.

**Table 4.** Results of Post-Test Questionnaire

| **Statement** | **Video-cued RTA** | | **Gaze-cued RTA** | |
|---|---|---|---|---|
| | **Mean** | **SD** | **Mean** | **SD** |
| Ease of verbalization | 4.17 | 0.56 | 4 | 0.83 |
| Supported remembering thoughts | 4.5 | 0.51 | 4.41 | 0.83 |
| Liked seeing the recordings | 4.38 | 0.8 | 4.29 | 0.9 |

**Table 5**. Participants' Comments

| **Comments on** | **Video-cued RTA Comment (Count[a])** | **Gaze-cued RTA Comment (Count[a])** |
|---|---|---|
| Seeing the recordings | Straightforward (3), Helpful (3) | Interesting (3), Distracting (2) |
| Verbalizing thoughts | Easy (1), Difficult (2) | Easy (3), Difficult (3) |
| The entire experience | Uncomfortable (1), Challenging (1) | Uncomfortable (1), Challenging (2) |

[a] Number of times a comment was mentioned by the participants.

The thematic analysis of why the participants preferred one method over the other identified a series of themes. We represent the most noteworthy ones below along with some participant quotations.

*Source of distraction*

All participants who preferred the video-cued RTA method commented that the traditional RTA method was less distracting than seeing their eye movement. Seeing the eye movement and verbalizing thoughts at the same time seemed to be a challenge to many participants. One participant commented that she "can concentrate on the thoughts about my actions better without knowing what my eye movements were." Another participant found it difficult to remember her thoughts while seeing the eye movement, "I found it very difficult to remember my thought process if I was concentrating on the eye movement. I would like to have seen the eye movement without having to think too."

Furthermore, others commented that seeing the eye movement made them think more about the technology itself rather than focusing on verbalizing their thoughts, for example, "Seeing the eye track made me think more about how the tracking worked than what I was thinking when I did the tasks." She also added that she was focused more on understanding her eye movement, "I kept wondering why my eyes were moving round so much." Another participant added that "I found it hard to verbalize what my eye movement was showing when I did not realize I was looking at something." The extra eye cue in RTA testing seems to be a source of distraction to many participants; however, we should take into account that none of the participants in this study were familiar with either the think-aloud protocol or eye tracking techniques. Therefore, it was clear that participants had made effort to verbalize their thoughts in general and this was a bit challenging when combined with the eye tracking data.

*Remembering details*

Most of the participants who preferred the gaze-cued RTA method found it easier to recall their thoughts in more detail compared to the other variant. When seeing the eye movement, participants were reminded of where they were looking and what they were looking for, which helped them to recall their actual thoughts in the form of steps. The following quotations are two examples of their comments: "Having the eye movement was a good reminder of what I was thinking, without it I ended by skipping ahead," and "To me, the eye movement reminded me of my focus at any particular point—this helped me recall my thoughts more accurately."

Another participant added that the eye movement reminded her of what she considered doing but did not: "The red dots and lines reminded me of more details of what I'd done and what I'd considered doing but didn't." Moreover, basic understanding of the eye tracking data helped some participants to remember more details. For example, "Liked it when red circles got bigger because it reminded me I had spent time there thinking," and "it helped to recall where I was looking and the process/steps/journey on the page and could remember why I did what I did. Also, seeing the areas, I looked at once or for a longer time helped to remember the important steps I took."

## Discussion

The following sections provide discussion on the usability problems we found, the RTA verbalizations, and the participants' experiences.

### *Usability Problems*

In agreement with the findings of Eger et al. (2007), the results of our study suggest that the gaze-cued RTA method can detect more problems than the video-cued RTA method. On the other hand, Elling et al. (2011) concluded that the total number of problems detected in both RTA conditions was not significantly different, and they argued that their results were different from the results of Eger et al. (2007) possibly because Elling et al. (2011) used more strict prompts and did not ask questions during verbalizations. However, in this study intervention was limited to "keep talking please" and "please tell me what were you thinking of at this stage," and no specific questions were asked.

Similar to the conclusion of Elling et al. (2011), both methods have been effective in detecting layout and design problems. However, in our study, the gaze-cued method had the advantage of detecting more navigation problems; a possible reason for this conclusion is that the eye movement helped participants to remember what they were looking for while performing a certain task. For example, one participant said after observing the steps of his eye movement, "Ah! Yeah, now I'm trying to find how to get back to shop," and the same participant in the video-cued RTA condition stayed silent for a while until he realized what he was doing, "I'm looking for [silence] yeah I found the find events [link]." The extra eye-cue also helped the evaluator to understand what the participant was looking for on the web page while the mouse cursor is fixed at one point of the page. Furthermore, the gaze-cued method detected more comprehension problems; we believe this is due to the large size of circles/dots (i.e., fixation duration), as well as the pattern of fixations at a particular area, which reminded the participants that they had spent a longer time looking at something.

In the case of the video-cued RTA method, more problems were discovered through observation only compared to the number of problems revealed through verbalizations. In contrast, in the gaze-cued RTA condition, the number of problems discovered through verbalizations only was slightly higher than the ones detected through observations. These results are similar to the findings of Elling et al. (2011). Measuring the severity levels of the detected usability problems suggested that both methods were effective in detecting critical and major problems, however, the extra eye-cue seems to be helpful in discovering minor and less critical ones. More research is needed to understand the relationship between problem severity and the extra eye-cue in think-aloud methods, perhaps by conducting a quantitative analysis of the eye tracking data.

### *RTA Verbalizations*

As expected, verbalization time in the gaze-cued RTA method was longer than the video-cued RTA method. This is due to the fact that the playback videos of the gaze-cued RTA method were slower than the video-cued RTA recordings, as it is recommended to see the playback video of

the eye movement at slow speed rather than a real-time speed. Despite the length of the gaze-cued RTA conditions being longer than the traditional RTA condition, the number of words produced per minute in the latter condition was more than the words produced per minute in the gaze-cued RTA condition. These results are different from the conclusions of Olsen et al. (2010), perhaps in their study there was no indication of the time factor when calculating the total number of words and also English, the language of the test, was not the first language of the participants in their study.

Our findings suggest that participants experienced more frequent silence periods (> 5 minutes) in the gaze-cued RTA condition, however, this did not impact the number of prompts required per minute in each RTA condition. Thus, more research is needed to understand why participants experience more silence in the gaze-cued RTA condition and whether this would change if participants receive proper training on eye tracking technology prior to usability tests.

In line with the primary findings of Hansen (1991), gaze-cued RTA verbalizations contained a higher percentage of *visual comments* compared with the video-cued RTA verbalizations. This increase in the visual comments was associated with a slight decrease in *manipulative comments*, which were higher in the video-cued RTA condition. There was no change in the numbers of *cognitive comment*s in both RTA conditions. Visual comments were linked to the discovery of more usability problems in comparison with manipulative comments, because participants were commenting upon their perception, such as "I had a little look around to see what is for sale," rather than describing behaviors that were evident from the video, such as, "So then I clicked on the charity link."

### *Participants' Experiences*

Participants expressed both positive and negative comments about the two RTA methods. Participants who preferred the video-cued RTA method found it more straightforward and less distracting than the other method, while others preferred seeing their eye movement because it helped them to remember details in addition to finding it more interesting than the video-cued RTA method. Even though the playback speed of the gaze-cued video was slow and the evaluator allowed the participants to pause the playback video when needed, participants in this experience seemed to be more distracted, sometimes confused, and excited than with the video-cued RTA method. Furthermore, some participants seemed to be keen to understand their movement rather than focusing on verbalizing their thoughts. However, as emphasized earlier, participants in our study were neither familiar with think-aloud protocol nor eye-tracking techniques.

## Limitations

This study has several limitations that should be considered when interpreting the results. First, due to time constraints, we used only two tasks per charity related website. Because our study examined a limited number of tasks, we could draw no conclusions about their effect on the results. Having more tasks, including complex ones, and testing other types of websites, could be used to compare between the two RTA conditions. Second, although all participants showed an interest in charities, this does not mean that these participants are actually representative users of both websites. Third, we did not analyze differences by participant characteristics, such as age, gender, or education level. The majority of the participants in this study were women older than 25 years old; having a better distribution of age and gender would be an advantage. Fourth, the evaluator's cultural background was different from that of the participants', which might have some influence on the results. Fifth, we designed the RTA sessions according to the actual practice of usability practitioners and the recommendations of previous research, and allowed participants to pause or change the speed of the recordings when needed. Thus, it would be useful to test whether the verbalizations might have been affected by the speed of the playback videos. Finally, none of the participants had any previous experience with think-aloud protocols, which might have affected their perception of watching their eye movement while verbalizing their thoughts. It would be helpful to invest more time in training participants for verbalizing their thoughts while watching their eye movement prior to the session day.

## Future Work

For future research, it would be useful to consider collecting quantitative eye tracking data (e.g., numbers of fixations and fixation duration) that can help in further evaluating the value of gaze-cued retrospective think-aloud methods. However, to date there is no standard scheme to interpret such eye tracking data in usability testing, which is another interesting area that researchers should study thoroughly to develop a standard schemer for interpreting eye tracking data in usability studies. Further research should also focus on measuring the influence of the type of the website (e.g., service and complexity) on the testing methods as well as the types of the designed tasks. Finally, similar research should consider studying the impact of having participants who are familiar with eye tracking data on the results.

## Conclusion

In this study, we carried out a within-subject experiment to explore the effect of the extra eye-cue on the traditional retrospective think-aloud usability testing method. We identified the effect on the detection of usability problems, participants' RTA verbalizations, and participants' experience. Our study shows that the gaze-cued RTA method can detect more usability problems compared to the video-cued RTA method. However, the gaze-cued RTA method is associated with longer sessions, in which participants may experience longer silence periods, and seeing the eye movement proved be a source of distraction to many participants who are not familiar with either eye tracking data or think-aloud methods.

This study particularly shows that both methods are effective in detecting critical and major problems, but the extra eye-cued information encouraged participants to verbalize more visual comments than manipulative comments, which consequentiality helped the evaluator to discover more minor and less critical problems. Both methods are effective in detecting layout and design problems, in addition, the gaze-cued RTA method detected more navigation and comprehension problems. These findings are helpful in deciding the choice of method for usability testing, considering time limits and type and severity levels of problems that the usability practitioner is interested in.

## Tips for Usability Practitioners

The following are tips for usability practitioners on what they should consider when deciding whether to use the traditional RTA method or combining eye tracking techniques with the RTA method.

- Apply the traditional RTA method if testing time is limited and critical. This is because the traditional RTA method is effective in detecting potential usability problems in a shorter time compared to the gaze-cued RTA method.
- Apply the gaze-cued RTA method if you are interested in a thorough understanding of how the users interact with your website as well as being interested in detecting minor usability problems (e.g., design and layout issues).
- Consider recruiting participants who are familiar with both think-aloud protocols and eye tracking data, or alternatively invest more time in training them, when using the gaze-cued RTA method. This can possibly reduce the reported distraction factor, and accordingly you may benefit from the advantages of this method.

## Acknowledgements

Thanks to the members of the School of Computer Sciences and the School of Psychology at the University of East Anglia for their support.

## References


Ball, L. J., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the peep method in usability testing. *Interfaces,* 67, 15–19.

Boren, T., & Ramey, C. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261–278.

Bojko, A. (2006). Using eye tracking to compare web page designs: A case study. *Journal of Usability Studies, 1*(3), 112–120.

Bojko, A. (2013). *Eye tracking the user experience: A practical guide to research*. Brooklyn, NY: Rosenfeld Media.

Cooke, L., & Cuddihy, E. (2005, July). Using eye tracking to address limitations in think-aloud protocol. *Proceedings of International Professional Communication Conference* (pp. 653–658). IEEE.

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.

Dumas, J., & Redish, J. (1999). *A Practical Guide to Usability Testing*. Intellect Ltd.

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007, September). Cueing retrospective verbal reports in usability testing through eye-movement replay. *Proceedings of the 21st British HCI Group Annual Conference on People and Computers* (pp. 129–137). British Computer Society.

Elling, S., Lentz, L., & De Jong, M. (2011, May). Retrospective think-aloud method: using eye movements as an extra cue for participants' verbalizations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1161–1170). ACM.

Elling, S., Lentz, L., & De Jong, M. (2012). Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions on Professional Communication, 55*(3), 206–220.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review, 87*(3), 215–251.

Goh, K. N., Chen, Y. Y., Lai, F. W., Daud, S. C., Sivaji, A., & Soo, S. T. (2013). A comparison of usability testing methods for an e-commerce website: A case study on a Malaysia online gift shop. *Proceedings of Information Technology: New Generations (ITNG), 2013 Tenth International Conference* (pp. 143–150). IEEE.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 1253–1262). ACM.

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica 76*(1), 31–49.

Hyrskykari, A., Ovaska, S., Majaranta, P., Rih, K. J., & Lehtinen, M. (2008). Gaze path stimulation in retrospective think-aloud. *Journal of Eye Movement Research 2*(4), 1–18.

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

Nielsen, J. (2012). Thinking aloud: The #1 usability tool. Retrieved February 2015 from http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

Olmsted-Hawala, E., Murphy, E., Hawala, S., & Ashnefelter, K. (2010, April). Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2381–2390). ACM

Olsen, A., Smolentzov, L., & Strandvall, T. (2010, September). Comparing different eye tracking cues when using the retrospective think aloud method in usability testing. *Proceedings of*

*the 24th BCS Interaction Specialist Group Conference* (pp. 45–53). British Computer Society.

Poole, A., & Ball, L. J. (2005). Eye tracking in human-computer interaction and usability research current status and future. In C. Ghaoui (Ed.), *Encyclopedia of Human-Computer Interaction* (pp. 45–53). British Computer Society.

Ramey, J., Boren, T., Cuddihy, E., Dumas, J., Guan, Z., van den Haak, M. J., & De Jong, M. D. (2006, April). Does think aloud work?: How do we know? *Proceedings of CHI'06 Extended Abstracts on Human Factors in Computing Systems* (pp. 45–48). ACM.

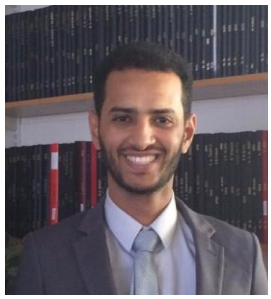
Robson, C. (2011). *Real World Research*. Wiley.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989), The validity of verbal protocols. Memory & Cognition. *Memory & cognition,* 17, 759–769.

van den Haak, M., De Jong, M., & Schellens, J. P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & information technology, 22*(5), 339–351.

## About the Authors

**Fatma Elbabour**
Ms. Elbabour is an assistant lecturer at the Faculty of Information Technology, University of Benghazi, Libya. She holds a Master degree in Information Systems from the School of Computer Sciences, University of East Anglia in Norwich, UK. Her current research interests focus on human computer interaction and usability.



**Obead Alhadreti**
Dr. Alhadreti is an assistant professor at the College of Computer, Umm Al-Qura University, Saudi Arabia. He has been involved in usability testing since 2009. His doctoral research focuses on the use of the think-aloud methods within usability testing. His interests involve usability evaluation, cultural usability, and user experience.

**Pam Mayhew**
Dr. Mayhew is a senior lecturer in the School of Computer Sciences at the University of East Anglia in Norwich, UK. She is interested in factors that affect the success of systems in general, and in usability issues and testing in particular. She supervises a number of PhD students in the area of usability.