# Comparing different eye tracking cues when using theretrospective think aloud method in usability testing

Anneli Olsen
+46 709 161695
*anneli.olsen@tobii.com*

Linnea Smolentzov
Tobii Technology AB, S-182 17 Danderyd, Box 743
*linneas@clemson.edu*

Tommy Strandvall
+46 8 5229 5132
*tommy.strandvall@tobii.com*

**Research has shown that incorporating eye tracking in usability research can provide certain benefits compared with traditional usability testing. There are various methodologies available when conducting research using eye trackers. This paper presents the results of a study aimed to compare the outcomes from four different retrospective think aloud (RTA) methods in a webusability study: an un-cued RTA, a video cued RTA, a gaze plot cued RTA, and a gaze video cued RTA. Results indicate that using any kind of cue produces more words, comments and allowsparticipants to identify more usability issues compared with not using any cues at all. The findings also suggest that using a gaze plot or gaze video cue stimulates participants to produce the highest number of words and comments, and mention more usability problems.**

*Eye tracking, usability testing, usability, retrospective think aloud,concurrent think aloud.*

## 1. INTRODUCTION

Eye tracking devices has become a popular tool in a wealth of areas, such as neuroscience, computer science, psychology and market research [4]. Eye tracking has also emerged as a promising method for detecting usability problems, especially in websites [1][3][6][14][16][22]. When using eye tracking in usability studies, it is important to select the most suitable methodology in order to extract relevant and useful data from the participants. The goal of this study is to compare four different retrospective think aloud methods in a website usability test: RTA without any cue, RTA with a video cue (screen video), RTA with gaze plot cues (superimposed eye movements on still images), and RTA with a gaze video cue (superimposed eye movements on a screen video). Previous research has, been done on using no cue, video replay and gaze video replay, but not on using a static gaze image as a cue when doing RTA[5] [10]. The paper is structured as follows: Initially, a brief introduction to eye tracking is given followed by a short presentation of the think aloud methodology as well as why RTA is seen as a suitable method in combination with eye tracking. This is followed by a walkthrough of what material and apparatus were used, what participants were selected for the study, the procedure of the testing and how the data collected was analyzed. The results are then presented and discussed and followed by a brief section containing the conclusions drawn from the study.

## 2. BACKGROUND

### 2.1 How Eye Tracking Works and What Is Measured

The basic idea behind eye tracking is that our eye movements can be used to make inferences about our cognitive processes [14]. An eye tracker follows the user's eye movements by reflecting infrared light onto the eye and then, using a geometrical model, determines the exact gaze point of the user [18]. Although eye tracking has been around since the 1800s, recent advances have made the devices easier to use, not only for the researchers, but also for the participant. The first eye trackers were rather invasive [11], a stark contrast to the remote eye trackers available today. Remote eye trackers, e.g. the Tobii T60/T120/T60 XT, allow participants to sit comfortably in front of a screen equipped with a built-in eye tracking device. To most users, the screen will look almost like a normal computer screen, making it a comfortable and familiar device to work with.
Most eye tracking studies aim at analyzing patterns of visual attention of individuals when performing specific tasks (e.g. reading, searching, scanning an image, driving, etc.). In these studies eye movements are typically analyzed in terms of fixations - a pause of the eye movement on a specific area of the visual field, and saccades - rapid movements between fixations [11]. This data is usually illustrated using gaze plots (or scan paths) which show saccades and fixations or aggregated heat maps which show the amount of or length of fixations [15].).

### 2.2 Think Aloud Methods

For usability research, eye tracking data should be combined with additional qualitative data because eye movements cannot always be clearly interpreted without the participant providing context to the data [10]. For example, longer fixations can mean a user found a particular area interesting [3], but it can also mean that they found the area difficult to interpret [10]. Hence, it is important to attempt to supplement eye tracking data with additional information gained from the participants about their experiences [ibid.].
Think aloud methods are often used when attempting to detect usability problems [8] [20] [22]. There are generally two ways to conduct a think aloud interview: either the participants are asked to verbalize their thoughts while they are doing tasks, i.e. concurrent think aloud (CTA), or the participants provide a description of their experiences doing the tasks after each or all of the tasks are completed, i.e. retrospective think aloud (RTA) [10]. Both are relatively simple methods of gaining insight into the participants' thought processes regarding task completion [21]. However, each of these methods offers its own set of problems or limitations which should be considered when selecting a methodology. These include:

- Think aloud processes may not be sufficient since certain cognitive processes are unconscious and participants may not be able to adequately verbalize their thought process [5].

- For CTA methods, an issue is that cognitive processes are quicker than verbal processes, so participants might be thinking about more than they are able to verbally express [5].

- CTA is more easily affected by reactivity; participants may perform better or worse in completing tasks due to the nature of the task, i.e. some tasks seems to be easier to do when CTA is used as the participant is forced to structure their thoughts which also means structuring their actions, while other appears harder as the cognitive workload increases as the participant has to both talk and interact with the computer simultaneously [20].

- CTA does not easily allow for measuring timing variables related to how long it takes to complete a task [20].

- CTA may potentially bias the participant's first impressions, whereas RTA may lead to forgetting their first impressions [2].

- Many participants forget to express their thought processes aloud when encountering difficulties interacting with the user interface if CTA is used [8].

- RTA relies on our highly fallible long-term memory [1], participants may be forgetting important steps in the task, or may be intentionally or unintentionally fabricating information [17].

### 2.3 Combining Think-Aloud and Eye Tracking

Using CTA in combination with eye tracking has proven to be less suitable as participants then produce eye movements which they would not normally do if completing their task on their own in their normal environment [12], e.g. looking away from the screen to describe something to the researcher or by focusing on certain areas of the screen while describing their thought processes regarding that area. Therefore, the RTA method is the recommended method when conducting usability tests where also objective eye movement data will be analyzed.
Since fallible memory and potential for fabrication can be problems when performing traditional RTA usability tests, a variety of cued RTA methods have emerged. In a cued RTA the user is presented with a form of replay of the interactions they previously performed in order to help cue their memory [21]. Replays could be, e.g., a video replaying their actions, screen shots, superimposed eye movements on a video, etc. This integrated approach to usability testing has proven to be a way to gain richer data from participants [13]. Presenting these visual stimuli serves as a way to get more detailed information, but also allows the participants to reflect upon their actions in a way they might not have been able to do otherwise [2]. The information gathered from the eye tracker accounts for much of the quantitative data needed, whereas the cued RTA provides qualitative data input from the participants.
Using a video cue that features eye movements (a gaze video replay) has been demonstrated as more effective at eliciting comments from users than an un-cued RTA [17]. Showing a playback of participants' eye-movements overlaid on a video showing the steps they took while completing a task has proven to be a successful way to elicit information from the participants and, in addition, allows for an accurate measure of other variables, such as task time [5][21]. The post-experience eye-tracked protocol (PEEP) method that utilizes playbacks of people's eye movements during RTA has shown to be potentially better than a video without eye movements when exploring new or complex environments [1]. One study showed that even if the participants stated

that they mostly relied on their memory when talking about a recently conducted task, they did find the video helpful as a reminder [8]. In addition, when using video as stimuli for cued RTA, recollections of the task turned out to be very accurate according to actual task performance, i.e. the video almost eliminated the risk of fabrication [ibid.].

## 3. CURRENTSTUDY

The goal of the current study is to compare four methods of RTA: RTA without any cue, RTA with a video cue (screen video), RTA with gaze plot cues (superimposed eye movements on still images) and RTA with a gaze video cue (superimposed eye movements on a screen video). The aim is to examine which method is more effective at eliciting comments from the participant and gaining information regarding usability problems found on a website. In addition, the study aims to explore the usefulness of a gaze plot as a cue when using RTA since this does not seem to have been explored in previous research. The dependent variables will be the total number of usability problems identified, the word count and the number of comments given by participants as this indicates how much, and about what participants talk in the four conditions.
The hypothesis is that using a cued RTA will result in more comments and words as well as the identification of a larger number of usability problems than when using an un-cued RTA. Additionally, using the RTA method with an eye tracking cue (gaze plot or gaze video) is believed to produce more comments and words, and will help participants to identify more usability problems compared with an un-cued RTA procedure.


*Figure1.Spotify homepage.*

### 3.1 Apparatus and Materials

The Tobii T120 remote eye tracker along with the Tobii Studio 2.0 (Enterprise version) software was used to record (in 60 Hz) and replay participants' eye movements. The software, Tobii Studio, allows the researcher to create gaze plots and play back a video recording of the screen, both with and without eye movements, needed for this study. The version of Tobii Studio used in this study also includes an automatic RTA recording function where the researcher can video and audio record the participant's reactions while showing the results from the previously recorded tasks, including gaze plots and video playbacks.
The website used in this test was www.spotify.com (see Figure 1). On the website users can register for a paid subscription and then download a software client that provides legal streaming of music. Participants in the test were given one task to complete on the website: Register for a monthly subscription with Spotify, complete the payment procedure (using a provided credit card) and then download the Spotify software. While completing this task on the website the participants' eye movements and the screen was recorded using Tobii Studio 2.0.

### 3.2 Participants

Opinions regarding the optimal number of participants included in a usability study vary, but generally a number between 5 and 15 participants is given (depending on the nature of the study) [7] [19]. The commonly used argument is that about 5 participants are needed in a usability study to identify 80% of the usability issues of a website [19]. As one of the objectives of this study was to compare the number and type of usability problems identified in the different conditions, the number of participants included was the same amount as would be included in a 'normal' usability study in a realistic, commercial setting.
In this study, four conditions were to be tested meaning that enough participants needed to be included in every group to allow comparisons between the different conditions. This means that 6 participants were recruited to test each condition. This means that, in total, 24 participants were included in this study.
Convenience sampling was used to recruit participants for the study. Data collection spanned over three days and the primary locations for data collection were a café and a hotel in Stockholm, Sweden, that provided access to a wide variety of potential participants. The participants were offered the incentive of either two lottery tickets or a gift card at a café. The conditions used when recruiting participants were that they did not currently use the Spotify program or website (as these were going to be tested), and that they liked music. In addition, as testing and analysis was conducted in English, only participants with moderate to good skills in the English language were recruited. The participants in the study represented several different nationalities. In order to ensure homogeneity among the participants, they were given a pre-test questionnaire asking questions such as: how often they listen to

music and if they have ever bought music online. One-way analysis of variances (ANOVAs) were conducted to look for any significant differences between the four groups in internet experience, previous experience with buying music online, listening to music on computers and using music software. None were found, indicating that the groups had similar levels of experience within the given factors. Each condition was tested with three male and three female participants except for the 'gaze plot' condition, which included four male and two female participants.
A randomized list of the four different RTA conditions, i.e. no visual cue, video replay, static gaze plot and video replay including gaze, was created prior to testing. Each person chosen to participate was included in whatever condition was next on the randomized list. In the end, each condition had been tested with six participants.

### 3.3 Procedure

Potential participants were initially asked screening questions to ensure that they were suitable for participation and were then introduced to the study. They were also asked to sign an audio and video consent and release form since their comments and eye movements were recorded during the test. During the actual testing the researcher followed an interview script as guidance. The first section of the session included an introduction and a brief explanation of the test. This was followed by a quick calibration of the eye tracker. Once the calibration was completed successfully, participants were provided with the tasks and instructed to begin completing the task using Internet Explorer. The participants were encouraged not to speak or think out loud while performing the tasks as the method chosen for the test was RTA.
The interview script was created in a way that attempted to ensure consistency between the four different test conditions. Upon completion of the tasks, participants in the un-cued RTA interview were asked to reflect and provide insight about the task (i.e. to register for a monthly subscription of Spotify) they had just completed while looking at a blank screen, i.e. no visual cue was provided. Participants in the three cued RTA conditions were shown either a video playback of the task they had just completed along with their eye movements superimposed on the screen, a static gaze plot showing their eye movements on each separate web page visited during task completion or a screen video without eye movements showing their user journey while completing the task. At the same time they were asked to talk about the task they had just completed. Users who were shown a video were also told they could fast forward, rewind or pause the video if they wanted. The video was by default shown at half speed from the start of the RTA interview. For all conditions, the RTA recording function in Tobii Studio 2.0.x was used to record the interview including audio and screen video. The participants could also use the mouse to point at things on the screen during the interview.

*Table 1: Categories used for categorizing usability problems.*

| | |
|---|---|
| **Layout** | **Inability to detect something in the screen that they need to find; Aesthetic problems; Unnecessary information** |
| **Terminology** | **Unable to understand the terminology** |
| **Feedback** | **User does not receive relevant feedback or it is inconsistent with what the user expects** |
| **Comprehension** | **Inability to understand the instructions given to them on the site** |
| **Data Entry** | **Problems with entering information** |
| **Navigation** | **Problems with finding their way around the site** |

*Table 2: Categories used for classification of participants' comments.*

| | |
|---|---|
| **Manipulative** | **Comments that express an action, e.g. "I enter my password in this box"** |
| **Visual** | **Depict what the user sees/ wants to see, e.g. "I am looking for the link"** |
| **Cognitive** | **The users interpretations, assessments and expectations, e.g. "Now I understand why the link wasn't clickable"** |

### 3.4 Measuring Usability Problems and Comments

Verbal transcripts were produced and analyzed after completion of the study. Identified usability problems and comments were picked out and finally compared. In previous usability research on website interface design and eye tracking, six usability problem categories were identified and defined [5] [20] (see Table 1). As the categorizations seemed well grounded in theory and experience as well as being suitable for the study at hand the same categorization was chosen to be used for this study. Previous research using the categorizations given above has shown that RTA with eye movement video cues have been particularly successful at detecting usability issues related to feedback and comprehension and has generally been proven to detect more usability problems than other think aloud methods [5].

Other research has focused on the amount of words and comments produced by RTA, dividing comments into the following categories [9][10] (see Table 2). By counting the number of words a comparable measure of how much the participants talked during the different conditions was collected. To expand on the potential findings regarding identified usability problems, classifying the comments given in relevant categories was believe do make the analysis richer and the conclusion more well grounded. As the categories shown in Table 2 had proven useful for other studies where RTA and eye tracking was studied, they were chosen to be used also in this study.
Previous research suggests that RTA in combination with eye movement video replay typically produce more cognitive comments than manipulative or visual comments [10].

### 3.5 Design and Data Analysis

The study was designed as a between-participant design study, i.e. each group of participants were subjected to only one of the independent variables respectively, with the cue condition (no cue, video cue, gaze plot, and gaze video) as the independent variable and word count, usability problem category (layout, terminology, data, entry, comprehension, feedback and navigation) and comments category (manipulative, cognitive and visual) as the dependent variables. One-way ANOVA and post-hoc analysis (Tukey's) were later conducted to look for significant differences between each group.

## 4. RESULTS

### 4.1 Word Count

The word count data was analyzed for normality and one outlier was found in the 'gaze plot cue' group (having a total word count more than three standard deviations from the mean). This participant was excluded from the word count analysis. Statistically significant differences were found between the groups that had been subjected to the different RTA conditions when analyzing the total number of words used per participant in the interview (one-way ANOVA $F(3,19)=4.358$, $p < .05$). A Tukey's post-hoc analysis revealed that the 'no cue' condition produced significantly fewer words than the 'Gaze Video Cue' ($p < .05$). Although no other groups revealed significant differences, Table 3 shows a trend toward the 'gaze video cue' and 'gaze plot cue' producing more words than a regular 'video cue' and 'no cue'.
The total interview time used for the RTA part of the study, indicated by length of the audio recordings, was also analyzed and it followed the same pattern as the word count results above. The average length of the 'no cue' interviews was 56 seconds. For the 'video cue' group, the interviews lasted on average 220 seconds while the averages were 225 seconds and 257 seconds for the 'gaze plot' and 'gaze video' groups respectively.

### 4.2 Usability Problems Identified

The number of unique comments and usability problems mentioned in the different groups were analyzed using the categorization model previously discussed. Table 4 and Table 5 illustrate the guidelines for categorizing comments and usability problems mentioned by participants. Each guideline is exemplified by actual verbatim quotes given by participants in the study.

*Table 3: Average and total number of words produced per participant in the different condition groups. N=23, one outlier removed in the gaze plot cue condition. Significant differences were observed between the no cue condition gaze video cue ($p<0.5$) condition, marked by * in the table.*

| | Average Number of Words | Total Number of Words |
|---|---|---|
| **No Cue*** | **47** | **282** |
| **Video Cue** | **157** | **942** |
| **Gaze Plot Cue** | **207** | **1033** |
| **Gaze Video Cue*** | **262** | **1571** |

*Table 4: Coding system used for categorizing comments*

| Comment Category | Definition |
|---|---|
| **Manipulative** | **Comments that express an action. E.g. "I selected the premium monthly subscription"** |
| **Visual** | **Depict what the user sees/wants to see. E.g. "I didn't see the optional thing."** |
| **Cognitive** | **The user's interpretations, assessments and expectations. E..g. "Then it was this field where I made the decision to click on this one."** |

*Table 5: Coding system used for categorizing usability problems mentioned by the participants*

| Problem Category | Definition |
|---|---|
| Layout | Inability to detect something that they need to find; Aesthetic problems; Unnecessary Information. E.g. "This could be larger" |
| Terminology | Unable to understand the terminology. E.g. "Premium what does it mean?" |
| Feedback | User does not receive relevant feedback or it is inconsistent with what the user expects. E.g. "When I corrected that one and then went on to the next one and tried to change the confirmed password, it didn't automatically mark up everything." |
| Comprehension | Inability to understand the instructions given to them on the site. E.g. "I wasn't sure if I was supposed to click the Visa or not." |
| Data Entry | Problems with entering information. E.g. "I spelled [the email] wrong, I didn't notice. They should probably have two fields for the email address if they really want it, because that would end up in the wrong place." |
| Navigation | Problems with finding their way around the site. E.g. "..because the sign up and buy premium [option being available], I would go with the 'sign up' [option] and that was wrong." |

As presented in Table 6 and Table 7, the results indicated a trend towards the 'gaze plot' cued and 'gaze video' cued groups identifying the highest number of unique usability problems, with the 'no cue' condition producing the fewest. A one-way ANOVA revealed a significant difference between the groups concerning the average number of comments yielded, $F(3, 20) = 3.981$, $p < .05$ and post-hoc tests (Tukey's) revealed that the 'no cue' group had significantly fewer comments than the 'gaze video cue' group ($p < .05$). However, for the average amount of identified usability problems, no statistically significant differences were found between any of the groups, although the table does indicate a trend toward more problems identified in the 'video cue', 'gaze plot cue' and 'gaze video cue' groups. Table 6 and Table 7 present the total number of unique comments and usability problems identified.

*Table 7: The total number of unique comments as well as the average number of comments made in the different groups and a breakdown of the unique comments per group.*

| | Total Com. | Av. Com. | Mani. | Cogn. | Visu. |
|---|---|---|---|---|---|
| No Cue | 3 | 0.8 | 1 | 0 | 2 |
| Video Cue | 24 | 4.8 | 9 | 6 | 9 |
| Gaze Plot Cue | 30 | 5.3 | 6 | 10 | 14 |
| Gaze Video Cue | 36 | 6.8 | 11 | 12 | 13 |

The different cues appeared to stimulate the participants in different ways when commenting on their behaviour. The two video conditions stimulated the participants to produce slightly more 'manipulative' comments than when a static gaze plot or no cue was used. This is likely because the participants were able to see the entire chain of events when interacting with the website (video) as compared with having to rely only on their memory (no cue) or when only seeing the individual pages separately (gaze plot). However, both eye movement conditions stimulated the participants to produce

*Table 6: The total number of unique usability problems mentioned as well as a breakdown of the different kinds of usability problems mentioned per participant group. Total numbers of usability problems are presented within the brackets.*

| | Total Number of Unique Usability Problems | Layout | Terminology | Data Entry | Comprehension | Feedback | Navigation |
|---|---|---|---|---|---|---|---|
| No Cue | 3(5) | 2(4) | 0 | 1(1) | 0 | 0 | 0 |
| Video Cue | 11(13) | 4(5) | 2(2) | 1(1) | 1(1) | 1(1) | 2(3) |
| Gaze Plot Cue | 12(20) | 2(3) | 4(9) | 0 | 5(7) | 0 | 1(1) |
| Gaze Video Cue | 12(14) | 5(7) | 1(1) | 3(3) | 2(2) | 1(1) | 0 |

slightly more 'cognitive' and 'visual' comments than the other two tested methods. One other general finding is that participants in the 'no cue' group tended to trivialize any problems they had and over-generalized the process as easy and problem free. This is also indicated by the fact that the number of usability problems mentioned as well as the number of comments in the 'no cue' group was very low compared to the other groups.

In addition, there were minor differences between the different conditions in terms of type of usability problems mentioned. Both 'video' conditions produced slightly more usability problem-related comments about the layout than did the 'gaze plot' and the 'no cue' group. However, participants in the 'gaze plot' condition commented somewhat more on usability problems categorized as 'terminology' and 'comprehension' than did the two 'video' conditions.

The 'gaze plot cued' RTA method has not been examined in depth in previous research. The results from this study indicated that using a gaze plot as a cue is a successful method of eliciting comments from the participant in a web usability test. Presenting the user with an image instead of a video allowed the participants to take as long as they needed to discuss each web page seen during the task completion. Even if the participants could stop the two video cues at any time, many didn't use this option very much. Hence, the possibility to reflect on details in their interaction such as comprehension of terms seemed to be more prominent in the gaze plot condition. Interestingly, the interviews in the gaze plot condition took almost as long as the interviews where a gaze video was replayed at half speed. Most participants did replay the entire video with short stops to describe certain elements. In the gaze plot condition, the participant was probed in much the same way as in the video cue conditions, but even though there was not such a concrete time illustration as when being replayed a video, the interviews still a similar average length as the interviews where a video cue was used.

The gaze plot cued RTA method also produced a higher number of words, comments and usability problem-related comments than the 'no cue' and 'video cue' RTA methods, but did not perform as well as the gaze video cued RTA method. Figure 2 shows one example of data that might not have been collected if the static'gaze plot' cue had not been used. This particular participant specifically noticed and commented on seeing his eye movements to the 'Help' button in the top menu, an observation that may have been missed in un-cued or just video cued format.


*Figure 2: Participant's eye movements on the Spotify 'Download' Page.*

## 5. CONCLUSIONS

The purpose of this study was to compare four methods of cued RTA: RTA without any cue, RTA with a video cue, RTA with a gaze plot cue, and RTA with a gaze video cue. The aim was to examine

*Table 8: Benefits of using the different RTA methods included in the study*

| No cued RTA | Video cued RTA | Gaze plot cued RTA | Gaze video cued RTA |
|---|---|---|---|
| | | **Stimulated participants to produce 'visual' and 'cognitive' comments** | **Stimulated participants to produce 'visual', 'cognitive' and 'manipulative' comments.** |
| | **Stimulated participants to produce 'manipulative' and 'visual' comments** | **Stimulated participants to comment on usability problems regarding 'terminology' and 'comprehension'** | **Stimulated participants to comment on usability problems regarding 'layout' and 'data entry'** |
| | **Stimulated participants to comment on usability problems regarding 'layout' and 'navigation'** | **The second best performing RTA method; produced the second highest number of words and comments** | **The best performing RTA method; produced the highest number of words and comments.** |
| **Produced significantly less data (comments and words) than any of the cued RTA methods.** | **Produced less data (comments and words) than eye movement cued RTA methods.** | | |

which method was slightly more effective at eliciting comments from the participants and gaining information regarding usability problems found on a website. The results presented in this paper suggest that using any cue will stimulate participants to provide a higher number of words and comments as well as help participants to identify more usability problems compared to when no cue is used. Additionally, by using gaze plots or gaze videos as cues, participants provided more feedback than when only using a screen video as cue; The method that produced the highest quantity of interview data, in number of comments and number of words, was the gaze video cued RTA method, followed by the gaze plot cued RTA method. Another interesting conclusion is that the gaze plot cued method proved to perform almost as well as the gaze video cued method, being especially good at producing visual and cognitive comments while simultaneously identifying the same amount of usability problems as the gaze video cued method.

Differences in the types of comments produced during the interviews and the types of usability problems mentioned were also observed. Eye movement cued RTA methods tend to stimulate participants to make more visual and cognitive comments, while video cued RTA methods produce somewhat more manipulative comments. No clear patterns were observed for the different categories of usability problems described in the different groups, but the participants in the cued RTA groups mentioned more usability problems than those in the non cued RTA group. Table 8 below summarizes the results from this study.

The goal of most usability testing is to identify usability problems and to gather relevant information from the participants. Consequently, it is beneficial to use a gaze plot or a gaze video as cue in an RTA interview to reach these goals as these two methods seem to stimulate the participants to give more feedback. Overall, using eye movements in combination with RTA proved to be a successful method for learning more about users' problems with a website. The eye movement cued RTA method provided both qualitative information in the form of the interview comments, and quantitative data from the eye tracker that could be analyzed in the context of qualitative feedback provided by the participants. The results from this research project can help usability researchers choose the most suitable RTA method for testing as the findings showed considerable differences in the outcomes of the different methods.

The value of gaze plots in RTA is an area that should be explored further in future research as it has only rarely been explored in the past. In addition, there is need for further research into the participant's experience of being subjected to the different cues investigated in this study (i.e., asking participants what they thought of the cue provided to them).

## 6. ACKNOWLEDGMENTS

Our thanks to everyone who contributed to this study.

## 7. REFERENCES

[1] Ball, T.J., Eger, N., Stevens, R., Dodd, J. 2006. Applying the post-experience eye-tracked protocol (PEEP) method in usability testing. Interactions, 67,15-19.

[2] Bartels, M. 2008. The objective interview: Using eye movement to capture pre-cognitive reactions. Qualitative Research Consultants Association: Views, 6, 3,58-61.

[3] Cowen, T., Ball, T.J., Delin, J. 2002. An Eye Movement Analysis of Webpage Usability. In People and Computers XVI - Memorable yet Invisible: Proceedings of the H CI 2002'(Tondon, UK), Springer-Verlag Ttd., Tondon, 317-335.

[4] Duchowski, A.T. 2002. A breadth-first survey of eye tracking applications. Behavior Research Methods, Instruments, & Computers, 34,455-470.

[5] Eger, N., Ball, T.J., Stevens, R., Dodd, J. 2007. Cueing
Retrospective Verbal Reports in Usability Testing Through Eye-Movement Replay. In Proceedings of the 21 St British CHI Group Annual Conference on H CI 2007: People and Computers XXI: HCI.butnotas we know it (Beijing, P.R. China), British Computer Society, 129-137.

[6] Ehmke, C., Wilson, S. 2007. Identifying Web Usability Problems from Eye-Tracking Data. In Proceedings of the 21st British CHI Group Annual Conference on H CI 2007: People and Computers XXI: H CI.but not as we know it (Beijing, P.R. China), British Computer Society, 119-128.

[7] Faulkner, T. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods, Instruments, & Computers, 3, 35, 379-383.

[8] Guan, Z., Tee., S., Cuddihy, E., Ramey, J. 2006. The Validity of the Stimulated Retrospective Think-Aloud Method as Measured by Eye-Tracking. In Proceedings of the SIGCHI conference on Human Factors in computing systems (Montréal, Québec, Canada), ACM, 1253-1262.

[9] Hansen, J.P. 1991. The use of eye mark recordings to support verbal retrospection in software testing. Acta Psycho lógica, 76,1,31-49.

[10] Hyrskykari, A., Ovaska, S., Majaranta, P., Raiha, K-I, Tehtinen, M. 2008. Gaze path stimulation in retrospective think aloud. Journal of Eye Movement Research, 2,4,1-18.

[11] Jacob, R.J.K, Karn, K.S. 2003. Commentary on section 4: Eye tracking in human-computer

interaction and usability research: Ready to deliver the promises. In The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research. Elsevier Science, Amsterdam, 573-605.

[12] Kim, B, Dong, Y, Kim, S, and Tee, K-P. 2007. Development of Integrated Analysis System and Tool of Perception, Recognition, and Behavior for Web Usability Test: With Emphasis on Eye-Tracking, Mouse-Tracking, and Retrospective Think Aloud. In Usability and Internationalization. H CI and Culture. (Beijing, P.R. China), Springer, 113-121.

[13] NAMAHN. Using eye tracking for usability testing. Namahn. 2001 Retrieved June 1,2009: www.namahn.com/resources/documents/note-eyetracking.pdf.

[14] Peyrichoux, I and Robillard-Bastien, A. Maximize Usability Testing Benefits with Eye Tracking. Getting a Measure of Satisfaction from Eyetracking in Practice - Workshop at CHI 2006 2006 Retrieved June 1, 2009: http://www.amber-light.co.uk/CHI2006/chi2006-eyetracking-workshop-position-paper-ipeyrichoux-arbastien-FINAT.doc. Position paper.

[15] Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 3, 372-422.

[16] Russel, M. 2005. Using Eye-Tracking to Understand First Impressions of a Website. Usability News, 1,1.

[17] Russo, JE. 1979. A software system for the collection of retrospective protocols prompted by eye fixations. Behavior Research Methods & Instrumentation, 11,2,177-179.

[18] TOBII TECHNOTOGY. What is Eye-Tracking? Tobii Technology. Retrieved June 1,2009:
http : //www. tobii. com/corporate/ey e_tracking/ what_is_ey e_t racking, aspx.

[19] Turner, Carl W, Tewis, James R, and Nielsen, Jakob. 2006. Determining Usability Test Sample Size. International Encyclopedia of Ergonomics and Human Factors, 3,3084-3088.

[20] van den Haak, M.J., de Jong, M.D.T., and Schellens, J. 2003. Retrospective vs. Concurrent think-aloud protocols: testing the usability of an online library catalogue. Behavior & Information Technology, 22, 5,339-351.

[21] van Gog, T., Paas, F, van Merrienboer, J.J.G, and Witte, P. 2003. Uncovering the Problem-Solving Process: Cued Retrospective Reporting Versus Concurrent and Tracking Retrospective Reporting. Journal of Experimental Psychology: Applied, 11,4, 237-244.

[22] Wulff, A. 2007. Eyes Wide Shut – Or Using Eye Tracking Technique to test a Web Site. International Journal of Public Information System, 2007:1, 1-12.