A gaze path cued retrospective thinking aloud technique in usability testing

Merja Lehtinen

University of Tampere
Department of Computer Sciences
Interactive Technology
Master's Thesis
Supervisor: Aulikki Hyrskykari
May 2007

University of Tampere
Department of Computer Sciences
Interactive Technology
Merja Lehtinen: A Gaze path cued retrospective thinking aloud technique in usability testing
Master's thesis, 53 pages, 4 Appendices
May 2007

The present study explores the differences and similarities of concurrent thinking aloud and gaze path cued retrospective thinking aloud techniques in usability testing. Eight users were asked to complete a set of tasks (usability testing) individually, whilst their eye tracking data was collected with a Tobii eye tracker. During the testing, four users were asked to verbalise their thoughts (thinking aloud) when they were carrying out the tasks and four users were not given instructions to do so. After completing the tasks, users were shown their individual gaze paths recorded during the test session and they were asked to verbalise their thoughts (retrospective thinking aloud) while they watched the replay. After the retrospective viewing, users were interviewed briefly. Firstly, the usability problems observed in the present study were observed and reported. The results were compared to an independent usability test conducted by students of a usability evaluation methods course. Secondly, operational comments produced by each user in concurrent and retrospective thinking aloud conditions were recorded and analysed. It was noted that the gaze path cued retrospective thinking aloud technique was no superior over traditional usability testing using concurrent verbalisation in terms of quality of the usability problems obtained. The second part of the analysis revealed that users did produce significantly more operational comments retrospectively than concurrently. The results suggest that the gaze path cued retrospective thinking aloud condition could be used in usability testing to reveal usability problems.

Keywords: Human-computer interaction, Usability testing, Thinking aloud, Retrospective thinking aloud, Eye tracking.

Index

1. Introduction

One of the most commonly used user testing method is usability testing with the thinking aloud technique. Usability testing has become popular, as it can give relatively objective information of the design flaws on the product. Usability testing involves users into the design process, allowing the developers to test the product in realistic situations. Usability testing gives information of the problem spots on a product and the time users spend completing the tasks. Usability testing does not, however, give answers to the questions such as why users are behaving the way they do.

In addition to get further information, typically users have been asked to verbalise their thoughts (in other words to think aloud) during the test session. Thinking aloud technique is a cost effective way to obtain verbal data, and it is argued to reflect users' concurrent thoughts (Rhenius and Deffner, 1990). Although thinking aloud is a way to gain insight into users' thought processes, it is not a technique without problems. Among its biggest flaws are the inconsistent practises to carry out the technique in usability tests (Boren and Ramey, 2000). Boren and Ramey pointed out that think aloud practises vary widely among practitioners, affecting the validity of studies using the technique.

However, there are other problems as well. For some users concurrent verbalisation may feel uncomfortable or unnatural (Nielsen, 1993). It may also affect users' performance, as overburdened cognitive processes slow down the task performance time (van Someren et al., 1994; Rhenius and Deffner, 1990). On the other hand, thinking aloud may have an opposite effect on task performance. Concurrent verbalisation may reveal inconsistent thoughts the user has, making the task easier for the user, and therefore decreasing the task time. Users may also perform differently as they would have without verbalisation (Nielsen, 1993).

Retrospective thinking aloud is a technique in which the users are asked to verbalise their thoughts after performing the tasks. By asking the users to think aloud retrospectively, many flaws of the concurrent verbalisation can be avoided. Typically, a video recorded task performance has been shown to the users to help them to memorise what they had been thinking and to recollect the interaction difficulties during the testing. Although this may be the most

used retrospective thinking aloud scenario, video recordings are not the only memory aid that can be used. Users can be shown their own gaze paths recorded during the testing. Hansen (1991) argued that a gaze path cued retrospective thinking aloud method is as valid a method as retrospective thinking aloud using video recording as a cue. In addition, it was noted that users in gaze path condition did produce slightly more problem focused verbal data than their counterparts (Hansen, 1991).

Using gaze paths as a cue has many advantages over video recording. The gaze paths are replayed to the user as an overlay on the recording of screen activity, allowing the users to see exactly where they had been looking during the task performance. Gaze paths indicate where users' attention was directed during the testing. Eye tracking technology has developed during the last few years making it easy to record eye movements and replay gaze paths to the users. However, gaze paths cued retrospective verbalisation has not been widely used in the usability tests. This may be because of lack of research on the technique. As there have been only a few studies concerning the gaze path cued retrospective thinking aloud technique (e.g. Hansen, 1991; Ball *et al.*, 2006), there is a need for further information. The preset experiment addresses some questions concerning the usability problems observed and quality of operational comments produced by users with the gaze path cued retrospective thinking aloud technique, trying to shed light into the usefulness of the technique.

2. Usability – what is it and how should it be measured?

It has become a widely acknowledged fact that every product should be designed in such a way that its usability is taken into account. But what is usability? Preece et al. (1994) define usability as

> *"a measure of the ease with which a system can be learned or used, its safety, effectiveness and efficiency, and the attitude of its users towards it" (p. 722).*

Usability can be thought of as an attribute of a product, just like functionality. Functionality refers into what can be done with the product, whereas usability refers to how people can work with the product (Dumas and Redish, 1993). Therefore usability research concentrates on users: how people find their way to use a product, do they find using it easy, and can the product help them to achieve the goals they have set. Usability studies are not interested in the product itself, but how the product could be developed to be more usable for the users. However, both usability and functionality affect the productivity of the user with a product.

International standard organisation (ISO) defines usability by using three concepts: effectiveness, efficiency and satisfaction (ISO 9241-11, 1998). According to the standard, users should achieve the given goal by using the product with effectiveness, efficiency and satisfaction. The ISO standard is, however, only one way to define usability. Another well known and regularly cited set of concepts to evaluate the usability was developed by Nielsen in 1993. According to Nielsen (1993), usability consists of five components, which should be taken into account in the evaluation process. These five components are *learnability*, *efficiency*, *memorability*, *errors* and *satisfaction*.

1. *Learnability* refers to how easy it is for users to accomplish basic tasks the first time they use a product or an application. In the context of web pages, learnability also refers to how easily the users are able to navigate and learn the basic commands on a web page when they use it for the first time. If the standard is met, the users can find information easily on the web page and the pages are well structured.

2. *Efficiency* refers into how quickly the users can perform tasks after they have learned how to use the product. When using the web, efficiency is not met if the users are not able to get the information they are looking for fast enough, or if they are not able to see whether the pages contain the information they require. Users should always know where they are at the pages and where did they come from.
3. *Memorability* refers to how easy it is for the users to re-establish proficiency after a period of not using the product. On web pages, users who visit the pages only occasionally should be able to navigate the pages easily and remember the basic structure.
4. *Errors* refer to the number of errors users make whilst using the product, how severe these errors are and how easily the users can recover from them. The web pages should also be designed in such way that the users would make as few errors as possible. If the user makes an error when using the web pages, she or he should be able to recover from it immediately.
5. *Satisfaction* refers to the pleasantness of the user experience. Instead of feeling frustrated, the users feel that they are in control when using the pages, and they are able to find information easily and navigate the pages freely. (Nielsen, 1993)

All these concepts are important when evaluating the usability of a Finnish web site specialised to deal used cars, Autotalli.com (www.autotalli.com). The site attracts several new users per day, and therefore learnability is crucial in order for the new users to be able to find the information they are looking for. After all, a new user will not become a regular user if she or he does not find the use experience easy and pleasant. The web site contains a vast amount of information, which has to be found easily in a short period of time by users with different user experience and background (efficiency). Although there are users who may visit the car broking pages regularly over a long period of time, typically the pages are visited periodically; at times when the user is looking for a new car or is trying to sell one. The structure and basic commands of the site must be easily remembered also by those users who do not visit the site regularly, hence the memorability is important. As any product, autotalli.com web site should be designed in a way that the users make as few errors as possible, and if they do any, they are able to recover from them easily. Lastly, satisfaction is a concept which can be measured by asking the users of their experiences when using the site. Pleased users are more likely to become regular users.

Autotalli.com has many advantages over other forms to deal used cars (i.e. newspaper advertisements or magazines specialised to cars), as growing number of possible car buyers have accessibility to the Internet. These growing numbers of users also mean that users do have different needs when using the site; some users are very familiar with for example search engines, whereas for some users the use experience may be the first. Also the ageing citizens are becoming more and more connected to the Internet, and their special needs should be taken into account for example in font size. In order to create a well working site, usability tests were conducted to reveal possible usability problems within the site. These findings will be discussed later.

2.1. Testing usability

Currently, the concept to test usability of a product, web site, or a software application has arisen from the need to take the user into account when designing well functioning products, complexity of the interfaces and the changing user situations. The tools used in this task are called usability engineering methods. Usability engineering methods include methods for design, modelling, and evaluation. Hence, the evaluation methods are categorised into two separate types of methods: inspection methods and user testing (Ovaska *et al.*, 2005). In the inspection methods, the usability specialists use their own expertise to study the pages in depth and report the development team of any usability problems they may find. Probably the most famous inspection method is heuristic evaluation, first developed by Nielsen and Molich (1990).

Another way to evaluate usability is to conduct usability tests, in which the real users use the product like it would be used in real life. The usability specialists' job is to run the tests (i.e. plan and conduct the test session and report the findings) and collect data of the problems the users have during the tests.

Although several usability methods are available, finding the most appropriate in terms of time, resources and goals set by the client can be challenging. Therefore there is always a need for a new, reliable and fast technique, which can produce accurate information on users' attention and behaviour when surfing the web site.

The present study concentrates on retrospective thinking aloud using eye tracking data, which is a relatively new technique, and it could be used parallel to usability testing and think aloud. Usability testing is a method in which the users´ performance is measured while they complete a set of tasks given by the experimenter. Think aloud refers to a technique in which the users are asked to verbalise their thoughts during the testing, while they are completing the tasks. It will also be investigated whether the retrospective verbalisation technique using eye tracking data could replace concurrent thinking aloud during the usability testing altogether.

2.2. Usability tests

As mentioned above, by usability testing the experimenter receive information of the users´ performance with a product. Test sessions are observed and analysed by the usability specialists, and the improvements are suggested to the developers. Usability testing has its focus on practical side of usability, whereas studies on methodologies of usability testing focus on issues such as validity and reliability of usability testing and development of new methods.

The main idea of usability testing is to observe users working with the product to help to define the problem spots of the product. Because of this, it can be distinguished from the inspection methods (for example heuristic evaluation). However, there are several other usability evaluation methods in which the users are observed, for example ethnography (see Anderson, 1992) or focus groups (see Bloor et al., 2001). Usability testing should not be confused with the terms 'usability' and 'testing' in general. Usability testing is a method and nowadays has its own guidelines of how it should be conducted.

As the field of HCI is relatively new, several of the usability methods have been still finding their forms. The first usability tests were conducted in late 1980's and since that the method has been widely used. The practices have not been established at all times among the usability experts, and almost any testing involving users have been called usability testing. Even the best run usability testing can produce misleading results if the participants are not part of the product's user group (for example experts instead of novices). (Dicks, 2002)

Although there are several variations of usability tests, every test should have the following five characteristics (Dumas and Redish, 1993):

- The primary goal is to improve the usability of a product. In addition to this, every test should have specific goals and concerns, which are set by the experimenter whilst planning the testing.
- The participants represent real users.
- The participants conduct real tasks during the tests.
- The participants conduct the tasks under observation, and all actions and verbalisations are recorded by the test leader.
- In order to reveal usability problems, the data is analysed after the study. The findings and recommended improvements are reported to the client (or development team).

Usability testing has become very popular, and according to Dicks (2002) its value has been threatened by misuse of the term. Dicks claimed that there might be misunderstanding in distinctions between usability studies and empirical usability testing. There is a lack of knowledge of the limitations and the proper methods for usability testing, hence the validity and reliability of a great number of studies is questionable indeed. Dicks pointed out that usability specialists should remember the four functions that usability tests are set to measure; ease of learning, usefulness, ease of use and pleasantness of use. In order to conduct high quality usability tests, none of these four functions should be forgotten or emphasised more than another.

2.2.1. Conducting a usability test

Usability testing can be used at any stage of the design process: to test a prototype of a product, as well as the end product. Parts of a product may also be tested. Usability testing is suitable for many different types of products, for example web-pages, software, electronic devices or mobile services (Koskinen, 2005). Although usability testing is typically conducted with only one user at the time, two or more users can be tested simultaneously. This is called paired-user testing or co-participation.

In order to reveal as many usability problems as possible in the given time limit, the testing must be planned carefully. Typically the client wants to investigate the usability of a new product or to improve an old product. The usability specialists should carefully address the problems and questions for which they are looking for answers. Rubin (1994) argued that there are nine

characteristics which should be taken into account when planning usability tests:

- *Why the usability test is conducted.* At this point, the usability specialists should consider whether the testing is the most appropriate method and investigate the goals of the testing.
- *The exact questions and usability goals.* At this point the usability specialists are setting the exact questions and goals. Exact goal could be, for example, whether the user prefers mouse over the keyboard.
- *Profiling the users.* The group of end users has to be well defined, as testing with other users does not produce realistic results, and therefore it can be a waste of time. If the product is aimed for novices, the testing should be run with novice users. Using experts at the tests would not give a realistic view of the problems arising with the novice users.
- *Procedure during the test session.* The procedure should be carefully planned and structured before the actual testing. After setting the goals and profiling the users, the methods and apparatus can be decided. A well structured and organised study helps the test leader to maintain conditions of each session similar, hence keep the testing reliable.
- *Tasks.* The tasks should represent situations from the real life settings. Tasks should be easy enough to be remembered by the users, yet they should not be so small that they become trivial and insignificant.
- *Test setting and equipment.* The test settings should represent real life settings. This is rarely possible as the testing is typically conducted at the laboratory, but even the laboratory environment can be decorated appropriately. This can also be helpful for the users, as they may feel the laboratory intimidating and unnatural. By ensuring naturalistic settings, the test leader can ensure that the setting does not affect the results. However, it must be noted that the laboratory settings offer the best equipment to record data, observe the users and evaluate the data.
- *Observing the test session*. For the best results, there should be at least two professionals to run each test session. The test leader should be with the user, handing the tasks and observing him or her. Another person should be managing the technical equipment and making sure that everything goes according to the plan.

- *Collecting the data*. At this point, the usability specialists should decide the ways to collect data. Typically the user is observed during the test, and the observations are written down with a paper and a pen. In addition to this, recordings can be made, for example, with a video recorder or/and an eye-tracking recorder.
- *Report*. Although usability testing produces a large amount of data, the client is rarely interested of it all. Therefore the usability specialists produce a report including the findings relative to the client and development team. The report should be clearly and consistently written, using many examples to clarify the problems.

2.2.2. Usability testing – the strengths and concerns

One of the biggest advantages of usability testing is the fact that well conducted testing is an objective method to collect data. It reveals the problems on the product and the time spent by each user completing the tasks. It is a valuable tool when making clear-cut design decisions about products. It brings the real users' views to the design process, and it may give answers for questions like "What problems did the users experience when performing registration?" or "How long did it take for the users to register?" (Kantner and Rosenbaum, 1997). It can be conducted with early versions of the product, therefore the valuable information from the users can be taken into account at early stages of the development process. The profiled users are tested either individually or in groups, under controlled conditions. Usability tests conducted in a real world setting can offer an excellent opportunity for the usability experts to observe how well the situated interface supports the real users' work environment (Jeffries et al., 1991). Usability testing produces quantitative data, and therefore usability testing can increase the credibility of usability evaluation.

Kantner and Rosenbaum (1997) also suggested that usability testing has a considerable psychological benefit for the usability specialists and developers observing the test. Seeing the user struggling with usability problems is more convincing than the opinion of usability evaluators, and users may also give hints about the possible solutions for dealing with the problem. Seeing the user complete the task, but not necessarily the way development experts were expecting or hoped for, may reveal that the user has difficulties with using the product, but not necessarily problems that would prevent the use.

Although there are several strengths on using usability testing, given that usability testing consumes time and resources, it may not always be the appropriate method. Especially, when testing web pages, which can be revised in fast cycles, the motivation to conduct usability tests may be low. For the test results to be of optimal use, the web site developers should refrain from making changes during the test period, which is not often the case (Kantner and Rosenbaum, 1997).

Another concern of usability testing is that it requires an experienced usability expert, preferably several professionals. As described above, the validity and reliability requirements of the tests should be met and the test should be carefully planned. Planning the tests as well as recruiting and testing with users is often time consuming and therefore expensive.

In order to obtain reliable results, the users' input during the test is crucial. The users should work with the task as they perform in real life, even though the tests are typically conducted in the laboratory, under observation. The test leader may decorate the laboratory to resemble, for example office environment, although it can never be exactly the same as the natural environment where the product is used. As well as the environment, the users may feel the test setting unnatural. They may feel pressure to please the test leader or pressure to complete the tasks as fast as possible.

User testing is a valuable tool to observe users' behaviour objectively. However, one of its biggest flaws is that though it is able to indicate the symptom of the problem, it is not able to identify the cause of the problem. Therefore there are three commonly known techniques to ask for users' subjective experiences. The first technique is to ask the users what did they do and why. This has to be done after the testing, as it cannot be done during the testing in order not to disturb the test and affect the results. Interviewing the users afterwards is probably not the best way to obtain information, as people tend to forget what they were doing and they may rationalise what they did (Ericsson and Simon, 1984).

Another technique is to ask the users to verbalise their thoughts during (thinking aloud) or after (retrospective thinking aloud) the test. The next section will introduce these widely used techniques and discuss their benefits and concerns.

2.3. Thinking aloud

Thinking aloud (or concurrent thinking aloud) is a usability technique in which the participant is asked to speak out his or her thoughts while performing the tasks (Boren & Ramey, 2000). Thinking aloud technique originates from experimental psychology, it was first described in 1945 by Karl Duncker who studied productive thinking (Nielsen et al., 2002). In the field of human-computer interaction, thinking aloud is able to offer insights to user's thoughts, not only providing answers to the "what" question, but also to questions "how" and "why". It has grown to be one of the most used techniques to collect data, and it has even been argued to be probably the most valuable usability engineering method (Nielsen, 1993). It does not only reveal the problems within the product, but it also reveals the mental models the users have of the product. Mental models are representations of the real world and they help humans to simplify the complex environment. Mental models do not only define how users think, but they also determine how they act (Sinkkonen et al., 2002). If the mental models are negative in nature, users find the user experience difficult or unpleasant. Thinking aloud also reveals users' subjective opinions and images of the product. Users' opinions and images of the product might be less valuable information in the sense of usability, but they greatly affect the marketing of the product.

2.3.1. Protocol analysis by Ericsson and Simon

There has not been a standard practice of how to use the thinking aloud technique in usability tests, but one of the most cited reference is the protocol analysis by Ericsson and Simon (Boren and Ramey, 2000). The protocol analysis was developed in the 1980's in the field of cognitive psychology, but it has become popular in several other fields of study, such as usability research and studies of reading comprehension.

The protocol analysis is a rigorous methodology, which assumes that participants are able to verbalise their thoughts from the working memory in a manner that does not alter the sequence of thoughts mediating the completion of a task. Therefore the verbalisations can be accepted as valid data of thinking (Ericsson, 2002). According to Ericsson and Simon (1994), the verbalisations should be categorised into three levels. How the verbalisations are classified is affected by how purely the participant is able to concentrate on the tasks during the test, or whether she is exposed to many external stimuli. They

argued that the interferences make the verbalisations less reliable data of thinking. The three levels of verbal data are seen in Table 1.

Level 1	This level includes the thoughts that do not have to be formed before verbalisation. For example participants who count out numbers whilst they solve mathematic problems produce speech at the same form that is internalised into their working memory.
Level 2	At this level the information has to be moderated before verbalisation. This level of data are, for example, pictures and abstract concepts that users need to first formulate into verbal form before verbalising them.
Level 3	The third level verbalisations include information which has to be cognitively processed, and it is not directly attached to the task. This can be for example a situation in which the participant is asked to search information from her long term memory before verbalisation. All external interruptions and requests form the verbalisations into third level data, an example of the data at this level would be to ask a user in a usability test to verbalise the function of a scroll down before using it. Hence the natural verbalisation of the information from the working memory may interfere.

Table 1. Three levels of verbalisations by Ericsson and Simon, 1994.

Hence the protocol analysis argues that verbal data can be accepted as a valid representation of participants' mental processes if internal (i.e. information from the long term memory) or external (i.e. comments made by the test leader) stimuli do not cause disturbances between the working memory and verbalisations. In order to record valid level one and level two verbalisations, the speech has to be collected reliably. (Ericsson and Simon, 1993)

Although protocol analysis is one of the most cited references among the studies concerning think aloud technique in usability testing, it still is rather problematic. According to Boren and Ramey (2000), the procedures to conduct thinking aloud in the studies have not met the strict standards set by Ericsson and Simon, or the procedures have not been reported at all. For example Nielsen (1993) cited the protocol analysis as a usable technique to conduct usability studies, but on the other hand, Nielsen (1993) prompts test leaders to

encourage the users to think aloud by asking for example "What do you think this message means?" (p.196). In order to avoid external stimuli to disturb the user's verbalisations, more appropriate prompt would be for example "keep talking" (Ericsson and Simon, 1993).

The main goal of the theory proposed by Ericsson and Simon (1993) has been to form a model to study verbal data in the field of cognitive psychology. Their work has concentrated on cognitive processes of humans processing thoughts into words. They argue that humans simply produce verbal data of the tasks they are performing at the present moment and that they do not have a need to express the reasons why they act the way they do, or the behavioural patterns which lay behind their actions. This assumption of a human as a task and verbally oriented individual has been criticised, as it has been argued that humans are not able to act separately of their emotions, surrounding environment and senses (Nielsen et al., 2002).

Among the usability specialists there have been inconsistent practices to conduct and report studies concerning the thinking aloud technique. All attempts to collect verbal data have been called thinking aloud, based on the work by Ericsson and Simon (1984). Despite the benefits of the thinking aloud technique, if the protocol is not carefully planned and carried out, the cognitive processes used to perform the tasks may actually be changed or distorted by verbalising them out loud. For example Wright and Converse (1992) argued that users who were asked to verbalise their thoughts during the testing committed fewer errors and consumed less task time than users in the silent testing condition. Hence, their findings suggested that cognitive processes changed during the thinking aloud condition, resulting in a change in task performance, and therefore the data collected from the thinking aloud group was biased. The theory by Ericsson and Simon (1984) provided an explanation to these findings. They argued that when the articulated information is directly available in short-term memory, the concurrent thinking aloud protocol does not change task performance. However, the changes in task performance are more probable when the users are asked to provide specific information available only in long term memory (see above for level 3 verbalisation). These findings should be taken into account when planning the use of the thinking aloud technique. Level 3 verbalisations provide the developers with information necessary to enhance the product's ease-of use, but on the other hand, the usability specialists must acknowledge that using verbalisations at this level may lead to significant method bias. (Wright and Converse, 1992)

In order to produce valid and reliable results, the theory and standardised practises are crucial. However, it must be recognised that the main aim of the usability studies is not to concentrate on cognitive models or practises, the main aim is to enhance the usability of products or web pages. The protocol analysis by Ericsson and Simon should not be used in the usability studies as the objects set for usability studies differ dramatically from the aims set to studies in cognitive psychology (Ilves, 2004).

2.3.2. How should think aloud be used in the usability tests

Thinking aloud is easily affected by the user's personality or the behaviour of the test leader. Some users find thinking aloud natural and easy, and their speech reveals both the ongoing cognitive processes during the test session, and the interpretations of the situations. However, some users find verbalising their thoughts during the test difficult and unnatural. In the worst case scenario, the user's attention is drawn into the thinking aloud instead of the task performance. Thinking aloud is not suitable for studies in which the users are either children or expert users. Children are not able to verbalise their thoughts whilst they perform the tasks, and the expert users perform automatically and in such great speed that they are not able to verbalise their actions. (van Someren *et al.*, 1994)

The role of the test leader is very important. The task is to make the atmosphere in the test session as easy and natural as possible. Many users have never taken part in such studies and therefore they may feel pressure to perform differently than they would in the real world. This pressure can be eased by introducing the laboratory and the equipment to the user, describing the study briefly, and by giving instructions for the thinking aloud technique.

Before starting the testing, the users should be given instructions how the thinking aloud is conducted. Ericsson and Simon (1993, p. 378) advise the test leaders as follows: the users should be told that the test leader is interested in their thoughts and reactions, not their ability to use the product. Therefore the users are asked to think aloud. Thinking aloud means that the user tells everything she or he may think from the moment of starting the task to the moment of finishing the task. The test leader should ask the users to act as they would do when working alone and speaking out thoughts they have in their mind, not make interpretations of their thoughts or not explaining their actions.

The users should also be instructed not to plan what they are saying, and they should try to keep talking throughout the whole testing. The test leader should tell users that they might be reminded to verbalise their thoughts if they forget to verbalise them during the test. (Ericsson and Simon, 1993)

The test leader should also make sure that the users have understood the instructions and have internalised the test procedure in terms of what is expected from them during the test session. The test leader should practise the technique with the users or illustrate it shortly with a simple task, for example by filling up a stapler. After illustrating the technique, the users should follow the test leader's example and perform a simple task using the thinking aloud technique. The test leader should encourage the users to express their thoughts during the testing. (Dumas and Redish, 1993)

Using the thinking aloud technique requires careful planning from the usability experts. Even after the careful consideration, problems may occur. Preece et al. (1994) argued that cognitive load of both completing the task and verbalisation as well as the interruptive role of the test leader may have an impact in the results. It has been also argued that verbalisation may become difficult in complex tasks, or the users may find the whole situation awkward and therefore fall silent (Preece *et al*., 2002). Thinking aloud may slow down the decision process and thus create greater opportunities because of taking more time to think. Concurrent verbalising may also direct attentional capacity away from the tasks. The users may behave the way they believe to be a more socially desirable, presenting their thought processes the way they think the test leader would like to hear (Kuusela and Paul, 2000). Nielsen (1993) argued that task times are not valid at the experiments using concurrent verbalisations.

As seen above, one of the most used usability engineering techniques has some serious problems. Therefore it might be crucial to try to find other techniques to replace it. The next section will introduce retrospective thinking aloud, which has been another variant of thinking aloud to collect users' verbal data.

2.4. Retrospective thinking aloud

Retrospective thinking aloud is a technique originally developed and used in the field of cognitive psychology. In this technique users are asked to verbalise their thoughts after completing the tasks immediately upon completion of the testing. As the verbalisation is completed typically after the test session

(instead of after every single task), the possible flaws of the thinking aloud method (i.e. thinking aloud during the test session might have a negative or positive effect on users' task performance) can be avoided. As the retrospective thinking aloud data is collected after testing, also quantitative measures, such as task time can be observed. In the retrospective thinking aloud technique, the focus is on having the users to explain the thinking and reasoning processes they had during the testing. Retrospective thinking aloud has been used often in conjunction with the concurrent thinking aloud technique, to supplement data gathered from it. (Ericsson and Simon, 1993)

Although the retrospective thinking aloud technique can obtain valuable verbal data on thought sequences employed in completing tasks, it has never become as popular as the concurrent thinking aloud technique. While concurrent thinking aloud has been known to have several serious problems which could have been avoided by using retrospective thinking aloud, the usability specialists have not been keen to change the technique. One reason to this may be that the validity of retrospective thinking aloud has been argued to be problematic. In 1993, Ericsson and Simon argued that the retrospective thinking aloud provides valuable data 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, they may be unable to provide an accurate account of the thinking and problem solving strategies they had whilst completing the task.

The recent work by Guan et al. (2006) sheds some light into this issue. They noted that most of the studies concerning retrospective thinking aloud were studies comparing it with other usability inspection techniques, for example with concurrent thinking aloud. They pointed out that

> *"no research has scientifically studied the validity of retrospective thinking aloud based on its most fundamental claim—that in retrospective thinking aloud people talk about what they really did in terms of their actual mental processes or performance. Thus the validity of retrospective thinking aloud in usability research is still in need of serious investigation."* (p. 1253).

By comparing users' verbalisations with their eye movements, they found retrospective thinking aloud to be valid and reliable. They argued that the technique provided a valid account of what people attended to in completing

tasks, the technique had low risk of introducing fabrications and its validity was unaffected by task complexity (Guan et al., 2006).

However, retrospective thinking aloud is not a technique without problems. For example, users may report their actions or thought processes in such a way they believe the experimenter would want to hear. Furthermore, retrospective thinking aloud may contain judgements or strategies which are more rational than they would be in a setting without thinking aloud. They may explain their actions in a fashion that they consider to be more systematic, rational, organised, well-thought or coherent (Kuusela and Paul, 2000). On the other hand, retrospective thinking aloud technique using video recording or eye tracking data of the test session as a cue allows the users to see their actions, revise what they were doing, how did they feel about it and possibly suggest some improvements. For a skilled usability specialist, the retrospective thinking aloud session does not only offer data of the problems on a product, but it offers a conversation between the users and the development team.

2.4.1. Retrospective versus concurrent thinking aloud

There have been few studies comparing the retrospective and concurrent thinking aloud. The studies have been concentrating on the differences in the quality of the data obtained with both methods. The results have been rather controversial. For example, Kuusela and Paul (2000) conducted an experiment in the field of cognitive psychology comparing concurrent and retrospective verbalisations during a decision making process. According to them, the concurrent protocol analysis outperformed the retrospective protocol analysis in revealing participants' thought processes. They went on to argue that while the participants in both conditions were given the instruction in the same fashion, participants in the retrospective condition were not able to recall their judgement processes as well as their counterparts in the concurrent verbalisation condition. In order to understand the results of the study by Kuusela and Paul, it must be noted that the participants in the retrospective condition were instructed to recall their decision making process without a cue (i.e. video or eye tracking recording of their actions during the decision making process). Kuusela and Paul themselves recognise that although uncued retrospective verbalisations suffer from frequent forgetting and fabrication problems, prompted ones may not have the problem.

Bowers and Snyder (1990) found out that the users produced more words in the concurrent thinking aloud condition than the users in the retrospective thinking aloud condition, but there was a difference in the quality of the verbal data. Users in the concurrent thinking aloud condition produced more words describing their actions during the testing. The users were more likely to read texts on the screen or give a description of their own actions, whereas the users in retrospective thinking aloud condition using video recordings as a cue were more likely to explain their actions or give suggestions how the product design could be enhanced.

A study by Capra (2002) suggested that there was no difference in describing critical incidents during or after the testing. In this technique, users take time out from using the interface to describe interactions that increase or impair their performance (critical incidents). Users in the study preferred concurrent incident reporting, but this did not affect their performance as the users did report as many incidents in both conditions. Capra, however, pointed out that an interesting difference between the two techniques is that retrospective thinking aloud requires more time than concurrent technique. In a retrospective condition, the users need to watch their performance on a tape afterwards. Capra also noted that retrospective sessions did not interfere with task performance thus allowing the usability specialists to collect objective usability measurements during the testing, such as time to complete the task. (Capra, 2002)

As seen above, there have been studies comparing the concurrent and retrospective thinking aloud, but one, single conclusion cannot be drawn. Both concurrent and retrospective thinking aloud have advantages and disadvantages over each other, and they both have their place in the usability studies. However, it might be useful to try to find new ways to use them. The present study will use both of the techniques in addition to eye tracking in a usability study. Eye tracking as a usability engineering method will be introduced in the next chapter, and the present study will be discussed in detail.

2.5. Eye tracking

Eye tracking is a method which collects data from the user's eye movements. The history of eye tracking is long, but its use in usability studies has been relatively short. Eye tracking data offers information of users' attention on a

product. It can reveal users' intentional and unintentional processes during the testing. The user may, for example, try to find a certain button on a web page and therefore look at the page for a longer time than expected. The test leader cannot know whether the user did not find the button, or found it but did not understand the content of it. On the other hand, if the user did not look at the button, it was probably ill-located in proportion to other elements on the web page, or for example flashing advertisements drew the user's attention away from it. (Karn et al., (1999)

Two of the most fundamental eye movements are *fixations* and *saccades*. The time an eye dwells on a target is called a fixation. During the fixation only very small area of visual information can be processed. This area is the same as an area of 1 cm by diameter at the distance of 57 cm from the eye. Because of the small area seen at a time, the eye must move often, approximately in every 300 ms. These very rapid movements after each fixation are called saccades. During a saccade, both eyes move simultaneously and the jerky movement ranges typically about 2-10 degrees of visual angle and lasts about 25-100 ms. Visual information is collected during the fixations, and because of the high velocity of saccades, the collection of visual information is suppressed while the eye is moving. To help the eye to move to the right direction after every fixation, peripheral information is collected during every fixation. (Goldberg and Wichansky, 2003)

Probably the most often collected data from the eye tracking studies are the locations and lengths of the fixations, lengths of the saccades and *gaze paths*. Eye movements which are used to process the visual environment (i.e. fixations and saccades) are called gaze paths (may also be called scan paths). A gaze path consists of sequences of fixations and saccades.

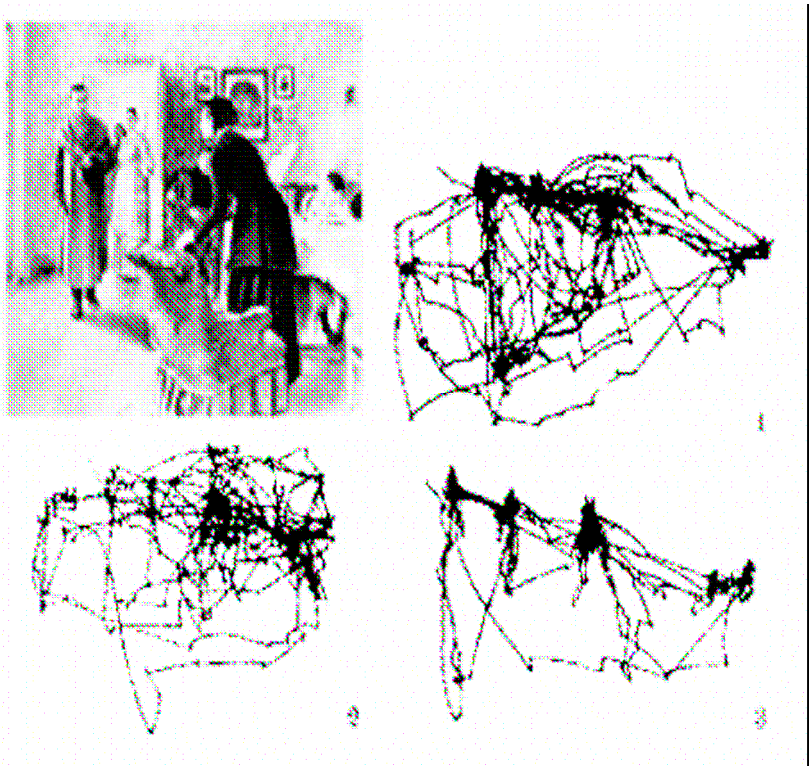


Figure 1. Different scan paths (Yarbus, 1967).

The gaze paths represent each person's individual way to perceive the target. Past experiences, need for information and motivation form the way individuals view, for example, a painting. In a study conducted by Yarbus (1967), the participants were asked to view Ilja Repin's painting called "They did not expect him". Figure 1 shows the first three of the scan path visualisations studied by Yarbus. In the first visualisation the participant was not given any instructions before viewing the painting. In the second visualisation, the participant was asked to rate the socio-economical background of the family in picture, and in the third visualisation the participant was instructed to rate the ages of each family member in the painting.

Fixations and saccades are not the only type of eye movements, however they are the most relevant for the usability studies. For further information of other types of eye movements, see for example Goldberg and Wichansky (2003).

2.5.1. Eye movements and cognitive processes

Just and Carpenter (1980) studied the link between eye movements and text comprehension. They suggested that eye stays fixated to a word until it is understood by the reader. Andreassi (1995) conducted a meta-analysis of several studies on cognition and eye movements. According to the analysis, there was a link between the learning strategies and the fixation durations. Students who had powerful learning strategies were found to have longer fixations than students with weaker strategies, though it must be noted that the fixation times were longer with all the students when they were presented lists of difficult words. The number of saccades and fixations grew at the cognitively demanding tasks, or when the participants were asked to view a picture whilst solving a problem. (Andreassi, 1995)

A link has been found between unconscious horizontal eye movements and natural emphasis on brain hemispheres. EEG measures suggest that activity in brain is emphasised either on a left or right hemisphere. Humans, whose left hemisphere is more dominant than the right hemisphere, move their eyes right during verbal tasks and visa versa. It is also noted that persons with left hemisphere dominance are right handed. People, who report to be right handed, move their eyes right during verbal tasks and left during spatial tasks. (Andreassi, 1995)

There have been studies linking other types of eye movements, such as changes in pupil size and eye blinking and cognitive processes. Changes in pupil size during emotional or cognitively demanding tasks have been reported (for example Aula and Surakka, 2002; Partala and Surakka, 2003). Due to problems with measurements, only studies conducted in the last 30 years can be seen as significant (Andreassi, 1995). One of the biggest problems in measuring pupil size has been the fact that the pupil reacts to light and the size changes in different lighting. It has been noted that pupil size decreases when the participant is tired, whereas, for example, feeling of fear or a cognitively demanding task increases the size. Increase in the pupil size reflects an increase in neural information processing in cognitive processes. Less increase can be observed during easier tasks. However, Andreassi (1995) pointed out that there was an increase in pupil size while a stimulus was expected. Although measurement of pupil size is not widely used in the usability studies, it may offer an opportunity for the researchers to study users' emotions or strain of the cognitive capacity whilst they use a product.

Although there are several ways to use eye movements in usability studies, one of the biggest benefits of eye tracking is probably to obtain information on to where users' attention is focused. According to Neisser (1967), focusing visual attention is a process with two stages; pre-attentive and focal attention. During the pre-attentive stage the human is able to perceive information from the whole visual field, whereas during the focal attention humans perceive information from one or few stimuli at most. Observing and structuring the visual field in general, as well as focusing attention into one target occur during the pre-attentive stage. Processes of attention during the pre-attentive stage are not conscious, and physical facts such as similarity, disparity or physical closeness have an effect in choosing the stimulus.

2.5.2. Advantages and disadvantages of eye tracking

Eye tracking is a method with many possibilities. Eye tracking produces objective and quantitative data. Yet the data (e.g. gaze paths) can be analysed qualitatively. Eye tracking can be used as the only method, or it can be used in addition to other methods. Collecting data is fast, and it can be done in real world settings (e.g. studies of pilots) in which the users' attention cannot be disturbed with, for example, interviews. Eye tracking offers information of users' unconscious eye movements (i.e. perception processes) and it may offer information on users' emotional state or cognitive load. It can also be used with user groups with special needs, for example with children. (Lehtinen, 2005)

However, there are some disadvantages in usage of eye tracking. First of all, there has been a lack of consensus on how eye movements could be linked to cognitive processes (Cowen, 2001). Nor does the method offer information on why the user behaves in certain manner. Eye tracking data offers answers to questions such as "what" and "when", but not "why".

Probably the biggest problem with eye tracking has been the limitations with technology. Problems may occur when using the trackers, as participants may need to be disqualified due to problems posed by heavy make-up, glasses, contact lenses or even eye colour. The number of disqualifications may affect the quality of the sample. Many disqualifications make the eye tracking studies time and resource consuming.

Also the equipments have been rather difficult to use. The devices have been expensive and required a lot of technical knowledge from the experimenters.

Due to the limitations of the eye trackers, the users have been restricted from moving naturally, even the slightest head movements may have caused problems with calibration and the whole testing session. However, there are new eye trackers available which are better suited for usability studies. For example, Tobii 1750 eye-tracker (Figure 2) allows the testing to be very naturalistic, as the tracker is located to the bottom of the screen. Users perform the tasks in the manner they would do with a regular PC. (Lehtinen, 2005)

Figure 2. Tobii 1750 eye tracker integrated with 17″ TFT display

2.5.3. Eye tracking in usability studies

Eye tracking is able to reveal information of users' visual search patterns and attention focusing. For example Aaltonen *et al*. (1998) used eye tracking technique to study how users read menus. By analysing the scan paths, they argued that users read menus with consecutive sweeps (Figure 3.).

Figure 3. Visualisations of user's scan path during a visual search (right) and normal reading situation (left) from menu (Aaltonen *et al*., 1998)

According to Goldberg and Kotval (1999), the length and duration of scan path reflects the effectiveness of users' visual search. A spatial scattering of the fixations reveals the area of user's visual search. Optimally, a visual search task has only few saccades and fixation on the target object. Simola (2004) argued that a scan path covering only small area indicates effective search, whereas a scan path scattered evenly to the interface indicates a poor design.

Some studies have used eye tracking successfully in terms of design suggestions. On the basis of eye tracking data, for example Goldberg et al. (2002) pointed out that users were more likely to choose buttons on the left upper corner, hence the important information should be placed there.

Pan *et al.* (2004) studied users' eye movements while they were reading a web site. They investigated whether users' gender or navigation order affected the eye movements. They argued that males had longer fixations than females, and that users' fixations were longer at the first pages, whereas the fixation times decreased when users were asked to navigate through several pages.

Examples presented above are only a fraction of studies conducted in the past few years. For further reference see Jacob and Karn (2003) for a list of 20 usability studies conducted in the last 50 years. The list provides a good sense of how usability studies have evolved during the history of eye tracking studies (Jacob and Karn, 2003).

However, the data provided by eye tracking is rather limited. As mentioned above, eye tracking data cannot give answers to questions, such as "why" or "how". Therefore the verbal think-aloud data is often collected.

2.6. Overview of the present experiment

The present experiment on using eye tracking in usability testing is described in detail in the next chapter. The experiment was motivated by the work of Hansen (1991). Hansen conducted an experiment, in which the usage of eye tracking data was compared to video recordings in retrospective thinking aloud. According to Hansen, eye tracking data was useful, as the users produced as many operational comments (i.e. users commented their actions, what they saw and their cognitive processes) as in the video recording condition. The experiment by Hansen (1991) suggested that a record of eye movements is as useful as regular video recordings in the retrospective thinking aloud technique. Yet the results did not reveal whether the technique was more powerful than the widely used concurrent thinking aloud technique.

Overall, there have been only few studies using eye tracking data and verbal protocols. One of the early studies was conducted by Russo (1978) in the field of consumer psychology. Russo compared recording eye fixations with four alternatives ways of collecting data from the users: chronometric analyses, information display boards, input-output analyses and (concurrent) verbal protocols. Several factors of each method were examined and the methods were compared. The factors under investigation were: detail revealed, informativeness, validity, range of settings, unobtrusiveness, ease of use and

cost. Interestingly, the study suggested that "verbal protocols are remarkably complementary with eye fixations" (p.569). This was due the differences between the methods, hence Russo (1978) argued that disadvantages of one method are compensated by the strengths of the other.

A recent study by Ball *et al.* (2006) suggested that using gaze paths as a memory cue in a retrospective think aloud technique is a useful tool indeed. Ball *et al.* argued that using gaze paths helped to identify significantly more usability problems than the think aloud technique, yet using gaze paths as a memory cue did not reveal more usability problems than the condition in which the users conducted retrospective thinking aloud using the playback of dynamic screen events (including cursor movements) that had arisen during task performance as a cue. However, they noted that gaze paths revealed more usability problems with certain search engines.

The present study was conducted by using eye tracking with concurrent and retrospective thinking aloud techniques. In a typical usability study, the users are instructed to think aloud during the testing and the retrospective verbalisations are collected by asking the users to view their video taped performance, or by using a screen as a cue. In the present study, the users were asked to complete a set of tasks (usability testing) individually. In the first condition, each user was asked to think aloud during the testing, whereas in the second condition, users were not given any instruction to verbalise their thoughts. After the testing, both groups were prompted to think aloud retrospectively while they watched their gaze paths collected with a Tobii 1750 eye-tracker (Tobii Technology, 2007).

Two hypotheses were made:

1. *The gaze path cued retrospective think aloud method used in the present study is able to reveal as many or more usability problems as were identified with traditional usability testing.*
2. *It is expected that users in retrospective condition who did not conduct think aloud during the testing were able to produce more comments than users in concurrent thinking aloud condition. It was also expected that the quality of comments were better in retrospective thinking aloud condition than in concurrent thinking aloud condition. Therefore it is hypothesized that retrospective verbalisation using gaze path as a cue, is such a powerful method that there is no need to use concurrent thinking aloud in the usability tests.*

3. Method

3.1. Design

Eight users took part in the present study. They were asked to conduct a set of tasks which were developed and used by students of the usability engineering methods course to test the same autotalli.com web-site in spring 2005. The usability problems observed within the present study were compared with problems found with regular usability testing by the students at the previous course. Four users were instructed to think aloud during the testing, whereas four of the users were given no instructions to verbalise their thoughts during the test session. However, they were not prohibited from doing so. The tests were conducted in a typical way, but in addition eye tracking data was collected with the Tobii 1750 eye-tracker.

After completing the set of tasks, all users were asked to view their individual scan paths of their performance, and they were encouraged to think aloud retrospectively whilst watching the recordings. All verbalisations were collected on tape, and they were analysed later. After the analysis the usability problems observed in the present study were compared to the usability problems found by the students in the usability engineering methods course.

The present study investigated whether eye tracking as a usability inspection method could reveal the same number or more usability problems than regular usability testing (conducted by the usability engineering methods course). The present study also investigated and compared the quantity and quality (according to Hansen's categorisations) of verbalisation in concurrent and retrospective thinking aloud conditions. Hence, the users were shortly interviewed whether they found gaze paths useful and easy to use as a memory cue.

3.2. Materials

The experiment included a usability test on the autotalli.com-web site. The site has several pages, yet the experiment focused on some of the most used pages. The following list shows the most used pages, and pages marked with * were the pages included in the present experiment:

- Front page*
- Search*
- Listings*
- Search results*
- Help
- "My pages" – pages with special features, required registration:
    - Watchdog
    - Favourites
    - Saved search results
    - User information*

The tasks were selected from the tasks used by the usability engineering methods course, their tasks covering all the pages listed above. However, some limitations occurred when selecting the tasks. The first problem was the number of tasks produced by the course. Students formed twenty two groups, each group developing twelve tasks. It was clear that not all the tasks were suitable for the present experiment, nor was it possible to use them all. In order to find the relevant ones to the present study, the experimenter chose tasks according to their relevance, to cover as many pages as possible and to fit the real user group of the site.

Another limitation was time. In order to avoid users to become tired and loose their ability to concentrate, the testing in all could take approximately an hour. Therefore all the pages listed above could not be tested in the present study. In order to be able to comment on their gaze paths, the eye tracking visualisation had to be shown to the users at half speed. Taking this into account, the retrospective thinking aloud session took twice the time than the testing. Results from the pilot tests suggested that completing the tasks could take approximately 15 minutes and retrospective think aloud approximately 30 minutes. The users were also introduced to the laboratory and they were asked to fill in forms before and after the testing. This took approximately 15 minutes.

The third problem arose during the pilot testing. Instead of handing the tasks to the user on paper, the tasks needed to be in a verbal form. This was due to the fact that after calibration the users' eyes were required to stay focused on the screen as much as possible. If the users' eyes were off screen too often or for long periods of time, there was a possibility of loosing the calibration. In order to avoid the problem, the tasks were formed to be short enough for the users to comprehend and remember, yet they had to be simple enough not to add users' cognitive work load. However, tasks possibly loading users' working memory were also used. For example, the users were asked to log into the site (with account name and password previously created by the experimenter). In this case the information was given to the users with a note attached to the upper corner of the screen. See Appendix 1 for the set of tasks used in the experiment.

3.3. Participants

Twelve users took part in this experiment. However, four users were disqualified due to various problems. One of the users did not show up, and three test sessions were discontinued due to technical problems. One of these unsuccessful sessions ended as problems occurred with data recording, and two of the sessions were ended due to frequent failure in calibrations. This was probably due to users' eye glasses (although there also were successful test sessions with users wearing glasses).

The ages of the remaining eight users ranged from 24 to 33 years, average being 30 years. Three of the users were male, five female. The real end user group of the autotalli.com-web site is dominated by male users. However, due to the problems discussed above, three male users were disqualified and due to the limitations with time, the experimenter was not able to recruit more male users. Two of the users had used the autotalli.com-site before, whereas six users were unfamiliar with the site. Seven of the eight users had valid driving licence, four owned a car, and three had a possibility to use someone else's car on a daily basis. One of the users had no car at all. Two of the users were planning to buy a new car in the near future.

All users rated their ability to use computers high, and all the users used Internet on a daily basis. Mainly they used Internet for searching information,

reading the news, receiving and sending e-mails, and electronic services, such as online personal banking.

3.4. Apparatus

Two PCs were used to run the experiment and record data. Users' PC had Windows XP and the Tobii eye tracker installed in it, and the users used Internet Explorer 6 to conduct the tasks. The experimenter used pen and paper to write down notes during the testing. Another computer connected to the Tobii PC was used to record a video of the whole session.

Tobii 1750 eye-tracker
Tobii 1750 eye tracker integrated with 17" FTF display was used as an eye tracker. ClearView eye gaze software was used to collect users' eye tracking data and display the gaze paths during the retrospective thinking aloud condition. Tobii 1750 eye tracker's sampling rate is 50 Hz. The users were required to use mouse and keyboard during the experiment.

Sony Handyman video camera
Sony Handyman video camera with 3.0 Megapixels was used to collect verbal data via a microphone placed on the users' work desk. It was also used to video record an overview of each test session.

Noldus Observer 5.0
Noldus Observer 5.0 was placed on the experimenter's PC to collect and analyse verbal data. Noldus Observer was used only to collect the video recordings.

3.5. Procedure

On entering the room, the laboratory and all the equipments were shortly introduced to each user. The users were asked to turn off their mobile phones and they were asked to sign an informed consent form (Appendix 2) and fill in a questionnaire asking some personal information related to the study (Appendix 3). The procedure was explained shortly, and the users were told that they had a right to quit the experiment at any time. The thinking aloud technique was explained to users in the concurrent thinking aloud condition

and they were allowed to practise it shortly. The users were instructed to think aloud as follows:

> *Thank you for taking part to my study. The present experiment does not observe your ability to use the site, but it is set to make the site better. I'm asking you to think aloud, which means that you should talk aloud all your thoughts during the testing: whatever comes to your mind. If you forget to think aloud, I might prompt you to do so. However, I cannot help you with the tasks.*

All users were allowed to ask any questions related to the experiment before starting to complete the tasks.

After the introduction, the user was asked to sit down at the users' PC. They were advised to pull the chair in a position where the user's face was in 70 centimetres distance from the screen. The users were told that the testing was recorded with the Tobii 1750 eye-tracker, as well as with a video camera. The users were advised to sit as still as possible during the testing. After the instructions, the Tobii 1750 eye-tracker was calibrated. Before starting the actual testing, users in the thinking aloud condition were reminded about verbalising their thoughts during the testing.

The testing started with the experimenter reading out the first task, and the tasks were given one at the time (Appendix 1). The experimenter used a timer to measure the time used for every task. The experimenter wrote notes on observed problems and actions during the test session. Some users forgot to think aloud during the testing, and in these cases the experimenter prompted the users to verbalise their thoughts during the testing. The task times were measured. After all the tasks, the Tobii 1750 eye-tracker was stopped.

At the second part of the study, the user was asked to view his or her own scan path from the PC with the experimenter. The users were instructed to think aloud retrospectively as follows:

> *Now we are watching the eye tracking recording together, and I'm asking you to think aloud. This means, that you should speak out all your thoughts when you see the recording. I might prompt you if you forget to think aloud while you do the tasks.*

Before starting the viewing, the experimenter explained the scan paths and fixations shortly to the user. The users were prompted to think aloud if they fell silent instead of verbalising their thoughts. The experimenter wrote notes during the retrospective thinking aloud condition.

At the end of each session, the users were asked to complete a short questionnaire (Appendix 4) and the users were shortly interviewed. The user was thanked for the participation and the main aim of the study was explained.

3.6. Coding

A transcript of users' verbalisations was divided into three groups: 1) concurrent thinking aloud condition, 2) retrospective thinking aloud with concurrent thinking aloud condition and 3) retrospective thinking aloud without concurrent thinking aloud condition. The total number of verbalisations was calculated in each group, and operational comments (Hansen, 1991) were investigated. Operational comments are the comments in which the users are verbalising their actions during the testing. Operational comments were divided into three categories: manipulative operations, visual operations and cognitive operations.

1. Manipulative operations
   For example:
   "I *write* down my name"
   "I could have *clicked* them all…"
   " Oh dear, I *entered* that wrong!"

2. Visual operations
   For example:
   "I *saw* it here somewhere!"
   "I'm *looking* at the picture"
   "I *read* it from the previous page"

3. Cognitive operations (interpretations, expectations, evaluations and specifications of action)
   For example:
   "I *remembered* it to be there…"
   "At this point I finally *realised* that there is a scroll bar there"

"I *figured out* that I couldn't find it there"

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…" This sentence was categorised as 1 manipulative and 1 cognitive operation comments. (Hansen, 1991)

4. Results

Two types of data were collected: usability problems investigated with the method used in the experiment, and the quantity and quality of words used in different conditions.

4.1. Comparison of usability problems found in the present study and by regular usability testing

In order to gain information of the usefulness of the present method, usability problems were investigated and reported. These findings were compared to problems found by usability engineering method course. It was predicted that at least as many or more usability problems could have been found with the method under investigation (hypothesis 1). As seen in Table 2, this was not the case.

	Number of usability problems observed by usability engineering course (66 users)	**Number of usability problems observed at the present experiment (8 users)**
General problems	4	3
General problems with navigation	6	5
Front page	5	5
Search	17	12
Listings	10	6
Search results	8	4
My pages	5	3
My pages – User information	8	6
Total number	**63**	**44**

Table 2. Number of usability problems observed by the usability engineering method course and the present experiment

At the second part of the usability problem analysis, it was investigated whether concurrent and retrospective thinking aloud conditions did reveal different usability problems. As Table 3 reveals, most of the problems would have been observed in both conditions (31 problems), but altogether 11 usability problems were observed only in retrospective think aloud condition and two problems in concurrent think aloud condition.

	Number of problems observed only in TA condition	**Number of problems observed only in RTA condition**	**Number of problems observed in TA and RTA conditions**	**Total number of problems observed**
General problems	0	0	3	3
General problems with navigation	0	1	4	5
Front page	0	2	3	5
Search	2	2	8	12
Listings	0	2	4	6
Search results	0	0	4	4
My Pages	0	2	1	3
My Pages - User Information	0	2	4	6
Total	**2**	**11**	**31**	**44**

Table 3. Number of problems observed in concurrent, retrospective and combined conditions.

However, retrospective thinking aloud technique using gaze paths as a cue did reveal additional information on the problems observed in concurrent thinking aloud condition. It was noted, for example, that the front page contained a large amount of information, including links to the other pages, *search by number* link, *log in* and *registration* links and several advertisements. In the first tasks the users were asked to log in, which turned out to be problematic as

several users had to search for the appropriate link. Retrospective thinking aloud cued with gaze paths revealed that six out of eight users viewed the upper right corner of the screen when trying to find the log in. Users commented their gaze paths:

> User 1: *"This site should have been done so that the logging in is placed at the upper right corner… that's the place where I tried to find it."*

> User 2*: "Quite often, at least at the sites I use the log in is placed at the upper right corner."*

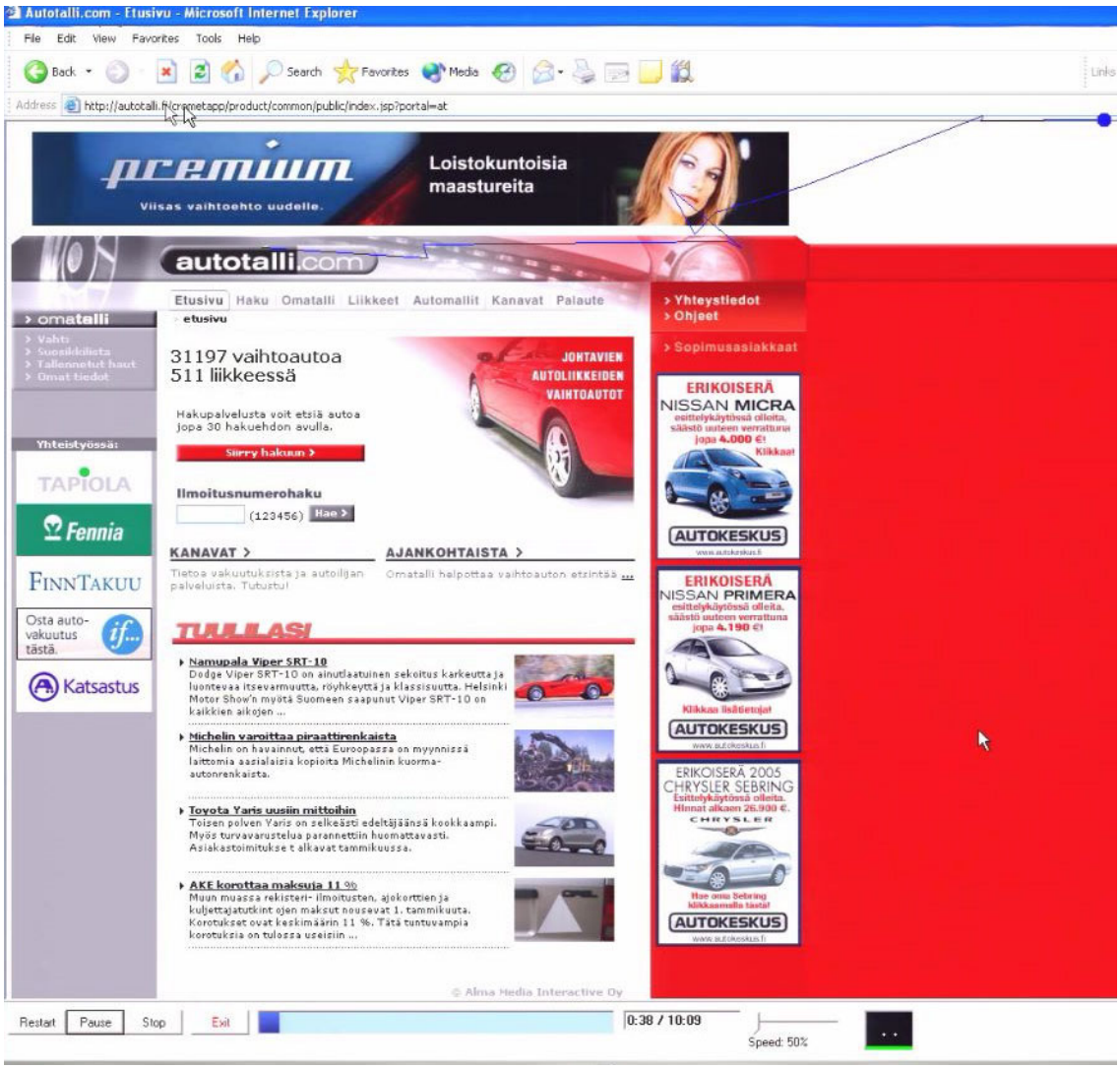


Figure 4. User's gaze path when searching for log in

At another case the gaze paths revealed that when users searched for log in from the front page their attention was drawn (i.e. their eyes fixated) into the *search by number* fill in box (Figure 5.). This indicated that the search by number function was oversized and therefore it should be redesigned.

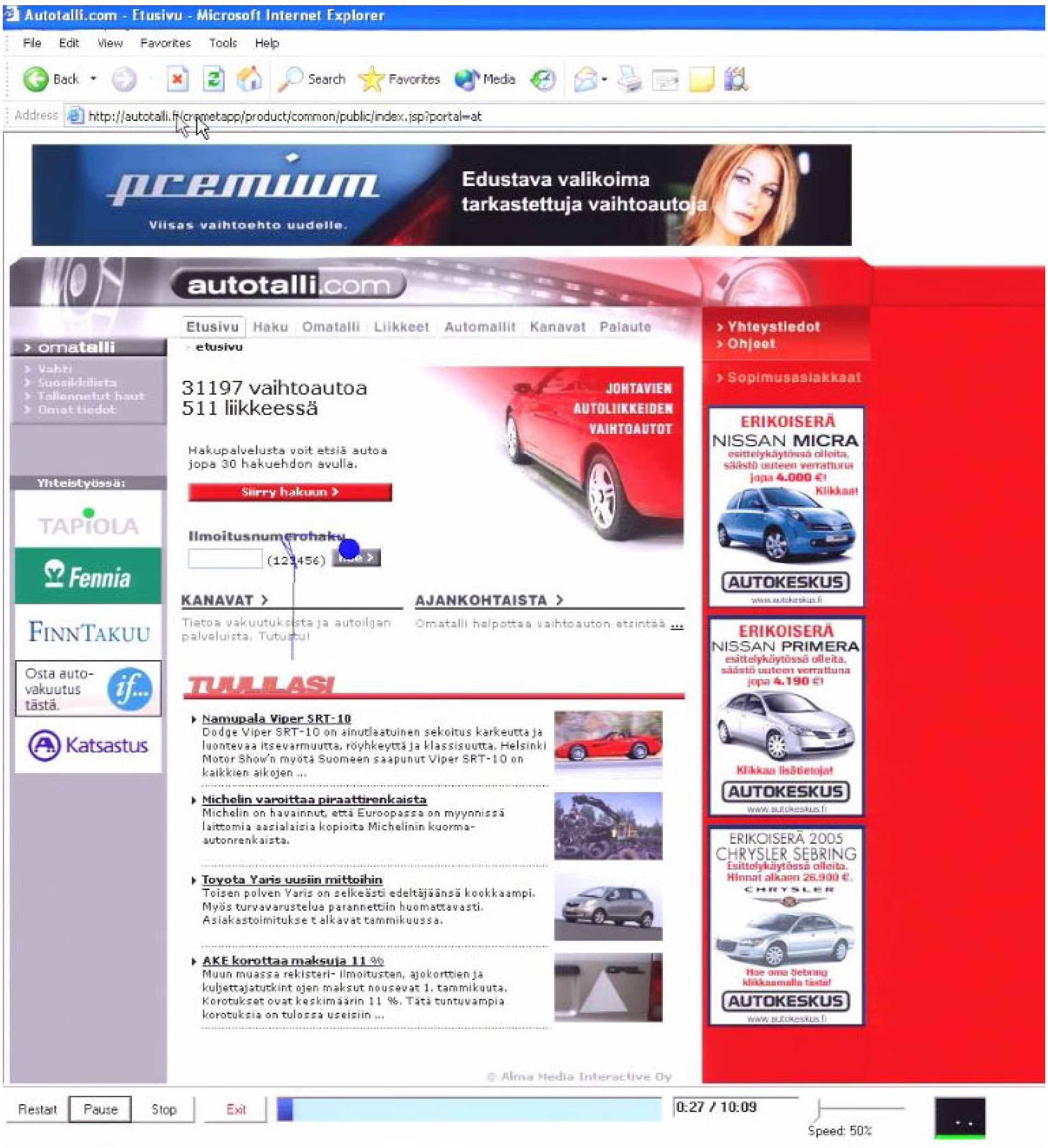


Figure 5. User's attention drawn to search by number function

It was also noted that users who conducted both concurrent and retrospective verbalisation commented their actions differently in the two conditions. In several cases, users noticed that they had problems to complete the task in concurrent thinking aloud condition, hence retrospectively they were able to

analyse why this might be the case and even suggest improvement ideas. In the following example, User 3 tried to find a check box from the basic and the specific search pages. However, the place of the check box moved on the pages when user navigated between the two search possibilities. User 3 commented the problem in the concurrent condition:

> User 3*: " Just a minute ago there was a box here... where... it just was here... oh, where did I see it? This is just too hard a task for me!"*

User 3 analysed the same problem retrospectively:

> User 3*: "In the end I realised that there was a scroll bar on the right and that some of the stuff is there. Then I found the check box. I think it's quite unbelievable that the check box was there and then all of a sudden it was not there! I remembered that I had seen it there, and it is a clear flaw that the screen views are changing when you choose the other search, it shouldn't be like that. Not very good at all."*

It was also investigated whether there were differences in usability problems when users did carried out both concurrent and retrospective thinking aloud. It was counted that 16 usability problems were observed in concurrent thinking aloud condition, whereas retrospective thinking aloud revealed 19 usability problems (i.e. 16 were the same problems that were observed in concurrent thinking aloud condition and three were new problems). There was no difference in the quality of problems observed with the concurrent and retrospective thinking aloud techniques, hence both techniques did reveal both serious and less serious design flaws of the site.

4.2. Comparison of the number and quality of words in the concurrent and retrospective thinking aloud conditions

At the second part of the analysis, the quantity and quality of words produced by the users in different conditions were analysed. In order to investigate whether the differences were significant, the data for each measure were analysed using a 2 x 2 x 3 analysis of variance (ANOVA). The factors were Condition, Stage and Category. Condition was a between-subjects factor with two levels (condition 1 concurrent thinking aloud, and condition 2 without

encouraged concurrent thinking aloud). Stage was a within-subjects factor with two levels (concurrent verbalisations and retrospective verbalisations); and Category was a within-subjects factor with three levels (manipulative, visual and cognitive). The adopted level of significance was $p < 0.05$.

	Total number of words	**Total number of operational comments**
CTA	1148	66
RTA (users conducting CTA)	3309	214
RTA (users not conducting CTA)	4136	267

Table 4. Total number of words and operational comments by users in three conditions

As the results in Table 4 suggest, users did verbalise their actions more in retrospective than concurrent think aloud condition. A significant main effect of Condition ($F_1 = 29.628$, $p < 0.005$) was observed suggesting the average amount of words was significantly different between concurrent and retrospective verbalisations, users producing significantly more words in total during the retrospective think aloud condition.

4.2.1. Frequency of operational comments

The operational comments in the experiment were categorised into three groups. The groups were: manipulative, visual and cognitive comments. Table 5 shows that 82% of the comments made by users who were asked to verbalise their thoughts during the testing were manipulative comments, i.e. they were commenting what they were doing at the time. The amount of cognitive comments in that group was only 4%. Forty two percent of the comments made by the users in the retrospective thinking aloud condition (without concurrent verbalisation) were manipulative, whereas 43% of their comments were cognitive. Both retrospective thinking aloud groups made the smallest amount of visual comments, whereas concurrent thinking aloud group made the smallest amount of cognitive comments. The percentage of visual comments was almost the same between the groups.

Table 5 also shows the means and standard deviations for the number of operational comments (i.e. manipulative, visual and cognitive comments)

produced by users at two different stages measured in the experiment (concurrent and retrospective verbalisations). The table reveals that the mean number of manipulative comments produced by users in the concurrent think aloud condition was 13.5, whereas the same users produced twice the amount of manipulative comments (mean 28) retrospectively. However, the users who were not asked to verbalise their thoughts during the testing produced the most manipulative comments in retrospective verbalisation (mean 29.8).

Users in the concurrent thinking aloud condition produced fewer visual comments (mean 2.3) than when they were asked to verbalise their thoughts while they watched their gaze paths (mean 7.5). Users in retrospective condition without concurrent verbalisation produced most visual comments (mean 10). Surprisingly, the users in concurrent verbalisation made averagely only one cognitive comment (mean 1), whereas the same users produced retrospectively more cognitive comments (mean 18) and the users conducting only retrospective verbalisation produced the most cognitive comments (mean 28.5).

	Concurrently			**Retrospectively**		
	Manipulative	**Visual**	**Cognitive**	**Manipulative**	**Visual**	**Cognitive**
CTA						
Mean	13.5	2.3	1.0	28.0	7.5	18.0
St.Dev.	8.9	1.0	2.0	10.5	6.6	7.0
%	82	14	4	53	14	33
With-out CTA						
Mean	1.0	0.5	0.3	29.8	10.0	28.5
St.Dev.	1.4	1.0	0.5	14.1	4.1	17.3
%	70	20	10	42	15	43

Table 5. Means, standard deviations and percentages for comments produced in different conditions (concurrent think aloud and without encouraged think aloud) and stages (concurrently and retrospectively).

A significant main effect of comment categorisation ($F_2 = 20.235$ $p = 0.001$) was observed suggesting the mean number of operational comments did vary significantly over manipulative, visual and cognitive categories. Significant interaction also occurred between concurrent and retrospective verbalisations and comment categories ($F_2 = 14.056$ $p = 0.001$). Chi-square (the χ^2 tests) was performed to analyse the differences between word categories, and significant

difference was found between the frequency of manipulative comments on concurrent and retrospective conditions (χ^2 = 8, df = 1, N = 8, *p = 0.005*). Significant difference was also found between the frequency of visual comments on concurrent and retrospective conditions (χ^2 = 4.5, df = 1, N = 8, *p = 0.034*). Furthermore, significant difference was found between the frequency of cognitive comments on concurrent and retrospective conditions (χ^2 = 8, df = 1, N = 8, *p = 0.005)*. As seen from Figure 6, the results suggested that users produced significantly more words in every category in the retrospective condition.

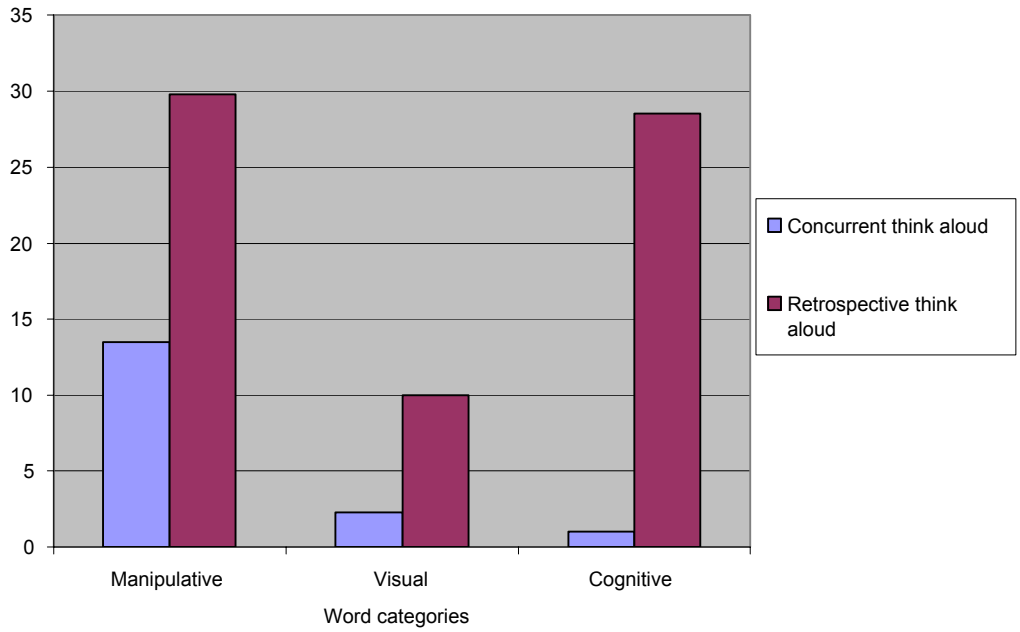


Figure 6. Mean number of operational comments in concurrent and retrospective verbalisations

The differences between users were also investigated by counting the number of operational comments made by each user. The number of operational comments did vary between the users, suggesting that some users did find verbalisation easier than others. Table 6 shows the number of operational comments made by each user. As seen from the table, for example User 2, who did conduct concurrent thinking aloud condition produced only one manipulative and one visual comment, whereas User 3 (also conducting concurrent thinking aloud condition) produced 22 manipulative and 2 visual comments.

CTA/ without CTA	CTA			RTA		
	Manipulative	Visual	Cognitive	Manipulative	Visual	Cognitive
U1 no	3	0	1	42	13	45
U2 yes	1	1	0	18	6	12
U3 yes	22	2	0	36	17	28
U4 yes	15	3	4	20	5	16
U5 no	0	0	0	12	6	9
U6 yes	16	3	0	38	2	16
U7 no	0	0	0	40	14	41
U8 no	1	2	0	25	7	19

Table 6. Operational comments produced by each user.

In order to investigate what users thought about the verbalisation techniques, they were interviewed shortly. Three out of four users who did conduct both concurrent and retrospective thinking aloud conditions answered that they preferred retrospective verbalisation. The one user who preferred concurrent verbalisation argued it to be more natural for her. Users did not find the concurrent thinking aloud disturbing or unpleasant, however, one user mentioned that *"…it probably made it (task performance) clearer for me"* (User 6).

4.2.2. Task times in concurrent thinking aloud condition and condition without thinking aloud

In order to investigate whether concurrent thinking aloud affected the time users spent completing the tasks, the task times were measured. Table 7 shows the average time spent by the users to complete each task in concurrent thinking aloud condition and condition without verbalisation.

	CTA	Without CTA
Task 1	22 sec.	22 sec.
Task 2	1 min. 32 sec.	1 min. 39 sec.
Task 3	33 sec.	30 sec.
Task 4	1 min. 14 sec.	36 sec.
Task 5	57 sec.	54 sec.
Task 6	2 min. 26 sec.	1 min. 34 sec.
Task 7	52 sec.	1 min. 34 sec.
Task 8	51 sec.	1 min. 10 sec.
Task 9	25 sec.	21 sec.

Table 7. Mean task times by users in concurrent thinking aloud condition and condition without verbalisation.

As seen at the Table 7, the task times did vary over both conditions. Tasks 4 and 6, for example, suggested that users conducting thinking aloud condition did require more time to complete the tasks. In case of Task 6 the difference may be due overburdened cognitive capacity, as the question was longer than many other questions (nine words). However, this was not always the case, as users who were not prompted to verbalise their thoughts had longer task times in Task 2, Task 7, and Task 8. Although differences in task times occur, it must be noted that these differences may be due the small number of users conducting the experiment.

5. Discussion

The results of the present study did not support the initial suggestion of the experimenter that the total number of usability problems obtained in this study would be as large or larger than the number of problems observed by the usability engineering methods course. There are at least a couple of obvious reasons why this might be the case. First of all, the number of usability problems observed by the present experiment and the usability engineering methods course are not compatible, as the students tested 66 users whereas the present experiment tested only 8 users. Also, the experimenter was probably less experienced in spotting the problems than the teachers in the course who have performed usability work in practice.

The results indicated that neither concurrent or retrospective thinking aloud conditions was superior over the other in order to reveal usability problems of the site. It was noted that a majority of the problems (31) could have been observed with either technique, whereas only two of the problems were observed only in concurrent and 11 of the problems in retrospective thinking aloud conditions. This suggests that retrospective thinking aloud is slightly more powerful in revealing usability problems than the concurrent thinking aloud technique. No difference in the quality of problems (i.e. seriousness of the problems) was observed between the techniques.

However, gaze paths did offer additional information on users' behaviour. It was noted that in some cases several users did look at certain part of the page to find information. Users' retrospective verbal protocols did confirm that they had been doing so. Gaze paths also offered additional information to the verbal protocols. It was noted that users' eyes fixated to the search by number link, and therefore the experimenter argued that the link was oversized compared to the other links on the page.

Users were able to see their eye movements at the retrospective thinking aloud condition, and it did raise comments on what they had been looking for or what and how did they try to find at the moment. Some users did also give design suggestions during the retrospective verbalisation. Three out of four users did find the retrospective thinking aloud condition more pleasant than concurrent thinking aloud. All the users reported that they felt it 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.

The second part of the study investigated whether the usability testing with retrospective verbalisation using gaze paths as cue could be a valuable method to be used instead of regular usability testing with concurrent thinking aloud. Quality and quantity of produced words were compared between retrospective thinking aloud with gaze paths and concurrent thinking aloud. The results suggested that users did produce significantly more words during the retrospective verbalisation. A significant difference was also noted between the operational comment categories (i.e. manipulative, visual and cognitive comments), suggesting that the mean amount of words did vary significantly over word categorisations. This suggested that users made significantly more manipulative comments than visual comments and significantly more cognitive comments than manipulative comments. Lastly, a significant main effect was noted between concurrent and retrospective verbalisations and word categories. Further statistical analysis revealed that users did produce retrospectively significantly more comments in every operational comment category: manipulative, visual and cognitive than in concurrent condition. These promising findings indicate that users in the gaze path cued retrospective thinking aloud condition made more, and better quality (i.e. cognitive) comments than their counterparts in the concurrent thinking aloud condition. These cognitive comments are especially useful for the purpose of usability studies, as the comments represent the thought processes users had during the testing.

The tasks times were measured to investigate whether users spent more time to complete tasks when they were asked to verbalise their thoughts simultaneously. Users in concurrent thinking aloud condition spent more time to complete the task five times, whereas users without verbalisation required longer time in three tasks. The task time was same between both conditions in one task. Nielsen (1993) argued that verbalisation may affect the task times both ways: some users may find the concurrent verbalisation unpleasant and cognitively demanding, whereas some users may benefit from the verbalisation, as they find that thinking aloud makes the tasks clearer and easier to complete.

There may be several reasons behind the promising results. It was noted that the users were clearly more relaxed in the retrospective thinking aloud

condition than in concurrent verbalisation condition. This may be due to the fact that the users felt that they were no longer under observation as they were during the usability test. It is possible that they felt more relaxed as they were allowed to move more freely than during the eye tracking when they were advised to avoid unnecessary movements. As the results show, users felt freer to comment their actions and make interpretations, judgements or explain their behaviour during the retrospective verbalisation. They also were more likely to provide improvement suggestions for the web site.

Many users 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 their actions), the retrospective verbalisation took only approximately 20 minutes, thus the users were able to concentrate throughout the whole test session.

However, it must be noted that the present experiment has some flaws, which may have affected the results. First of all, the small number of users may have affected the results. The total number of words and operational comments did vary greatly among the users. The limited number of users also shows in tasks times, as one very slow user affects the average task time greatly.

The number of operational comments may also be affected due the time spent completing the retrospective thinking aloud condition. As the pilot studies showed that the gaze paths had to be viewed in half speed, the condition did last twice as long as the concurrent thinking aloud condition. Therefore the users had more time to comment on their actions. The analysis also revealed that in the retrospective thinking aloud condition the experimenter behaved in a more conversational manner than might have been appropriate. Both of these factors affected the number of words produced in retrospective condition, however, these flaws were taken into account when calculating the operational comments (i.e. comments not related to the experiment were not included).

The results suggested that retrospective thinking aloud condition using gaze paths as a cue was a useful technique for finding usability problems. The analysis of operational comments suggested that users did produce significantly more operational comments (i.e. manipulative, visual, and cognitive) in retrospective verbalisation than in concurrent thinking aloud

condition. It was also noted that users did find gaze paths as a useful memory aid and easy to follow and interpret. In addition to this, the quality of the Tobii eye tracker provided an easy and unobtrusive way to collect eye tracking data, not disturbing the users' attention and allowing the users to concentrate on the tasks. To sum up, the gaze path cued retrospective thinking aloud proved to be a useful technique to reveal usability problems and produce high quality verbal data.

The use of gaze path cued retrospective thinking aloud technique in usability testing seems rather promising. It has many advantages over concurrent thinking aloud, as users' attention is not drawn to the verbalisation and they may concentrate purely on the tasks. This is a benefit, as concurrent thinking aloud cannot be used in tests with expert users or children, as they are not able to verbalise their thoughts while they do the tasks. Expert users simply work too fast to be able to verbalise their thoughts, and for children verbalisation is cognitively too demanding. However, both of these user groups are growing as computers are getting more and more common. This notion may serve as a guide to future research that will examine the effects of factors such as expertise or age (children) on gaze path cued retrospective thinking aloud technique in usability testing.

6. References

Aaltonen, A., Hyrskykari, A. & Räihä, K. (1998) 101 Spots, or how users read menus? *Proc. of Human Factors in Computing Systems (CHI 1998),* ACM Press, 132-139.

Anderson, R.J. (1992). Representations and requirements: The value of ethnography in system Design. *Human-Computer Interaction,* 9, 152-182.

Andreassi, J. L. (1995). *Psychophysiology: Human Behavior and Psychological Response,* 3rd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Aula, A. & Surakka, V. (2002). Auditory emotional feedback facilitates human-computer interaction. X. Faulkner, J. Finlay & F. Détienne (eds.) *People and Computers XVI: Memorable Yet Visible, Proc. of HCI 2002,* Springer-Verlag, 337-349.

Ball, J. L., Eger, N., Stevens, R., Dodd, J. (2006). Applying the post-experience eye-tracked protocol (PEEP) method in usability testing. *Interfaces,* 67, 15-19.

Bloor, M., Frankland, J., Thomas, M., & Robson, K. (2001). *Focus Groups in Social Research.* London: Sage.

Boren, 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. *Proceedings of the Human Factors Society 34th Annual Meeting,* 1270-1274.

Capra, M. G. (2002). Contemporaneous versus retrospective user-reported critical incidents in usability evaluation. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting,* 1973-1976.

Cowen, L. (2001). An eye movement analysis of web-page usability. Unpublished Masters Thesis, Lancaster University, UK.

Dicks, R. S. (2002). Mis-Usability: On the uses and misuses of usability testing. *Proc. 20th Annual International Conference on Computer Documentation (SIGDOC 2002),* ACM Press, 26-30.

Dumas, J. S., & Redish, J. C. (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex Publishing Corporation.

Ericsson, K. A. (2002). *A protocol analysis and verbal reports on thinking.* http://www.psy.fsu.edu/faculty/ericsson/ericsson.proto.thnk.html (6.8.2006)

Ericsson, K. A. & Simon, H. A. (1984). *A protocol analysis: verbal reports as data.* Mass: MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). *A practical guide to usability testing.* Norwood, NJ: Ablex Publishing Corporation.

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., Lewenstein, M., Scott, N. & Wichansky, A. M. (2002). Eye tracking in web search tasks: Design implications. *Proc. of 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 J. Höynä, R. Radach & 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 of Human Factors in Computing Systems,* ACM, New York, 1253-1262.

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

Ilves, M. (2005). Ääneenajattelu. S. Ovaska, A. Aula, & P. Majaranta (toim.), *Käytettävyystutkimuksen menetelmät,* 187-208. Tampereen yliopisto, Tietojenkäsittelytieteiden laitos B-2005-1.

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

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the ACM SIGCHI'91,* New Orleans, 119-124.

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehending. *Psychological Review,* 87, 329-354.

Kantner, L. & Rosenbaum, S. (1997). Usability studies of WWW sites: Heuristic evaluation vs. laboratory testing. In *Proceedings of SIGDOC'97,* ACM, New York, 153-160.

Karn, K. S., Ellis, S. & Juliano, C. (1999). Workshop: The hunt for usability: Tracking eye movements. *Proc. of Human Factors in Computing Systems (CHI 1999),* ACM Press, 173.

Koskinen, J. (2005). Käytettävyystestaus. S. Ovaska, A. Aula, & P. Majaranta (toim.). *Käytettävyystutkimuksen menetelmät,* 187-208. Tampereen yliopisto, Tietojenkäsittelytieteiden laitos B-2005-1.

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

Lehtinen, M. (2004). Katseenseuranta. S. Ovaska, A. Aula, & P. Majaranta (toim.). *Käytettävyystutkimuksen menetelmät,* 223-236. Tampereen Yliopisto, Tietojenkäsittelytieteiden laitos B-2005-1.

Neisser, U. (1967) *Cognitive Psychology.* New York: Appleton-Century-Crofts.

Nielsen, J. (1993). *Usability Engineering.* New York: Academic Press, Inc.

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

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces, *Proc. ACM CHI'90 Conference*, 249-256.

Ovaska, S., Aula, A., Majaranta, P. (2005). *Käytettävyystutkimuksen menetelmät.* Tampereen Yliopisto, Tietojenkäsittelytieteiden laitos B-2005-1.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. W. & Newman, J. K. (2004). The determinants of web page viewing behaviour: An eye-tracking study. *Proc. Eye Tracking Research & Applications (ETRA 2004),* ACM Press, 147-154.

Partala, T. & Surakka, V. (2003). Pupil size variations as an indication of affective processing. *International Journal of Human-Computer Studies,* 59(1-2), 185-198.

Preece, J., Rogers, Y., Sharp, H. (2002). *Interaction Design: Beyond Human-Computer Interaction.* John Wiley & Sons.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T. (1994). *Human-Computer Interaction.* Harlow: Addison-Wesley.

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

Rubin, J. (1994). *Handbook of Usability Testing.* New York: Wiley.

Russo, J. E. (1978). Eye fixations can save the world: A critical evaluation and a comparison between eye fixation and other information processing methodologies. H. Keith Hunt (Ed.), *Advances in consumer research*, 5: 561-570. Ann Arbor, Michigan: Association for Consumer Research. http://forum.johnson.cornell.edu/faculty/russo/Eye%20Fixations%20Can%20Save%20the%20World.pdf [20.2.2007].

Simola, J. (2004). Silmänliikkeiden mittaus käytettävyystutkimuksessa. Adage Oy.
http://www.adage.fi/artikkelit/silmanliikkeiden_mittaus_kaytettavyystutkimuksessa.html. [20.2.2007]

Sinkkonen, I., Kuoppala, H., Parkkinen, J. & Vastamäki, R. (2002). *Käytettävyyden psykologia.* Helsinki: Edita, IT Press.

Tobii Technology. http://www.tobii.se [19.2.2007]

Van Someren, M., Barnard, Y., & Sandberg, J. (1994). *The Think Aloud Method. A Practical Guide to Modelling Cognitive Processes.* London: Academic Press. http://www.swi.psy.uva.nl/usr/maarten/Think-aloud-method.pdf (5.8.2006)

Wright, R. B. & Converse, S. A. (1992). Method bias and concurrent verbal protocol in software usability testing. *Proceedings of the Human Factors Society 36th Annual Meeting*, 1220-1224. Atlanta, Georgia.

Yarbus, A. F. (1967). *Eye movements and Vision.* New York: Plenum Press.

Appendix 1: Usability testing tasks

1. Mene sivulle autotalli.com
2. Kirjaudu sisälle palveluun. Tunnukset näet tässä (annettu käyttäjälle)
3. Selvitä montako Audi A3 merkkistä autoa on myynnissä Pirkanmaalla
4. Tarkenna hakua niin, että haet vain kuvalliset ilmoitukset
5. Vaihda salasanasi (annettu käyttäjälle)
6. Etsi kaikki ohjaustehostimella varustetut matkailuautot. Lue ääneen halvimman hinta
7. Vertaile näistä kahta uusinta vierekkäin
8. Etsi auto kuvitteellisesta lehti-ilmoituksesta saamallasi ilmoitusnumerolla 1103884. (numero annettu käyttäjälle)
9. Kirjaudu ulos palvelusta
10. Tarkastele etusivua, mitä mieltä olet

Appendix 2: Informed consent form

Käytettävyyslaboratorio
Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
33014 Tampereen yliopisto

ulab@cs.uta.fi
Käyntiosoite
PinniB h.1067-8
fax 215 6070

LUPA KÄYTETTÄVYYSTESTIN VIDEOIMISEEN

Toimin tänään testaajana käytettävyyslaboratoriossa järjestettävässä testissä, joka on osa Käytettävyyden arvioinnin menetelmät –kurssin harjoitustyötä. Testin järjestäjät ovat kertoneet minulle testitilanteen videoinnista ja testin järjestelyistä.

Testissä nauhoitettua videota käytetään ainoastaan testattavan sovelluksen käytettävyyden analysointiin. Materiaalia ei käytetä ilman erikseen pyydettävää lupaa muihin tarkoituksiin.

Annan luvan videointiin.

Käytettävyyslaboratoriossa ___. ___. 2005

Nimikirjoitus _____

Nimen selvennys _____

Rastita seuraava ruutu, jos annat luvan videomateriaalin esittämiseen Käytettävyyden arvioinnin –menetelmät kurssilla opetustarkoituksessa.

[ ] Testin aikana nauhoitettua videota saa näyttää muille kurssin osallistujille opetuksen yhteydessä.

Testin järjestäjät täyttävät

Järjestävän ryhmän nimi _____

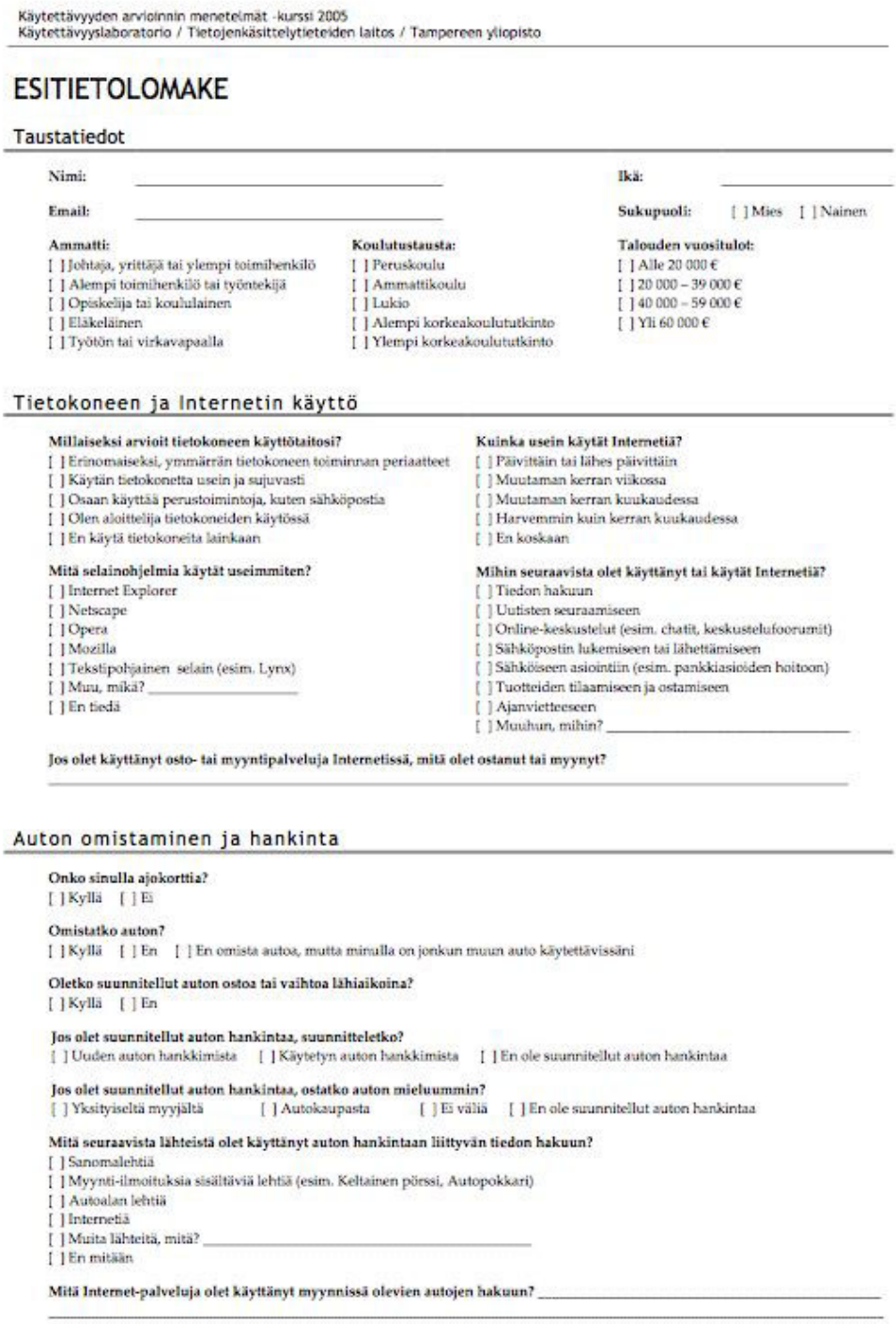
Testattu sovellus **autotalli.com**

Käyttäjä saapui kello _____ ja lähti _____

Nauhoituksen kesto noin _____ minuuttia

Videonauhan numero **1** **2** **3**

Appendix 3: Questionnaire of users' personal information



Käytettävyyden arvioinnin menetelmät -kurssi 2005
Käytettävyyslaboratorio / Tietojenkäsittelytieteiden laitos / Tampereen yliopisto

ESITIETOLOMAKE

Taustatiedot

Nimi: _____ **Ikä:** _____

Email: _____ **Sukupuoli:** [ ] Mies [ ] Nainen

Ammatti:
[ ] Johtaja, yrittäjä tai ylempi toimihenkilö
[ ] Alempi toimihenkilö tai työntekijä
[ ] Opiskelija tai koululainen
[ ] Eläkeläinen
[ ] Työtön tai virkavapaalla

Koulutustausta:
[ ] Peruskoulu
[ ] Ammattikoulu
[ ] Lukio
[ ] Alempi korkeakoulututkinto
[ ] Ylempi korkeakoulututkinto

Talouden vuositulot:
[ ] Alle 20 000 €
[ ] 20 000 – 39 000 €
[ ] 40 000 – 59 000 €
[ ] Yli 60 000 €

Tietokoneen ja Internetin käyttö

Millaiseksi arvioit tietokoneen käyttötaitosi?
[ ] Erinomaiseksi, ymmärrän tietokoneen toiminnan periaatteet
[ ] Käytän tietokonetta usein ja sujuvasti
[ ] Osaan käyttää perustoimintoja, kuten sähköpostia
[ ] Olen aloittelija tietokoneiden käytössä
[ ] En käytä tietokoneita lainkaan

Kuinka usein käytät Internetiä?
[ ] Päivittäin tai lähes päivittäin
[ ] Muutaman kerran viikossa
[ ] Muutaman kerran kuukaudessa
[ ] Harvemmin kuin kerran kuukaudessa
[ ] En koskaan

Mitä selainohjelmia käytät useimmiten?
[ ] Internet Explorer
[ ] Netscape
[ ] Opera
[ ] Mozilla
[ ] Tekstipohjainen selain (esim. Lynx)
[ ] Muu, mikä? _____
[ ] En tiedä

Mihin seuraavista olet käyttänyt tai käytät Internetiä?
[ ] Tiedon hakuun
[ ] Uutisten seuraamiseen
[ ] Online-keskustelut (esim. chatit, keskustelufoorumit)
[ ] Sähköpostin lukemiseen tai lähettämiseen
[ ] Sähköiseen asiointiin (esim. pankkiasioiden hoitoon)
[ ] Tuotteiden tilaamiseen ja ostamiseen
[ ] Ajanvietteeseen
[ ] Muuhun, mihin? _____

Jos olet käyttänyt osto- tai myyntipalveluja Internetissä, mitä olet ostanut tai myynyt?
_____

Auton omistaminen ja hankinta

Onko sinulla ajokorttia?
[ ] Kyllä [ ] Ei

Omistatko auton?
[ ] Kyllä [ ] En [ ] En omista autoa, mutta minulla on jonkun muun auto käytettävissäni

Oletko suunnitellut auton ostoa tai vaihtoa lähiaikoina?
[ ] Kyllä [ ] En

Jos olet suunnitellut auton hankintaa, suunnitteletko?
[ ] Uuden auton hankkimista [ ] Käytetyn auton hankkimista [ ] En ole suunnitellut auton hankintaa

Jos olet suunnitellut auton hankintaa, ostatko auton mieluummin?
[ ] Yksityiseltä myyjältä [ ] Autokaupasta [ ] Ei väliä [ ] En ole suunnitellut auton hankintaa

Mitä seuraavista lähteistä olet käyttänyt auton hankintaan liittyvän tiedon hakuun?
[ ] Sanomalehtiä
[ ] Myynti-ilmoituksia sisältäviä lehtiä (esim. Keltainen pörssi, Autopokkari)
[ ] Autoalan lehtiä
[ ] Internetiä
[ ] Muita lähteitä, mitä? _____
[ ] En mitään

Mitä Internet-palveluja olet käyttänyt myynnissä olevien autojen hakuun? _____
_____

Appendix 4: Questions asked at the interview

Ryhmä ääneenajattelulla

1. Mitä mieltä olet, onko oman toiminnan ja ajatusten kommentointi helmpompaa testin aikana vai sen jälkeen?
2. Kummassa tilanteessa oli helpompi kertoa ajatuksia?
3. Kumpi kommentointitilanne toi enemmän ajatuksia mieleesi?
4. Miten arvelet ääneenajattelun vaikuttaneen testitehtävien suorittamiseen?
5. Mitä mieltä olet katsepolusta, auttoiko se sinua palauttamaan mieleen ajatuksia?
6. Oliko katsepolkua vaikea tulkita?
7. Onko sinulla muita ajatuksia tai kommentteja?

Ryhmä ilman ääneenajattelua

1. Mitä mieltä olet katsepolusta, auttoiko se sinua palauttamaan mieleen ajatuksia?
2. Oliko katsepolkua vaikea tulkita?
3. Onko sinulla muita ajatuksia tai kommentteja