Journal of Eye Movement Research
2(4):5, 1-18

Gaze Path Stimulation in Retrospective Think-Aloud

Aulikki Hyrskykari
University of Tampere

Saila Ovaska
University of Tampere

Päivi Majaranta
University of Tampere

Kari-Jouko Räihä
University of Tampere

Merja Lehtinen
University of Tampere

For a long time, eye tracking has been thought of as a promising method for usability testing. During the last couple of years, eye tracking has finally started to live up to these expectations, at least in terms of its use in usability laboratories. We know that the user's gaze path can reveal usability issues that would otherwise go unnoticed, but a common understanding of how best to make use of eye movement data has not been reached. Many usability practitioners seem to have intuitively started to use gaze path replays to stimulate recall for retrospective walk through of the usability test. We review the research on think-aloud protocols in usability testing and the use of eye tracking in the context of usability evaluation. We also report our own experiment in which we compared the standard, concurrent think-aloud method with the gaze path stimulated retrospective think-aloud method. Our results suggest that the gaze path stimulated retrospective think-aloud method produces more verbal data, and that the data are more informative and of better quality as the drawbacks of concurrent think-aloud have been avoided.

Keywords: Think-aloud protocol, gaze path stimulated retrospective think-aloud, usability testing

Introduction

Eye tracking is an increasingly popular method in usability evaluation (e.g. Bojko, 2006). However, it has not been as fruitful as one could expect (Jacob & Karn, 2003), and a systematic methodology for the use of eye trackers in usability evaluation has not emerged. Interpreting the recorded data is intricate. Although eye tracking tells us what users look at, it does not tell us why.

For example, a prolonged gaze to some widget does not necessarily mean that the user does not understand the meaning of the widget. The user may just be pondering some aspect of the given task unrelated to the role of the widget on which the gaze happens to dwell. Similarly, a distinctive area on a heat map is often interpreted as meaning that the area was interesting. It attracted the user's attention, and therefore the information in that area is assumed to be known to the user. However, the opposite may be true: the area may have attracted the user's attention precisely because it was confusing and problematic, and the user did not understand the information presented. A conclusion is that the data recorded by eye tracking seems to call for the user's interpretation.

In usability tests, such interpretive information is commonly obtained through the *think-aloud* (TA) method (Nielsen, 1993; Boren & Ramey, 2000; Van den Haak, De Jong, & Schellens, 2003). It is a way to gain insight of the user's cognitive processes during the use of a product. Usually the verbalizations are collected during the execution of the tasks in a usability test (Denning, Hoiem,

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Simpson, & Sullivan, 1990). We call this the *concurrent think-aloud* (CTA) method.

Concurrent think-aloud is still the predominant data collection method in usability testing (Nielsen, Clemmensen, & Yssing, 2002). However, its shortcomings are well known. Many users find thinking aloud difficult and it makes them feel uncomfortable (Nielsen, 1993). Since we think much faster than we are able to verbalize our thoughts, "thinking aloud" is actually an unreasonable demand. Verbal protocols measure mainly conscious thoughts that can be easily verbalized (Ericsson & Simon, 1993); for example, automated processes are hard to transfer into think-aloud. Some participants may be unable to think aloud while performing a cognitively demanding task, or their verbalizations may be very brief and procedural (Branch, 2001). Consequently, the verbalizations are incomplete (Wilson, 1994). In addition, thinking aloud probably affects the user's task performance (Nielsen et al., 2002; van Someren, Barnard, & Sandberg, 1994; Guan, Lee, Cuddihy, & Ramey, 2006). An obligation to verbalize the performed processes may slow down the normal behavior with the product and even change the steps of execution from the ones the user would take in a normal situation.

One of the suggested ways to overcome these problems is to use the think-aloud method after the test, not during it. The participant is asked to carry out the given tasks without an obligation to think aloud, which hopefully makes the interaction with the product more natural. After the task the participant gives a verbal report of the task session. We call this the *retrospective think-aloud* (RTA) method. In retrospective think-aloud it is common to prompt the participant with visual reminders of the task, hence the terms "stimulated RTA" (Guan et al., 2006) and "cued retrospective reporting" (van Gog, Paas, van Merriënboer, & Witte, 2005) have also been used. Other phrases found in the literature include "retrospective testing" (Nielsen, 1993), "think after" (Branch, 2000), and Post Experience Eye tracking Protocol (PEEP) (Maughan, Dodd, & Walters, 2007).

The usual way to stimulate RTA is to present the user a playback of the test session. It typically includes a video of the screen showing the mouse movements and possibly an inserted video of the user. Adding an overlaid replay of the user's gaze path to the video playback can further facilitate the users' recall of their thoughts during the test (Hansen, 1991; Ball, Eger, Stevens, & Dodd, 2006).

We compared the quality of the verbal data received in a usability test during CTA versus the data gathered retrospectively after the test while watching a playback of the session augmented with an overlaid gaze path animation. We found that users produced significantly more verbal data in RTA than in CTA. Analyzing the quality of the verbalizations revealed that in RTA the comments reflected more cognitive operations whereas in CTA the emphasis was on manipulative comments.

In the following sections, we first provide a review of research on retrospective think-aloud, eye tracking and the combination of these in usability testing. We then introduce our own experiment, summarize the results of the experiment, and discuss how the method should be used in usability testing. We conclude by summarizing the work.

Previous Research on Think-Aloud and Eye Tracking in Usability Evaluation

Studies of Retrospective and Concurrent Think-Aloud Protocols

The data obtained with concurrent think-aloud protocols refer to the actual use of the product and not the participants' judgment of its usability, and that kind of data have high face validity—CTA helps to find real usability problems (Van den Haak et al., 2003). In usability evaluation the main goal is to identify those areas of a design that need refinement, especially areas that do not work as anticipated or cause problems to the users (Ebling & John, 2000). Of the empirical data sources, the think-aloud protocol is one of the best sources for identifying usability problems (Ebling & John, 2000).

However, even if the think-aloud protocols can be very rich in diagnostic and evaluative information, they can also be contrived, biased, and misleading. They require the user to constantly verbalize what he or she is thinking, expecting, deciding, etc. Verbalizing while doing can be very demanding and can change the course of one's behavior (Rhenius & Deffner, 1990). Russo, Johnson, and Stephens (1989) call a protocol *reactive* if verbalization changes the primary process or lengthens it. According to their work, collecting a concurrent think-aloud protocol can affect the actual test situation, change

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

the course of events or even improve performance in some tasks. When compared with a silent condition, producing a concurrent verbalization lengthens the task performance. Similar observations are common in usability testing in general (e.g., Nielsen et al., 2002; Van Someren et al., 1994).

To avoid the problem of trying to do two things, the primary test task and think-aloud, at once, the method of retrospective report can be employed. The user first performs the task without verbalization and then either produces the think-aloud protocol from memory or reviews a recording of the interaction and comments on the events as they are replayed. When the retrospective report is given right after the task session, the user still has part of the information in short-term memory (Ericsson & Simon, 1993). Using a supporting video playback or other memory cues to stimulate the retrospective report helps the retrieval of information also from long-term memory.

Another concern expressed by researchers of protocol analysis (e.g., Russo et al., 1989; Ericsson & Simon, 1993; Guan et al., 2007) is *veridicality* of the protocol—that is, if the protocol reflects the original primary task. RTA, for example, might contain errors of omission or commission: some thoughts might be left out or fabricated without a link to the actual thought processes of the primary task.

Users may report their actions or thought processes in a manner they believe the test leader wants to hear. Especially the probes that the experimenter presents to them have an impact on how the protocol evolves (Ericsson & Simon, 1993). Furthermore, the verbalizations may contain judgments or strategies which are more rational than the participants actually applied while carrying out the test. They may explain their actions in a fashion that makes them look more systematic, rational, organized, thoughtful or coherent (Kuusela & Paul, 2000).

Ericsson and Simon (1993) argued that when the verbal protocol is based on unsupported memory recall RTA provides valuable data only on simple tasks, but that the technique is not valid in lengthy and complex tasks. They went on to argue that the users' cognitive processes may have changed so dramatically after completing the task that they may be unable to provide an accurate account of the thinking and problem solving strategies they had whilst completing the task. Similarly, Kuusela, & Paul (2000) point out that retrospective protocols are incomplete regarding the steps taken and information considered during the decision-making process. While good at describing the decision outcome, they lack the details of the process, and those can be better obtained with a concurrent think-aloud protocol (Kuusela & Paul, 2000).

The veridicality concern is important for research on memory processes and problem solving strategies. However, in the context of our study veridicality issues do not emerge, since the retrospective think-aloud protocol is cued by the video playback to aid recall.

Despite all the problems in RTA, Guan et al. (2006) found RTA to be valid and reliable when they compared participants' retrospective verbalizations with a captured record of their eye movements during the test. According to their research, RTA provided a valid account of what people attended to in completing tasks, the technique had low risk of introducing fabrications (reports of events that in fact did not occur), and its validity was unaffected by task complexity.

Bowers and Snyder (1990) found that the users of CTA produced more words than the users of RTA, and that there was a difference in the content of verbalizations. In CTA, users were more likely to read texts on the screen and describe their own actions. In RTA, the users were more likely to explain their actions or give suggestions on how the product design could be enhanced. The study of Bowers and Snyder did not, however, use the gaze path to stimulate the retrospective verbalization.

If researchers want to enhance the completeness, reliability and validity of the data collected with the verbal protocols, they should collect both the concurrent and retrospective protocols, since they complement each other (Taylor & Dionne, 2000). It is common that the CTA is not complete, and reviewing it with the participant gives additional information (Van Someren et al., 1994). However, a big problem with the retrospective account is that it takes substantially longer than the concurrent think-aloud method (Norman & Murphy, 2004). As noted above, CTA also prolongs the evaluation, and if the techniques are used together the effect is pronounced.

Eye Tracking in Usability Tests

Goldberg and Wichansky (2003) discuss areas where eye tracking has been used. Usability evaluation of products under construction is one of them (Goldberg & Kot-

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

val, 1999). Eye tracking has been applied successfully to produce design suggestions.

For instance, based on eye tracking data Goldberg, Stimson, Lewnstein, Scott, and Wichansky (2002) point out that users are more likely to choose buttons in the upper left corner; hence the important information should be placed there. Similarly, Pretorius, Calitz, and Van Greunen (2005) indicate how eye tracking data give added value to conventional usability evaluation methods: in some tasks when all the users were able to complete the task and gave the correct answers in time, the gaze paths showed that the users struggled in finding the requested information because of the cluttered screen layout. Penzo (2006a; 2006b) discusses web form design and gives design recommendations based on findings of gaze paths over different layouts. Bojko (2006) compares web page designs using heat map visualizations (more about heat maps below) of the recorded eye tracking data. The distributions of fixations illustrated by heat maps helped in deriving conclusions from design details. Nielsen (2007) reports on studies done by the Nielsen-Norman group where eye tracking has been used to detect the user's reading patterns on web pages to aid the page designers. For example, people generally spend more time on the top left area than other parts of the web page and they tend to ignore areas where advertisement banners are typically located (a phenomenon called "banner blindness").

However, the data provided by eye tracking itself is limited, if not augmented with a verbal explanation either by the test participant, or an expert of the task who can interpret the gaze path (Seagull & Xiao, 2001). As pointed out by Duchowski (2006), the traditional metrics given by eye tracking software are *low level measures* and their relationship with usability findings has not been established: how are fixations supposed to elucidate user satisfaction, effectiveness and efficiency, the common (ISO 9241) usability goals? From eye tracking data it is difficult to derive answers to questions such as "why" or "how".

It should be noted, however, that the traditional eye tracking metrics have been used in usability evaluation, too (for reviews, see Jacob & Karn, 2003; Poole & Ball, 2006; Ehmke & Wilson, 2007). For example, Nakamichi, Shoma, Sakai, and Matsumoto (2006) found that the moving distance and speed of the users' gazing points increase in web pages with low usability, suggesting that, in general, the user spends more time in searching for information (and user interface elements) than focusing on it. Furthermore, fixation duration has been found to have a strong link with cognitive processes, for example in reading studies (Rayner, 1998). More research on this is clearly needed in the context of usability studies.

A common way of analyzing web pages with eye tracking is to visually overlay on top of the page fixation maps (Wooding, 2002) or an "attentional landscape" (Velichkovsky & Hansen, 1996), most often called "*heat maps*" (Maughan et al. 2007; Nielsen, 2007), but also "*hot spots*" (Duchowski, 2006), and "*inverted density distributions*" (Schiessl, Duda, Thölke, & Fisher, 2003). A heat map is generated by an analysis tool after the session has ended, and indicates which parts of the page are most looked at. It can be based on the eye movements of one participant, or it might combine input from all of them.

Another solution is to define *Areas of Interest* (AOI) based on the parts of the web page, and then count how many of them are visited by the user. The participant's fixations should match the important AOIs. Johansen and Hansen (2006) point out that people in general have good memory of what AOIs they have looked at, but they do not necessarily remember the order in which they focused on each one, nor do they remember seeing a logo even though they had looked at it.

In some studies certain kind of gaze paths have been associated with potential usability problems. For instance, Duchowski (2006) presents a screen shot augmented with the participant's fixations on the screen, and the gaze path shows the participant visually searching for the "Edit" button in the wrong parts of the window. Duchowski concludes that the participant loses time based on this inefficient search caused by the non-standard placement of the button.

More specifically, Ehmke and Wilson (2007) found that usability problems are connected to certain sequences of eye movement patterns, resulting from the different coping strategies applied by users when they encounter a usability problem. For example, if the expected information is missing, it typically induces a gazing pattern with many short fixations across the page, whereas missing functionality causes a high number of fixations on a certain area followed by less spatially dense fixations across the page. However, the patterns of

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

eye gaze behavior differ across users, and while the thought processes might be linked to the eye gaze behavior, one user might not encounter the same usability problem as others.

While the eye-movement patterns may indicate potential usability problems for the experienced evaluator, a remaining challenge is that the gaze data are hard to interpret by the evaluator without discussions with the user. In spite of the emerging understanding of the relationship of eye-movement metrics and usability problems, we are far from any automatic linkages between eye-movement patterns and usability findings (Ehmke & Wilson, 2007).

Often the analysis based on the participant's gaze path is delayed to take place only after the participant has left the usability laboratory. Next we look at studies where the eye tracking method is used together with the more commonly applied think-aloud method.

Studies Combining Eye Tracking and Think-Aloud

Retrospective protocols based on memory recall only differ from CTA protocols both in the number of segments of verbalization produced and the content of the segments (Kuusela & Paul, 2000). However, when the production of the RTA protocol is supported with a memory aid, for instance, video playback with an overlaid gaze path, memory is not a limiting factor any more.

Overall, only few studies use both eye tracking data and verbal protocols. One of the early studies was conducted by Russo (1978) in the field of consumer psychology. Russo compared eye fixations with four alternatives to collect data from the participants: chronometric analyses, information display boards, input-output analyses and concurrent verbal protocols. Interestingly, the study suggested that "verbal protocols are remarkably complementary with eye fixations" (p.569). This was due to the differences between the methods; hence Russo (1978) argued that the disadvantages of one method are compensated by the strengths of the other.

Hansen (1991) compared the quality of retrospective verbalizations with and without showing an overlaid gaze path. Hansen found that the users were able to recall their thoughts during the task execution more precisely when a playback of the session was supplemented with the overlaid gaze path. In other words, Hansen (1991) compared two conditions of RTA; retrospective verbalization when the users were presented a plain playback video of the session and when the users were presented a gaze path augmented video playback. Hansen's study introduced a nice method for analyzing the user comments. However, since it compared the two RTA methods, it was not surprising that the additional information, the gaze path, helped the user to remember the steps and the thoughts during the task session.

Even the CTA protocol does not give information about the user's thoughts especially in situations where the user is puzzled. Cooke and Cuddihy (2005) used eye tracking to find additional information about the actions that are not verbalized by the user in CTA. Their method was to manually produce a written transcript describing the gaze paths together with the verbalizations and mouse movements of the participant, and use this transcript for further analysis after the session. They found that the gaze path transcript is useful: it gives hints about the hesitations and also expectations about where information should be located. Such information is not available in the CTA protocol or even in observational data.

However, it is not clear how many problems could be identified without the think-aloud, just observing the user's behavior. Ehmke and Wilson (2007) calculated the number of problems found through CTA, observation and RTA, but they do not indicate how many unique problems were found in each condition, only that each of the conditions produced a high number of findings, and many of the problems were duplicates that were found in many types of data.

Ball et al. (2006) compared CTA with two versions of RTA: cued with a playback with or without (the participant's own) eye movements. They found that using gaze paths as a memory cue in RTA is useful indeed. Using gaze paths in the retrospective verbalization revealed significantly more usability problems than the think-aloud technique during the test. Eger, Ball, Stevens and Dodd (2007) provide an extended report of the same study. They found that the gaze augmented RTA helped in identifying more usability problems than CTA or RTA with plain screen playback in certain cases. They found a strong interaction between a search engine (Infomagnet vs. Google) and the cue type (gaze-augmented vs. plain screen playback). This suggests that the gaze-augmented playback is especially useful in evaluating more complex (search) environments.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

In the following sections we describe our study on gaze path stimulated RTA. We were especially interested in the possible differences between the quality of the verbal data retrieved in CTA and gaze path stimulated RTA. We applied the methodology for categorizing verbalizations used by Hansen (1991) in his early study. The main distinction to his study is that he compared two different RTA conditions, while we compared the traditional CTA and the gaze path stimulated RTA to address the following questions:

1. Will the participants using gaze path stimulated RTA be able to produce as many or more useful comments than users using CTA?
2. Do we find fewer or more usability problems with gaze path stimulated RTA than with CTA?
3. How will the need for capturing the user's gaze path and collecting retrospective verbalizations affect planning and running usability tests?

Experimental Setup

We performed a traditional usability test of a Finnish car brokerage web site (Autotalli.com) with eight test participants. We gave the test participants eight tasks, varying from maintaining their own profile in the service to searching for cars with certain properties. Half of the participants were counseled to think aloud. The other half performed the assigned tasks without verbalizing their thoughts. The gaze paths of all participants were recorded and they were asked to give a retrospective verbalization of their recalled thoughts during the playback of the test session. The gaze path was overlaid on the video replay (see Figure 1).

Apparatus

The experiment was set up in a usability laboratory on a desktop computer with Windows XP. The participants used Internet Explorer 6 to carry out the tasks. They were required to use the mouse and keyboard during the experiment. Gaze data were recorded using the Tobii 1750 eye tracker with its hardware integrated within the casing of the participant's 17" LCD display. The sampling rate of the Tobii 1750 eye tracker is 50 Hz, and it needs to be calibrated for each user. Tobii's ClearView eye gaze analysis software was used to display the gaze path on the participant's display during the retrospective think-aloud session.

Another PC was used to capture a video of the session. The room setup was arranged so that the user's facial expressions could be captured with a video camera, and a microphone placed on the participant's desk was used to collect verbal data. The video and audio data of the whole session were captured with Noldus Observer 5.0 running on the experimenter's PC. The experimenter sat next to the participant and used pen and paper to write down notes during the test. No other test personnel or observers took part in the tests.

Participants

Eleven persons were recruited to take part in the experiment, however, three of them had to be omitted. One session was ended as problems occurred with data recording, and two of the sessions were ended due to frequent failures in calibration. This was probably due to eye glasses, although some test sessions with participants wearing glasses were successful.

The remaining eight participants were from 24 to 33 years, in average 30 years old. Three of them were male and five female. The real end user group of the Autotalli.com web site is dominated by male users, but due to the problems discussed above, three male participants that were originally recruited had to be left out of the study.

Two of the participants had used the Autotalli.com site before. Seven of the eight participants had a valid driving license, four owned a car, and three had a possibility to use somebody else's car on a daily basis. One participant had no car at all. Two of the participants were planning to buy a new car in the near future.

All participants rated their ability to use computers high, and all used Internet on a daily basis. It was mainly used for searching for information, reading the news, receiving and sending e-mails, and accessing electronic services, such as online personal banking.

Procedure

On entering the room, the laboratory and the equipment were introduced to each participant. They were asked to sign an informed consent form and fill in a background questionnaire on their Internet use. Then the experimenter explained the procedure briefly, and the participants were told that they had the right to quit the experiment at any time. Thinking aloud was explained to the participants in the think-aloud condition and they were allowed to practice it briefly. Participants were al-

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

lowed to ask any questions related to the experiment before starting the tasks.

After the introduction, participants were asked to sit down in front of the display equipped with the eye tracker, about 70 centimeters away from the screen. They were advised to sit as still as possible during the test but their movements were not restricted in any way. After the instructions, the Tobii 1750 eye tracker was calibrated. Before starting the actual test, participants in the think-aloud condition were reminded about verbalizing their thoughts. Those who were not asked to think aloud were only asked to tell aloud when they had finished working on a task. However, they were allowed to speak if they wanted to.

The test started with the experimenter reading out the first task. The tasks were given one at a time. The tasks were a collection of simple fact finding search tasks, comparison of the search results, and logging into the site to update user information (see the section *Qualitative Analysis* for a listing of the tasks and their execution times). Some of the tasks had simple numeric answers while others just asked the participant to find a certain page within the site. All tasks could be completed within the site. When the task involved using the keyboard, the experimenter gave a post-it note with the correct words (e.g., user password) in writing. This was done to ease the recall of the needed words. The note was placed on the top right bevel of the display to minimize off screen glances which may worsen the eye tracker's calibration.

The experimenter used a timer to measure the time used for each task, and took notes of the observed problems and actions. Some participants forgot to think aloud during the test, and in these cases the experimenter prompted them to verbalize their thoughts. After all the tasks were completed the Tobii 1750 eye tracker was stopped.

In the second part of the study, all participants were asked to view their gaze paths with the experimenter. No audio was available on the playback. While watching the recording the participants were asked to think aloud or comment what they were thinking about during the test. The first two seconds of the recording were shown first

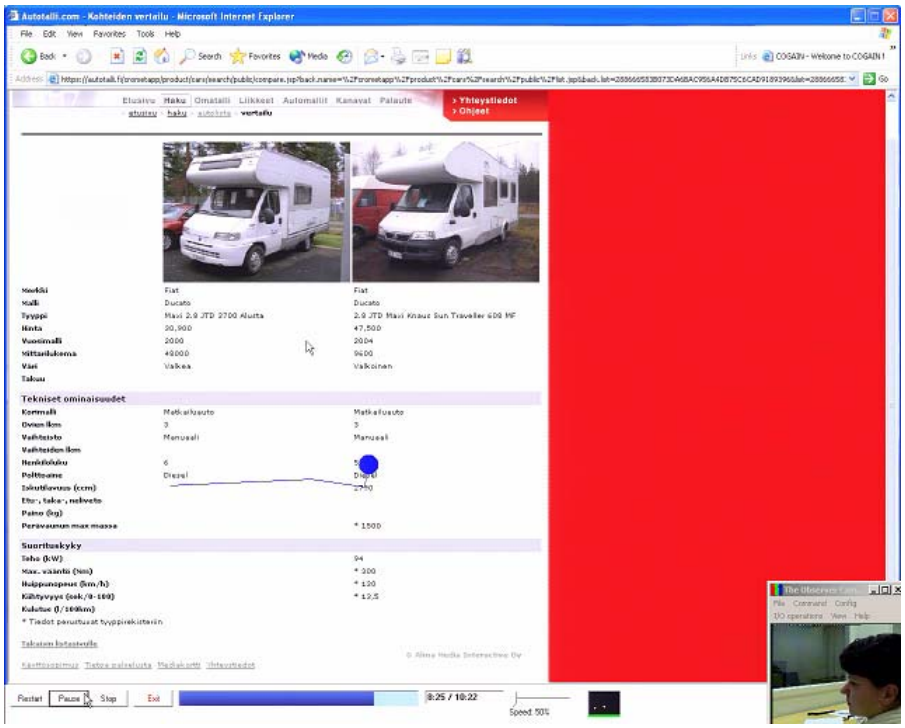

Figure 1. *A screen capture of the recorded video.*

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Before continuing the viewing, the experimenter stopped the replay and explained gaze paths and fixations briefly to the participant. Eye movements are so fast that the recording needs to be shown in half speed, and two seconds of the recording already included several fixations.

The web site under evaluation has fixed width pages that fill only partially the maximized window in Figure 1. The participant is shown in the small insert only in this still image captured from the video recording of the situation. The participant did not see the video image but only a recording of the screen together with the gaze path. The controls at the bottom of the screenshot are from the ClearView program used for viewing the recordings.

The participants were prompted to think aloud if they fell silent instead of verbalizing their thoughts. The experimenter wrote notes during the retrospective think-aloud. Also this part was captured on video.

At the end of each session, the participants were interviewed and asked to fill in a questionnaire. The participant was thanked for his or her participation and the main aim of the study was explained.

Design

A qualitative analysis of participants' verbalizations was done to learn about the usability problems found in the test. The findings were collected from the think-aloud protocols and categorized into problems found with CTA and RTA.

The verbal protocols produced by the participants were also analyzed quantitatively. They were first coded according to the coding scheme of verbalizations suggested by Hansen (1991): manipulative, visual, and cognitive comments. More elaborate categorizations appear in the literature; see, for instance, Guan et al. (2006) and Van Gog et al. (2005). However, the simpler categorization points out the differences more clearly.

The quantity and quality of words produced by the participants in different conditions were analyzed using a 2 x 2 x 3 analysis of variance (ANOVA). The factors were CTA, Stage, and Comment Category. CTA was a between-subjects factor with two levels (with and without concurrent thinking aloud). The second factor, Stage, was a within-subjects factor with two levels (concurrent verbalizations and retrospective verbalizations). The last factor, Comment Category, was a within-subjects factor with three levels of comments (manipulative, visual and cognitive); the categories are explained in the *Quantitative Evaluation* section.

Results

We were interested in studying how the think-aloud method affects the participants' observations in a usability test. By presenting the playback with an overlaid gaze path, we hoped to get the same information from the user that we would get with the concurrent think-aloud method. We first present the results from the usability evaluation point of view, i.e., the number of and quality of usability problems found in different conditions. We then analyze in detail the verbalizations of the participants to get insight into the kinds of comments made in each case.

Qualitative Analysis

The usability test sessions where the participants worked on the test tasks took from roughly 9 minutes to nearly 16 minutes (Table I). In general, the participants in the concurrent think-aloud condition did not spend more time on the tasks. As a whole, no large differences in task times were found between the conditions. In tasks 1, 4 and 7 a post-it note with the data required was attached to the casing of the monitor to ease recall.

We analyzed the test findings to produce a list of the usability problems users experienced in the site. Most parts with which the participants had problems were brought up in CTA as well as in RTA. Some examples of usability problems found can be seen on the main search window (Figure 2) of the site. Many participants experienced various usability problems on this page.

In Figure 2, the user is currently visually scanning the page (Task 3) and wondering how to restrict the search to cover only advertisements with pictures. The gaze path is currently moving close to the correct check box on the right, which was very hard to find. Many users explained in RTA that they actually did not notice the check box, though according to their scan paths they seemingly fixated on it. This is an example of a usability problem found with the help of the gaze path and the use of RTA.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Table I
Test tasks and their duration in seconds per participant (P1-P8).

	Task durations during the test in seconds							
	Think aloud				No concurrent think-aloud			
Tasks	**P2**	**P3**	**P4**	**P6**	**P1**	**P5**	**P7**	**P8**
0. Go to the page Autotalli.com	10	25	30	20	25	15	15	20
1. Log in to the site. The test account is Auto Talli, password asuntoauto.	80	64	130	100	60	70	240	30
2. How many Audi A3-cars are for sale in Pirkanmaa?	20	50	20	40	30	25	40	45
3. Restrict the search to show only ads with a picture.	30	130	50	65	30	40	35	30
4. Change your password.	35	60	60	60	60	55	40	40
5. Find motor caravan cars with power steering. Read aloud the price of the cheapest of them.	50	120	300	120	110	70	70	120
6. Compare two newest of the motor caravan cars.	40	40	100	60	15	30	20	15
7. Find the details corresponding to this newspaper advertisement number.	25	80	30	60	55	75	25	120
8. Log out of the site.	10	50	30	10	10	30	10	25
9. Tell your comments about the main page.	150	120	90	70	40	15	50	80
Whole duration of working on the tasks (min:sec)	10:00	14:25	15:40	12:05	10:50	9:10	12.00	10:40
Total length of session (test+retrospection, min:sec)	25:00	45:40	47:30	36:30	31:20	28:00	36:20	31:50

For instance, after entering the other search criteria P4 spent nearly 10 seconds just looking at the page and even opened the list box next to the check box, but she did not talk at all while searching for the functionality.

RTA revealed design flaws in the buttons at the bottom of the page in Figure 2, too. The buttons do not appear active since they are colored gray, and therefore the participants did not find the functionality. The second button is “Clear the fields”. A participant explained: “When I did not find a button for emptying the fields, I just went to the main page – perhaps the fields would be cleared that way.” [P3, RTA] The participants also explained their conceptual models of these buttons quite clearly: “Yeah, the [first] button is red and everything, but it is like one of the steps telling me where I am going.” [P1, RTA] Without the think-aloud, the reason for navigating between pages might not have been recognized as a usability problem relating to the buttons.

The analysis of usability problems (Lehtinen, 2007) found in the various conditions revealed that most of the problems could be observed both in the CTA and RTA conditions. Of the 44 usability problems encountered by the participants, 31 problems were present in both retrospective and concurrent think-aloud protocols. The participants talked about the problems they encountered during RTA, since only two problems that could be observed in the videotaped sessions were not talked about by the participants in the RTA protocol. However, the RTA condition revealed eleven new problems that could not be found in the observation of user behavior and the CTA protocol.

The participants also commented on their actions differently in the two conditions. When in the test the users had problems in finishing a task, they talked about not knowing what to do, but in the retrospection they were able to analyze why it was difficult and even suggest improvement ideas.

When the users saw their gaze paths displayed, they were enthusiastic and motivated to tell about their thoughts and reasons for looking at various parts of the screen. There was more time to talk, since the playback was at half speed, which enabled also deeper analysis of the site in general. Sometimes the discussion covered also their experiences in related sites with which they compared the site under evaluation. In general, there was much more talk in the playback session than in the task execution stage.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

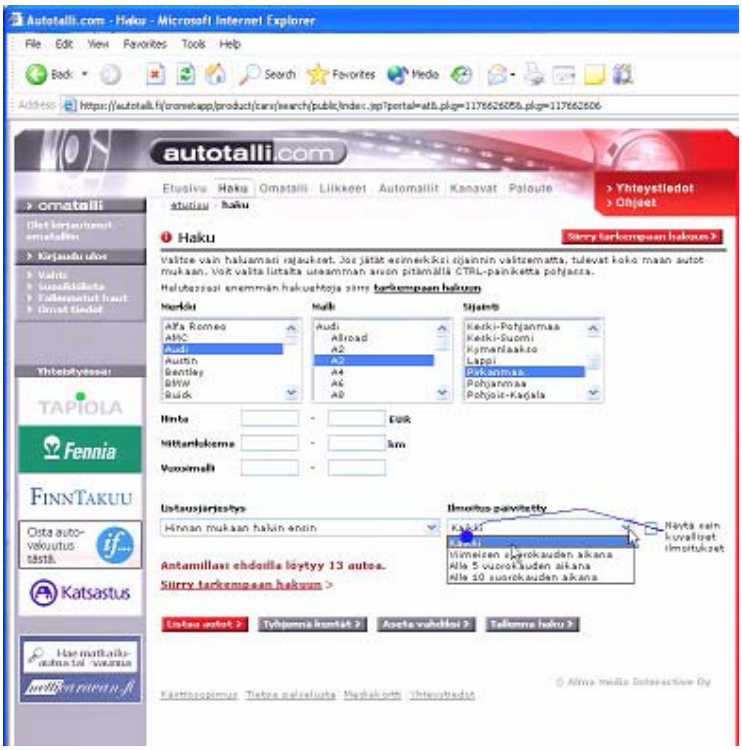

Figure. 2. *The main search page with an overlay of the gaze path indicating a problematic area.*

In addition to actual usability problems, RTA was good at revealing expectations of web design. The participants explained their earlier experiences while watching the gaze path. For instance, mental models about where to find the logout functionality differed based on their earlier experiences: some looked for the button in the top right corner "where it commonly is" while some others found it easily from the left side menu based on their earlier experiences with a banking site ("Kirjaudu ulos", the second dark grey button from the top in the left menu in Figure 2). Such explanations would hardly emerge during concurrent think aloud.

The participants found the gaze path easy enough to interpret but noted that their gaze shows lack of concentration on the tasks since it is bouncing around so fast. Some of them also expressed amusement about their prolonged searches for some functionality: "Eagle eye is asleep now". Thus, while the retrospective sessions produced more talk, not all of the talk was focused on the actual explanations of behavior.

Quantitative Evaluation

A preliminary analysis of the data was presented by Lehtinen, Hyrskykari, and Räihä (2006). We first computed the word count of the verbal data recorded in each of the conditions. Then the operational comments (Hansen, 1991), i.e., the participants' verbal expressions on behavior or operations, were extracted from the data. After that we used the coding presented by Hansen to compare the types of think-aloud data in the three different conditions. Hansen categorized the comments to manipulative, visual, and cognitive operations.

Manipulative operations are the ones expressing performance. Some examples of manipulative operations in our data are

"I *write* the name into this field",

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

.

Table II
Number and percentages of participant comments in the conditions

	Words	Comments	Comment Categories		
			Manipulative	Visual	Cognitive
Concurrent think aloud (CTA)	1148	66	82%	14%	4%
RTA after first doing CTA	3309	214	53%	14%	33%
RTA without preceding CTA	4136	267	42%	15%	43%

"Of course, I could have *clicked* all of those…", or
"Oh, I *gave* an erroneous input…".

Visual operations reflect perceptual activity, like

"I *saw* it here somewhere…",
"Then I *look* for a picture of the car…", or
"I *read* it from the previous page …".

Cognitive operations reflect interpretations, evaluations, expectations and specifications of action, like

"I *remember* seeing it before",
"Now I finally *understand* that there is a scroll bar on the right", or
"I *found* out that I can't make a search from this page".

Many of the phrases included several verbs, falling into different categories, for instance: "then I *went* back to the account page, and *saw* the right button there…" The sentence was categorized into both manipulative and visual operation comments. The results of the analysis are shown in Table II.

Retrospective think-aloud supported with gaze path playback produced significantly more verbal data than the original think-aloud method. A significant main effect ($F(1,6) = 29.6$, $p < .0001$) of Stage (concurrent vs. retrospective verbalizations) on the average number of words was observed, with participants producing significantly more words in total during the retrospective think-aloud stage (Lehtinen, 2007).

Moreover, 82% of the comments made in CTA during the testing were manipulative comments, i.e. the participants were commenting what they were doing at the time, whereas the share of cognitive comments in that group was only 4%. In RTA the percentages were almost equal: 42% of manipulative comments vs. 43% cognitive comments. In each condition, the share of visual comments was almost the same between the groups.

The results in Table II suggest that participants did verbalize their actions more in retrospection than in concurrent think-aloud. For statistical analysis of the data, the number of comments in each condition is summarized in Table III. (Without CTA means that the participant was not instructed to think-aloud concurrently; nevertheless, they may have expressed some comments, as shown in the table.)

We then compared the type of the comments between RTA with gaze paths and CTA. A significant difference was found between the operational Comment Categories (i.e., manipulative, visual, and cognitive comments), $F(2,6) = 20.2$, $p < .0001$, suggesting that the mean number of comments did vary significantly over comment categorizations. Participants made significantly more manipulative comments than visual comments and significantly more cognitive comments than manipulative comments. Moreover, there was a significant interaction effect between Stage and Category ($F(2,6) = 14.1$, $p < .0001$). In RTA the participants produced significantly more comments in every operational comment category: manipulative, visual, and cognitive, than in the concurrent condition. On the other hand, no effect was found for the CTA condition, i.e. whether the participants carried out concurrent think-aloud or not did not significantly affect the number of comments in the different categories.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Table III
Means and standard deviations for number of comments produced in different conditions and stages.

			Mean	St. Deviation
CTA	Manipulative comments produced concurrently	CTA	13.5	8.9
		Without CTA	1.0	1.4
	Visual comments produced concurrently	CTA	2.3	1.0
		Without CTA	0.5	1.0
	Cognitive comments produced concurrently	CTA	1.0	2.0
		Without CTA	0.3	0.5
RTA	Manipulative comments produced retrospectively	CTA	28.0	10.5
		Without CTA	29.8	14.1
	Visual comments produced retrospectively	CTA	7.5	6.6
		Without CTA	10.0	4.1
	Cognitive comments produced retrospectively	CTA	18.0	7.0
		Without CTA	28.5	17.3

Discussion

The experiment we performed brought us a fair deal of experiences on gaze path stimulated RTA. The experiences vary from very practical observations on issues of designing such a usability test to observations on problems and advantages of the method.

Participants

Eye tracking itself poses several new demands for usability testing. Already Aaltonen (1999) noted the need to recruit extra participants just in case the calibrations do not work out for all of them. The generations of trackers have substantially improved since 1999 in this respect, as noted by Duchowski (2006). Nevertheless, we lost the data from nearly one third of our participants, even though the tests were run in 2005 with eye tracker hardware purchased in 2004. Furthermore, the test requires an experienced moderator to overcome the initial obstacles of eye glasses or wrong viewing angle.

Mobility Restrictions

Even with the state-of-the-art trackers it still is questionable if the exact calibration of the tracker endures lengthy periods of time of looking aside the screen and potential changes in the head or body posture. Thus, the tasks of writing with the keyboard, reading paper material not on screen, or even looking at the test moderator should be minimized when using eye tracking. The participant may feel extra tension if this requirement of staying still is emphasized before the test. However, it might be necessary since the device does not warn audibly if calibration is lost or if the user has moved so that the tracker's camera no longer sees the eyes.

Duration of the Test

In addition to reserving time for the calibration(s), one should of course reserve the additional time for retrospective walk through. The time needed is most likely at least twice or three times the length of the original task execution, since the gaze movements are so fast that they cannot be commented on in real time. For some participants, this certainly limits the duration of the test. Time is a precious resource for the usability practitioner, too, as emphasized by Denning et al. (1990). Getting a retrospective report of the actions in the test not only lengthens the time spent with each participant but also makes it harder to engage designers as observers in the laboratory.

When collecting the retrospective comments, Russo (1979) suggested that the replay speed could be controlled by the test participant; it might be useful to slow it down by factors of two, four or even eight. In our tests, the replay speed was set at start and not changed during the session. Some retrospective reports have been collected at faster than original speed, for instance, Norman and Murphy (2004) fast forwarded the video playback and stopped only at those events the participant wanted to explain. We found that the gaze path overlay would be

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

too distracting at such a speed, and the faster pace would diminish the opportunity of talk-aloud.

Many of our participants were surprised to see their eye movements, which may have affected the large number of cognitive comments. The users seemed to explain their actions to themselves as well as to the experimenter. Although the gaze paths needed to be viewed in half speed (in order for the users to be able to comment on their actions), the retrospective verbalization took only about 20 minutes. Our users were able to concentrate throughout the whole test session, but we do not have data on any longer tests.

We could considerably shorten the time during the test, if the test leader could mark the points where the user seems to have problems in performing the task and then run only those gaze paths in the RTA session. This demands a tool for marking and replaying only the problematic points efficiently.

Instructions for the Participants

The main advantage of moving into retrospective reporting is that during the test no think-aloud protocol is needed. This enables the test participant to concentrate on the actual test task. We noted, however, that for some of our participants it was not clear if they were also allowed to explicate their thoughts during the test. Clearly these new methods require a thorough rephrasing of the test instructions.

As we have pointed out based on previous literature, CTA is reactive, and it also changes the eye gaze behavior making it harder to interpret the actual thought processes. Russo et al. (1989) think that even collecting the RTA data after the primary task may affect the primary task, if the participant knows in advance that an RTA protocol requiring memorization or explanations is to be produced. That is why the test leader should try to make the participant feel relaxed and ease the possible prejudice against eye tracking.

Role of the Auditory Data

When planning the tests, we chose not to play the audio captured during the test task execution—we thought it would disturb the retrospective verbalization (while in some cases it might have helped the participant to recall the thoughts). When running the retrospective think-aloud sessions, we had problems in keeping in pace with the actual test tasks. They were read aloud during the test and did not appear on the screen in any way; without the verbalization, we were not sure of when they actually were presented to the participant. We need to use a timer during the test session to enable prompting when a new test task is moved to in the test, or we need to show the test tasks on screen. A tool that allows the experimenter to set (different types of) markers in the data during the test would be useful in marking the start and the end of each task.

Quality of the Data

When testing usability each operational comment made by the participants is valuable, but especially the cognitive comments give information that usually cannot be deduced from the user's observed interaction with the product.

Guan et al. (2006) note that the eye tracking data is not similar to the comments made during RTA: omissions (in their data nearly 44%) indicate spots where the participant's gaze path shows a fixation in an area of interest but the participant does not tell about that in RTA. When the participant has worked on difficult tasks such omissions occur more often than with simple tasks. Contrary to our method, Guan et al. did not show the gaze path to the users but they reported their RTA based on a video captured of the screen showing only their mouse pointer movements. In our RTA, the participants could see their fixations and saccades (and also mouse movements) in the replay of the recording which may have helped in avoiding such omissions.

Capra (2002) suggests that participants themselves could produce written reports of critical incidents that they encounter during a test session either contemporaneously within each task, or retrospectively after the session. Of course, giving written reports cannot be compared with think-aloud; the writing task adds a burden to the participant and takes even more time from the participants in the lab. However, Capra learned that the participants wrote about things that might be unavailable for an observer—which complies with our experiences on RTA. Capra encouraged the users to think aloud so that they would better remember the incidents afterwards when watching the recording. We did not find this a problem, but our sessions were shorter than Capra's, and cued by the gaze path replay.

Our study points out (in Tables II and III) that in the concurrent think-aloud condition the relative amount of

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

cognitive comments is considerably low, while in the retrospection their amount is clearly higher. Thus, we claim that gaze stimulated RTA gives better quality data for the usability analysis.

Special User Groups

For some user groups retrospective verbalizations might work out better than CTA. For instance, in their work with older adults, Dickinson, Arnott, and Prior (2007) point out that for some members of this user group thinking aloud during the usability test is challenging, but RTA does not always succeed either. The participants had processing and memory difficulties and could not recall the steps they had just performed. However, Dickinson et al. applied RTA successfully in another eye tracking experiment to collect the older beginners' initial understandings of web pages. The RTA protocol was collected after the participant had been looking at a web page silently for 20 seconds. Dickinson et al. note, though, that the second exposure to the web page contributes to learning, therefore potentially confounding the experimental results.

Culture may also have an effect. To think aloud during tasks may be more difficult for easterners than westerners due to the cultural differences. Kim (2002) had westerners and easterners solve reasoning problems while they were thinking aloud, and found that talking impaired Asian Americans' performance because they tend to use internal speech less than European Americans. Evers (2002) found that verbalization is easier for North Americans than for Japanese users who feel uncomfortable verbalizing their thoughts. We do not know if RTA would be any easier than CTA for them, but at least the results from the task execution would be more reliable without the unreasonable requirement to think aloud during the test. It has also been found that cultural background affects non-verbal behavior, such as gestures, in a usability test situation (Yammiyavar, Clemmensen, & Kumar, 2008). It would be interesting to study whether similar differences can be observed in gaze behavior.

Veridicality

It can happen in the retrospection that the participant considers some usability problems encountered during the test trivial or just forgets to mention about them, or on the contrary, makes them even more difficult than they were. Though these affect the veridicality of the protocol, it is not catastrophic from the usability evaluation point of view, since the test participant is helping in finding the problematic areas of the interface—the focus is on the findings, not on the performance of the participant.

Conclusions

Analyzing eye tracking data using statistical methods is laborious. Using heat maps to get an overview of the data loses information of the gaze path. Using eye tracking to help the participants to recall the task session is a technique that maintains full information of the gaze path, and can help in producing information that cannot be obtained by traditional techniques alone.

Already the increased amount of data received from gaze-stimulated RTA suggests that it works better than conventional think-aloud. However, it is even more noteworthy that the original task session can be performed in a more natural way without interruptions, which, presumably, corresponds to more natural behavior than the data received with the traditional CTA method.

Our observations justify us to expect that when using gaze path playback, RTA may produce more and better quality data than CTA. Gaze paths did offer additional information on users' behavior. In retrospection, users were able to see their eye movements while working on the test tasks, and it did raise comments on what they had been looking for or what they were trying to find at the moment. Three out of four users found the retrospective verbalization more pleasant than concurrent thinking aloud. All the users reported that it felt easy to follow the gaze paths, and eye movements did offer an excellent aid to recall their thoughts afterwards. Hence, some users were quite enthusiastic to see where they had been looking at in each task.

There may be several reasons behind the promising results. It was noted that the users were clearly more relaxed in the RTA condition than in the CTA condition. Probably the users felt that they were no longer under observation as they were during the usability test. Moreover, during RTA they were allowed to move more freely than during the initial eye tracked session; at that time they were advised to avoid unnecessary movements. As the results show, users felt at ease to comment on their action, to make interpretations or judgments, and to explain their behavior when commenting their (gaze) behavior retrospectively. They also were more likely to pro-

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

vide improvement suggestions for the web site. However, our results are only preliminary due to the small sample size. For instance, the increase in verbalizations might be due to more talkative participants in the RTA condition, and not caused by the retrospection itself. As Taylor and Dionne (2000) point out, the data collected in a retrospective report are influenced by many things, for instance the instructions given in eliciting the report, the nature of the task, the experience level of the participant, and the specific questions and probes given by the experimenter. Instructions and probes in RTA need further research.

Acknowledgements

We would like to thank the reviewers, Linden Ball and Ben Tatler, for their comments, which resulted in significant improvements to this article.

References

Aaltonen, A. (1999). Eye tracking in usability testing: Is it worthwhile? 'Usability & eye tracking' workshop at CHI'99, ACM Press.

Ball, L.J., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the PEEP method in usability testing. *Interfaces 67*, Summer 2006, 15–19. Retrieved September 24, 2008, from http://www.psych.lancs.ac.uk/people/uploads/LindenBall20070323T100155.pdf

Bojko, A. (2006). Using eye tracking to compare web page designs: A case study. *Journal of Usability Studies 1*(3), 112-120. Retrieved September 24, 2008, from http://www.upassoc.org/upa_publications/jus/2006_may/bojko_eye_tracking.pdf

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

Bowers, V.A. & Snyder, H.L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors Society 34th Annual Meeting*, 1270–1274.

Branch, J.L. (2000) Investigating the information-seeking processes of adolescents: the value of using think alouds and think afters. *Library & Information Science Research 22*(4), 371–392.

Branch, J.L. (2001) Junior high students and Think Alouds: Generating information-seeking process data using concurrent verbal protocols. *Library & Information Science Research 23*(2), 107–122.

Capra, M.G. (2002). Contemporaneous versus retrospective user-reported critical incidents in usability evaluation. *In Proceedings of the Human Factors Society, 46 th Annual Meeting*, 1973–1977. Retrieved September 24, 2008, from http://www.thecapras.org/mcapra/work/Capra_HFES2002_UserReportedCIs.pdf

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

Denning, S., Hoiem, D., Simpson, M., & Sullivan, K. (1990). The value of thinking aloud protocols in industry: A case study at Microsoft Corporation. In *Proceedings of the Human Factors Society 34th Annual Meeting, 1990*, 1285–1289.

Dickinson, A., Arnott, J., & Prior, S. (2007). Methods for human-computer interaction research with older people. *Behaviour & Information Technology 26*(4), 343–352.

Duchowski, A.T. (2006). High-level eye movement metrics in the usability context. Position paper, *CHI2006 Workshop Getting a Measure of Satisfaction from Eyetracking in Practice*, April 23, 2006. Retrieved September 24, 2008, from http://www.amberlight.co.uk/CHI2006/pos_papr_duchowski.pdf

Ebling, M. R. & John, B.E. (2000). On the contributions of different empirical data in usability testing. In *Proceedings of Symposium on Designing Interactive Systems*, DIS'00, 289–296.

Eger, N., Ball, L.J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. In L.J. Ball, M.A. Sasse, C. Sas, T.C. Ormerod, A. Dix, P. Bagnall, & T. McEwan (Eds.), *People and Computers XXI - HCI... but not as we know it: Proceedings of HCI 2007.* Swindon: The British Computer Society.

Ehmke, C. & Wilson, S. (2007) Identifying web usability problems from eye-tracking data. In L.J. Ball, M.A. Sasse, C. Sas, T.C. Ormerod, A. Dix, P. Bagnall, & T. McEwan (Eds.), *People and Computers XXI - HCI... but not as we know it: Proceedings of HCI 2007.* Swindon: The British Computer Society.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Ericsson, K.A. & Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data.* Revised edition. Cambridge, MA: The MIT Press.

Evers, V. (2002). Cross-cultural applicability of user evaluation methods: A case study amongst Japanese, North-American, English and Dutch users. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. New York: ACM Press, 740–741.

Goldberg, J.H. & Kotval, X.P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics 24*, 631–645.

Goldberg, J.H., Stimson, M.J., Lewnstein, M., Scott, N., & Wichansky, A.M. (2002). Eye tracking in web search tasks: Design implications. In *Proceedings of Symposium on Eye Tracking Research & Applications (ETRA 2002).* ACM Press, 51–58.

Goldberg, J.H. & Wichansky, A.M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In R. Radach, J. Hyönä, & H. Deubel. (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research.* Amsterdam: Elsevier Science, 493–516.

Guan, Z., Lee, 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 (CHI '06).* ACM Press, 1253–1262.

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

Jacob, R.J.K. & Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In R. Radach, J. Hyönä, & H. Deubel. (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research.* Oxford: Elsevier Science, 573–605.

Johansen, S.A. & Hansen, J.P. (2006). Do we need eye trackers to tell where people look? In *Extended Abstracts of CHI 2006 (Work in Progress),* ACM Press, 923–926.

Kim, H. S. (2002). We talk, therefore we think? A cultural analysis of the effect of talking on thinking. *Journal of Personality and Social Psychology 83*(4), 828–842.

Kuusela, H. & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology 113*(3), 387–404.

Lehtinen, M. (2007). A gaze path cued retrospective thinking aloud technique in usability testing. M.Sc. Thesis, University of Tampere, Interactive Technology. June 2007. Retrieved September 24, 2008, from http://tutkielmat.uta.fi/tutkielma.phtml?id=17069

Lehtinen, M., Hyrskykari, A., & Räihä, K.-J. (2006). Gaze path playback supporting retrospective think-aloud in usability tests. In *Proceedings of the 2nd Conference on Communication by Gaze Interaction – COGAIN 2006, Turin, Italy*, 88–91. Retrieved September 24, 2008, from http://www.cogain.org/cogain2006/COGAIN2006_Proceedings.pdf

Lin, T. & Imamiya, A. 2006. Evaluating usability based on multimodal information: an empirical study. In *Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI '06)*. New York: ACM Press, 364–371.

Maughan, L., Dodd, J., & Walters, R. (2007). Video replay as a cue in retrospective protocol …Don't make me think aloud! Poster presented at the *Second Scandinavian Workshop on Applied Eye-Tracking (SWAET 2007)*, Lund.

Nakamichi, N., Shoma, K., Sakai, M., & Matsumoto, K. (2006). Detecting low usability web pages using quantitative data of users' behavior. In *Proceedings of the 28th International Conference on Software Engineering (ICSE'06)*, ACM Press, 569–576.

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

Nielsen, J. (2007). Eyetracking research. Retrieved September 24, 2008, from http://www.useit.com/eyetracking/

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads? Reflections on the think-aloud technique. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction (NordiCHI 2002),* 101–110.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Norman, K.L. & Murphy, E. (2004). Usability testing of an Internet form for the 2004 overseas enumeration test: A comparison of think-aloud and retrospective reports. In *Proceedings of the Human Factors Society 48th Annual Meeting, Human Factors Society*, New Orleans, LA. Retrieved September 24, 2008, from http://lap.umd.edu/lap/Papers/Tech_Reports/LAP2004TR04/LAP2004TR04.pdf

Penzo, M. (2006a). Evaluating the usability of search forms using eyetracking: A practical approach. *UX-Matters* January 23, 2006. Retrieved September 24, 2008 from http://www.uxmatters.com/MT/archives/000068.php

Penzo, M. (2006b). Label placement in forms. *UXMatters* July 12, 2006. Retrieved September 24, 2008, from http://www.uxmatters.com/MT/archives/000107.php

Poole, A. & Ball, L.J. (2006). Eye tracking in human-computer interaction and usability research: Current status and future prospects. In Ghaoui, C. (Ed.). *Encyclopedia of Human Computer Interaction*. Idea Group Publishing. Retrieved September 24, 2008 from http://www.alexpoole.info/academic/Poole&Ball%20EyeTracking.pdf

Pretorius, M.C., Calitz, A.P., & Van Greunen, D. (2005). The added value of eye tracking in the usability evaluation of a network management tool. In *Proceedings of the 2005 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries.* ACM International Conference Proceeding Series, vol. 150, 1–10.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin 124*(3), 372–422.

Rhenius, D. & Deffner, G. (1990). Evaluation of concurrent thinking aloud using eye-tracking data. In *Proceedings of the Human Factors Society 34th Annual Meeting*, Human Factors Society, 1265–1269.

Russo, J.E. (1978). Eye fixations can save the world: A critical evaluation and a comparison between eye fixation and other information processing methodologies. In H.K. Hunt (Ed.), *Advances in Consumer Research, 5*, 561-570. Ann Arbor, Michigan: Association for Consumer Research. Retrieved September 24, 2008, http://forum.johnson.cornell.edu/faculty/russo/Eye%20Fixations%20Can%20Save%20the%20World.pdf

Russo, J.E. (1979). A software system for the collection of retrospective protocols prompted by eye fixations. *Behavior Research Methods & Instrumentation 11*(2), 177–179.

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

Schiessl, M., Duda, S., Thölke, A., & Fisher, R. (2003). Eye tracking and its application in usability and media research. *MMI-Interaktiv 6*, 41–50.

Seagull, F.J. & Xiao, Y. (2001). Using eye tracking video data to augment knowledge elicitation in cognitive task analysis. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*. Retrieved September 24, 2008, from http://hfrp.umm.edu/alarms/12.%20Seagull_Xiao_CTA_HFES%202001%20Camera%20Ready%20v2.pdf

Taylor, K.L. & Dionne, J.P. (2000). Accessing problem solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology 92*, 413–425.

Van den Haak, M., De Jong, M.D.T., & Schellens, P.J. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology 22*(5), 339–351.

Van Gog, T., Paas, F., van Merriënboer, J.J.G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied 11*(4), 237–244.

Van Someren, M., Barnard, Y., & Sandberg, J. (1994). *The Think Aloud Method. A Practical Guide Modelling Cognitive Processes.* London: Academic Press.

Velichkovsky, B.M. & Hansen, J.P. (1996). New technological windows into mind: There is more in eyes and brains for human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'96)*, New York: ACM Press, 496–503.

Wilson, T.D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science 5*(5), 249–252.

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692

Wooding, D. (2002). Fixation maps: Quantifying eye-movement traces. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA 2002)*, New York: ACM Press, 31–36.

Yammiyavar, P.G., Clemmensen, T., & Kumar, J. (2008). Influence of cultural background on non-verbal communication in a usability testing situation. *International Journal of Design 2*(2).

DOI 10.16910/jemr.2.4.5

ISSN 1995-8692