Concurrent and retrospective usability test methods in a realistic setting

Daalmans, B.J.

Award date:
2011

Concurrent and retrospective usability test methods in a realistic setting

A graduation report by
ing. Bo Daalmans

Technical university of Eindhoven
Faculty:
Industrial Engineering & Innovation Sciences
Supervised by
dr. Wijnand IJsselsteijn
&
dr. Wouter van den Hoogen

ABN AMRO Bank N.V.
Department: Internet
Supervised by
Tibor Sisarica
&
ir. Jennie Huijboom

22-09-2011

Content

Content 2
1. Introduction 3
1. Introduction 3
2. Theoretical frame 5
2.1. Usability testing methods 5
2.2. User testing 5
2.2.1. Think aloud usability test methods 6
2.2.2. Evaluating the use of concurrent and retrospective usability research methods. 7
2.2.3. Research location 9
2.3. Categorising usability problems 10
2.4. Using questionnaires in usability testing 11
2.5. Summary 13
3. Research questions 14
4. Method 16
4.1 Study design 16
4.2. Participants 16
4.3. Settings 17
4.4. Materials 18
4.5. Measurements 20
4.5. Procedures 21
4.6. Analysis 23
5. Results ..... **Fout!Bladwijzer niet gedefinieerd.**
5.1. Influence of usability Method & Location on the number of Usability problems found ..... **Fout!Bladwijzer niet gedefinieerd.**
5.1.1. Usability problems in Category 2.1 "User friendliness" ..... **Fout!Bladwijzer niet gedefinieerd.**
5.2. Influence of usability Method & Location on WEQ ratings **Fout!Bladwijzer niet gedefinieerd.**
5.3. Influence of usability Method & Location on qualities, beauty and goodness user ratings ..... **Fout!Bladwijzer niet gedefinieerd.**
5.3.1. Pragmatic qualities ..... **Fout!Bladwijzer niet gedefinieerd.**
5.3.2. Hedonic qualities – stimulation .. **Fout!Bladwijzer niet gedefinieerd.**
5.3.3. Hedonic qualities – Identification **Fout!Bladwijzer niet gedefinieerd.**
5.3.4. Beauty ..... **Fout!Bladwijzer niet gedefinieerd.**
5.3.5. Goodness ..... **Fout!Bladwijzer niet gedefinieerd.**
6. Discussion ..... **Fout!Bladwijzer niet gedefinieerd.**
7. Conclusion ..... **Fout!Bladwijzer niet gedefinieerd.**
8. References 54
Appendix 57
Appendix 1 57
Questionnaire 1: 57
Questionnaire 2: 59
Appendix 2 61
Appendix 3, instructions 63
Appendix 4, Uncovered usability problems, that where used in the comparison. 64

1. Introduction

People are increasingly using Internet for more and more aspects of their lives. The competition between organizations is rapidly moving from services in face to face situations to digital services that can be used regardless time and space. This development makes it very easy for customers to compare suppliers and to switch from one supplier to the other. Companies that can offer easy accessible products and services will win the battle for the client. In this high competition environment, the quality and usability of the company websites has become a critical success factor. To ensure the usefulness and user friendliness of a website or web product different research methods can be applied. The research field in which these research methods are evolved is called Human Computer Interaction (HCI).

Within the field of HCI user centred design is one of the major topics. In user centred design extra care is taken for the needs of the end user in each stage of the design process. Each stage has own methods to test the users needs. In early stages for example, paper prototyping is a useful test method. In the late stages focus groups can be used to see how the product is received by the users. Usability testing is especially useful for testing a product during the later stages of development. The main goal of usability testing is to uncover usability problems by the end user.

This research focuses on methods of usability testing as they are used in so called usability labs. Usability labs are artificial environments especially designed to conduct usability tests. In this research there is an emphasis on two kinds of usability test methods, Retrospective Think Aloud with Eye tracker (RTE) and Concurrent Think Aloud (CTA) methods. These methods will be extensively explained in the theoretical frame. The methods exist since 1980 when Ericsson and Simon conducted and reported on the first usability test. In this research we will look at the strength and weaknesses of these test methods and discuss when they can best be used.

Nowadays a lot of research is done in the field of usability testing and usability test methods. Many of these researches take place in a university environment, with students as users. Nielsen and Pernice (2010) argue in their book that academic papers are irrelevant for commercial design. They contend that the difference in setting influence the results of usability tests. The websites tested are generally non-commercial. In this research the relevance of the university research environment in usability test methods is tested, using a commercial website. The same usability tests are conducted in a laboratory environment at the Technical University Eindhoven, and in a more realistic environment, in the

usability lab of a large Dutch bank (ABN AMRO). The results of both environments are compared with each other.

Several questionnaires are conducted in order to measure the user's experience before and after the usability test. Besides the two locations, the usability tests are conducted with two research methods. The answers to the questionnaires and the results of the test give the experimenter a good view of the influence of method and location on the results and the user experience.

2. Theoretical frame

2.1. Usability testing methods

In the field of usability testing, a variety of methods can be used. These methods can be divided into six classes (Zhu, Vu & Proctor, 2005). These classes are:

- Expert methods – this class exists of tests in which an expert tests the web-product, for example in an expert review or a heuristic evaluation.
- User methods – this class tests involves users or potential users. Examples of test in this class are concurrent en retrospective think aloud methods.
- Prototyping – this class tests paper prototypes (lo-fi) or digital prototypes (hi-fi). This happens typically in a spiral way where prototypes are continuously are enhanced, corrected and tested.
- Observing – users are observed while using the product. This observation can be done by people but also by sensors that measure skin humidity etc.
- Interviewing and questionnaires – using different kinds of questionnaires by which the experience of users on a website can be measured. These questionnaires are often conducted online.
- Web-based methods – web tools analyse user behaviour and can detect user problems. Also the so-called A-B testing in which two variants of a site are simultaneously online for different users. A web analytic tool detects the effectiveness by measuring time or sales numbers.

In this research two of these six methods are used: user testing and questionnaires. The primary focus lies on the difference between two user methods, the Concurrent Think Aloud method (CTA) and the Retrospective Think Aloud method with eye tracker data (RTE). Two questionnaires are used to evaluate the influence of the two user test methods on users perceived qualities from the tested website.

2.2. User testing

User testing exists of tests in which users are asked to use the tested product. When users work with products, practical problems become visible. User testing is a process that often takes place in iterative design cycles and agile work processes. Problems that are uncovered during user testing can then be fixed; hopefully this results in a more user-friendly product. Different methods of usability testing are used. In this research we focus on think aloud methods.

2.2.1. Think aloud usability test methods

The most used method of usability testing with users is the CTA method as described by Ericsson & Simon (1980). In this method users simultaneously execute assignments while using the product and thinking out loud, commenting on the process and product. The test is being guided by an experimenter. Experimenters support the user though (different) tasks while observers make notes.

CTA is one of the most common usability tests and has evolved over the past 30 years. Still, few issues are related with the CTA usability test. The basics of the CTA method are designed by Ericsson & Simon (1980). On the basis of empirical studies, they concluded that CTA results are valid if conducted under strict conditions. The tasks may not be too difficult for the user; this could lead to a stop of the verbalization of the user's thoughts. But the task may also not be too easy, because then the user might perform the tasks somewhat automatically and thereby be unconscious of his or her thoughts. This can alter the verbalizations leading to biased observations. Also it is very important for the validity of the data that the experimenter does not intervene in the process and only stimulates the user to think out loud and continue the process.

An alternative think aloud method is the retrospective think-Aloud (RTA) (Eger et al, 2005). In this method users first finish their tasks in silence. Users are not asked to reflect on their progress and their experiences during the task, so they can concentrate on completing their task. When the tasks are finished, a screen recording of their performance is shown (Guan, Lee, Cuddihy & Ramey, 2006). On the basis of these recordings feedback and comments are given. This way the user can perform the task in a more natural way and give his feedback afterwards. This reduces the influence of concurrent verbalization on the process.

In some research the RTA methodology observations and verbalizations are split apart. In the first run observations of problems are made. In the second run the user verbalizes his issues and a second list of problems is found. Research suggests that verbalisations result in a larger amount of detected problems compared to only observations (Riel, 2007). In her research Riel found that adding verbalized problems led to more problem detections in the categories content and layout. For the problems categorised by navigational problems, the verbalizations did not seem to lead to more uncovered problems. Navigational problems are easier to observe.

A relatively new method is Retrospective Think-Aloud with Eye tracker (RTE); this method is quite similar to the RTA method. The only difference between the methods is that in the RTE the playback video also shows the eye movements of

users (Eger et al, 2007). The eye movements are first recorded using an eye tracker, and then added to the video. The added value of these movements is to help users remember what they where looking at on every moment of the video. The term RTE is not used in all literature; in some literature the RTE method is part of the RTA method and are named the same.

2.2.2. Evaluating the use of concurrent and retrospective usability research methods.

In this chapter the user usability methods CTA and RTE are compared. Previous research (Haak et al, 2006; Haak, et al 2007; Haak, et al 2009; Riel, 2007; Eger et al, 2007; Burg, 2008) gives us good ideas about the pro's and con's of these methods. A definite best research method however has not yet emerged.

Problems with CTA

A problem connected to the use of the CTA is the fact that the natural task performance is interrupted by questions of the experimenter and comments stated during the test. Users think more about steps to take while participating in a CTA test (Eger, Ball & Dodd, 2007). This leads to a distorted view of the users advancements. This distortion is due to the fact that some problems can be caused or prevented by unnatural behaviour of users or influenced by the questions of the experimenter. Synchronically thinking out loud can change users thoughts into a more structured way, which can change the users performance in either a good or a bad way (Haak et al 2003; Eger et al, 2007).

Benefits of RTE over CTA

In their research Haak et al, discussed a lot of positive properties of retrospective research methods in user usability testing (Haak, et al, 2003). RTE is a retrospective method, developed to let the test user perform the assignments of the usability test as natural as possible. Users can perform tasks at their own pace. This way, the result reflects the natural process better compared to the CTA condition. Using retrospective methods, users have some time to reflect on their experience. This can lead to insights of individual problems and bigger structural problems. Also walking through a system a second time, gives other opportunities to discover additional potential problems.

Another big plus of retrospective methods is that the time it takes to complete a task is measured in a representative way (Nielsen & Pernice, 2010). In contrast, the execution time measured in synchronic Think Aloud methods as CTA, can hardly be called representative for real task execution time. This is due to the interference of the experimenter and verbalization of steps taken while tasks are executed. The time it takes to complete a task in a natural way can be used to

detect and define problems and compare designs. In retrospective methods, the product (web site) is used in a more natural way next to the CTA method, the time recorded in retrospective methods can be used as problem identifying data. A long execution time for a small process indicates a problem.

Problems with RTE

Retrospective methods have the downside that the tests take longer, due to the fact that everything has to be run twice, first to do the task and later to comment on the process (Haak et al, 2003)(Eger et al, 2007). Users might get more tired during a long test and attention can fade away. It is important that test users give a lot of feedback and tell their thoughts about the system. When being more tired, users do not remember or verbalize problems so well.

Another weakness of retrospective methods is that users can forget their initial problems during the task (Haak et al, 2003). It is hard to remember all the problems and questions encountered. This is the reason why video and eye tracking are used in the newer retrospective usability methods. A video of the performance of the task helps users to remember their thoughts and problems they encountered.

In retrospective think aloud method verbalizations can be inaccurate due to the lack of ability of users to observe and verbalize their own mental processes, especially thoughts (Eger e.a., 2005; Guan e.a., 2006; Haak e.a., 2006).
People think faster then they can speak (Cooke, 2005). This leads to information loss in both methods. When people are speaking about one problem, another passing problem can slip their mind and will never be verbalized. In the CTA method a user can stop with their tasks to verbalize his thoughts and then continue. In the RTE condition users barely pause the recording, even if they have the possibility to do this.

Riel (2010) concludes that the eye movements shown in RTE confuse the users. The researcher suggests that this is because users encounter their eye movements for the first time, distracting them from their main task, verbalizing their thought and encountered problems.

Which is the best method for uncovering usability problems?

Researchers state that eye movements provide more insights into search processes and explore more usability-problems compared to the CTA method without eye-tracking data (Eger, et al, 2007). Based on this research, Eger e.a. concluded that RTE and RTA have methodological benefits compared to the CTA methodology in usability testing. The difference found in this research could as well have been caused by the method as by the added value of eye tracking data

and observations during the research. The CTA method used in the research of (Eger, et al, 2007) is without the use of eye tracking observations. In other research (including this one) the CTA method also contains eye tracking observations.

Haak et al (2009) uncovered more usability problems in the retrospective method then in the CTA method. They conducted a research between three different research methods of which one was the CTA method and one was a retrospective method. With 80 students conducting five assignments on multiple websites, they compared their results with their previous research conducted in 2003 (Haak et al, 2003).
In 2003 the test was conducted on only one website (a library catalogue). This research was conducted with 40 students. They got specific tasks that where quite complex. Result of both test resulted in the same conclusion. Although in the retrospective methods slightly more problems where uncovered, both researches found no significant difference in problem detection per method. This means that the differences in uncovered problems could merely be chance. In the CTA method generated significantly more suggestions and explanations for problems.

In final thesis researches dedicated to find differences between user usability methods, no significant distinctions where found in problem detection capabilities of usability test methods. (Gombert, 2009; Haas, 2009; Nell, 2009). Differences in the way problems are found where detected. More problems are observed by observation in the RTE method, while more problems were observed by user verbalizations were found in the CTA method.

2.2.3. Research location

So far most researches has been conducted with students in a university environment (Haak et al, 2003; Haak et al, 2006; Haak, et al 2007; Haak, et al 2009; Riel, 2007; Burg, 2008; Pothuis, 2009). User usability test can be conducted in many locations. Usability test can be conducted at user's homes or at their workspace. These are called field studies. Tests can also be done in a usability lab at bigger companies, specialized usability research bureaus, or at universities. These are called laboratory studies.

Nielsen & Pernice (2010) state in their research that academic papers are irrelevant for commercial design. They argue that the research environment & the research participants of research in a university influence the results. A field study can be more representative since the natural environment in which a product is used differs from a laboratory environment. Laboratory environments

have the benefits that most environmental factors can be controlled. The differences found in a variable can be appointed to the manipulation (Nielsen & Pernice, 2010).

In previous research Kaikkonen, et al (2005) researched the number of usability problems found in a laboratory study and a field study on location with a mobile phone application. With 20 users the number of problems found did not differ, however three problems where observed more often in the field study.

Another difference between usability test locations is the usage of the rules. In traditional usability testing the experimenter is only allowed to ask a limited number of questions. In a realistic setting there is not enough time to perform the tests with 40 users. The priority is focused on major problems and on alternative ideas for design problems (Kaikkonen, et al, 2005).

A difference between two laboratory research environments (big company versus university) has to do with the psychological perspective. People are tended to give socially desirable responses (Crowne & Marlowe, 1960 and Ross & Mirowsky, 1984) while being in a building of a bigger company with its own usability research laboratory. Positive remarks about this company can be considered as socially desirable. People may give more positive answers concerning ABN AMRO when in an ABN AMRO environment.

2.3. Categorising usability problems

In this research two research methods are compared with each other. In order to compare the results of the two usability tests the detected problems are first categorised. This way the power of the comparison increases and the test results can be compared to other literature for they also categorised their results in the same categories (Cooke, 2005; Haas, 2009; Riel, 2010).
Elling, Lentz & De Jong (2007) describe three main categories: navigation, content and layout. The three main categories in the research of Elling, et al (2007) are divided into sub categories. These categories and sub categories are designed to evaluate websites on different levels. The evaluation is done in a questionnaire called the web evaluation questionnaire (WEQ). In the usability researches mentioned above the categories are used to define and group the usability test results (detected problems). The categories defined are shown in table 1.

Table 1, Main and Sub Categories (WEQ)

Main Category	Sub Categories
Navigation:	1, Ease of use – to what extent does the user experience the website as being easy to use? 2, Structure – in it easy to find one's way through the website? 3, Hyperlinks – to what extent is the title of the link self explanatory, and does the user understand it. 4, speed – to what extent does the user experience the speed as slow or fast. 5, Search engine – to what extent does the search engine help to succeed in finding the correct information.
Content:	1, Relevance – in what way does the provided information correspond to the requested information. 2, Comprehensibility – the way the information is understandable. 3, Comprehensiveness – the way the information is experienced as complete.
Layout:	1, layout – experience of colour usage, backgrounds, frames and pictures.

2.4. Using questionnaires in usability testing

The WEQ questionnaire itself gives detailed insights on how users experience the quality and user friendliness of a website (attachment 2). It would be interesting to combine usability testing with questionnaires in order to get more feedback on the experienced quality of the tested product/website. So far no research has been conducted on the effect of different usability test methods on the ratings in this kind of questionnaires. Using these questionnaires would for instance be extra interesting when two designs are compared with each other using a usability test. In some research questionnaires are used in order to measure users experience during the test (Haak et al, 2003 & Haak et al, 2009). No questionnaires have been conducted specifically on how participants/users experience the tested product/website.

When questionnaires are used to support user usability testing, Hassenzahls questionnaire would be a good addition. In the Hassenzahl questionnaire not only usability is tested, but also visual appeal. Hassenzahl is one of the lead researchers in the field of hedonic qualities in digital products. In a lot of papers between 2000 and 2007 he designed and refined questionnaires to measure the usability and other aspects of a digital product. He argues that not only the usability of a product but also the visual appeal leads to the quality of a digital product. Hassenzahl states that products with a higher visual appeal motivate

users to try harder to use the product, in this way increasing the products usability and usefulness.

Hassenzahl (2004) uses measurements of Pragmatic quality (PQ), Hedonic qualities; stimulation (HQS) and identification (HQI), beauty and goodness in order to test digital products. The different qualities are explained.

Pragmatic quality (in early research called ethnographic quality) exists out of the quality related to traditional usability, namely efficiency and effectiveness. A website that is easy to use and is useful gets a high usability rating and thereby has a high PQ rating.

The visual appeal of a product is called Hedonic quality (HQ). In later research two kinds of Hedonic qualities emerge: the Stimulation (HQS) and the Identification (HQI) variant. Hedonic qualities exist out of aspects less obviously related to the beauty or goodness of a website. Originality, innovativeness and beauty are factors influencing the HQ.

A product with a high Hedonic quality – stimulation (HQS) rates the level to which the user is stimulated to use the product and keep using the product. The HQS rating depends mainly on the feeling a product gives the user. An original modern product will receive a high HQS rating.

Hedonic quality – identification (HQI) addresses the human need to express one's self. Using objects including digital products one can achieve their desired self-presentation. This is relevant because individuals want to be seen in specific ways by relevant others (Hassenzahl, 2004).

Beauty and Goodness are measured in order to get a complete view of someone's opinion on the digital product.

Hassenzahl summarizes the different qualities as:

> *A product can be perceived as pragmatic because it provides effective and efficient ways to achieve behavioral goals. Moreover, it can be perceived as hedonic because it provides stimulation by its challenging and novel character or identification by communicating important personal values to relevant others.* (Hassenzahl, 2004, P 322)

2.5. Summary

The most commonly used method in usability research is the Concurrent Think Aloud (CTA) method (Eger, et al, 2007). Research about usability tests suggests that this might not be the best method. Since the existence of usability tests, more different test methods have been developed. Retrospective methods have been developed to counter some of synchronic think aloud method problems. The essence of retrospective methods is to let the user perform the tasks in a natural way. One of the new retrospective methods is the "retrospective Think-Aloud with Eye tracker" (RTE). This method could be developed due the existence of eye trackers. Differences between the results of these methods are minimal.

The location where a usability test is conducted can have an influence on the test results. Although hardly anything is known about the differences between different laboratory settings in usability testing.

Questionnaires could be beneficial in usability testing. Using questionnaires in combination with user usability testing would give a greater spectrum of results. So far no research has been conducted to test the influence of usability testing on questionnaire results.

3. Research questions

Several researches have been conducted to compare usability research methods. All these researches are based on information gathered from one or two different websites. Only few minor differences between different methods have been obtained. All researches have been conducted by external usability experts, in a laboratory environment and cover entire websites.

The scope of the research is directed towards specified usability tests as they are being used within big companies. The content of the tests is reduced to specific parts of The ABN AMRO website. Two test methods are compared (CTA and RTE) and both methods are conducted in two different laboratory environments: a university environment and a big company environment. Results from the different test locations will help to put results into perspective. The research question is the following:

What is the influence of usability test methods and location in usability testing?

This question is answered by answering the following sub questions:

1) What is the influence of usability test method on the number of usability problems uncovered per category, depending on usability test method and location?
2) What is the influence of usability test method and location on the WEQ results per category?
3) What is the influence of usability test method and location on the Quality questionnaire results?

Hypotheses, sub question 1

What is the influence of usability test method on the number of usability problems uncovered per category, depending on usability test method and location?

I hypothesize that the RTE method will find more problems in the navigational category compared to the CTA method. This hypothesis is based on the suggestion of Nielsen and Pernice that users in a laboratory setting who verbalize their thoughts are more stimulated to structure them, and by doing that find their way better compared to users in a realistic setting. (Nielsen & Pernice 2010)

Hypotheses, sub question 2

What is the influence of usability test method and location on the WEQ results per category.

I hypothesize that the ABN AMRO location will have a positive effect on all category ratings compared to the university location. The usability lab is located within the ABN AMRO dialogue house. I expect that users will give more socially desirable answers in the company environment, resulting in higher scores in the WEQ.

Also I hypothesize that the CTA and RTE conditions will have no influence on the WEQ ratings.

Hypotheses, sub question 3

What is the influence of usability test method and location on the Quality questionnaire results?

I hypothesize that the ABN AMRO location will have a positive effect on all quality ratings compared to the laboratory environment. This for the same reason as discussed in sub question 2.

Also I hypothesize that when participating in the CTA method users will reflect on their own progression, therefore they will experience the usability problems more conscious. This can result in the CTA method having a negative effect on the PQ rating.

4. Method

4.1 Study design

Normally a usability test is conducted with about 5 to 8 users. Nielsen (1994) found that with about five users 85% of the usability problems is found, and with 8 users around 95% of the usability problems is detected. In this study 50% percent of the tests are conducted in the CTA condition, the other 50% in RTE condition. In both test the same product is tested: existing parts of the ABN AMRO site. Calculations on questionnaire results are done with a total of 48 users.

In order to be able to compare the results per method, per location and with other literature, the detected problems are categorized. When problems are categorised a better comparison can be made because differences in results can be compared per category. The categories are divided in three main categories. These are: accessibility, content and design. Definitions of these categories can be found in appendix 1.

The problems detected in each setting (Company vs university and CTA vs RTE) are analyzed with an ANOVA with Method and Location as two fixed factors. The research is conducted with two dependant variables (location, method), resulting in four different conditions.

Generating the four conditions:

- ABN AMRO, CTA method (F_CTA)(4p)
- ABN AMRO, RTE method (F_RTE)(4p)
- TU/e, CTA method (L_CTA)(20p)
- TU/e, RTE method (L_RTE)(20p)

4.2. Participants

Users in the realistic environment were arranged by an external bureau. They were paid 40 euro’s to participate in the test for 1,5 hours. Users in the laboratory environment were contacted by mail through a database of the Technical University Eindhoven, with students that participate in experiments in order to earn some money. They received 15 euro’s for their participation.

Table 2, the general properties of the samples

General profile	ABN AMRO	TU/e
Average age	25.8	22.3
Male\Female\%male	6 \ 2 \ 75%	25 \ 15 \ 63%
Internet usage	100% Every day	100% Every day
Customer of ABN AMRO	25%	17%

Table 3, the educational level of the sample

Average educational level (highest finished study)	ABN AMRO	TU/e
HAVO	0	4
VWO	1	20
MBO	1	0
HBO	1	9
WO	5	7

The average of educational level is slightly different in both environments. This is because the highest finished study is asked. The users in the TU/e environment were mostly students who had not yet finished their studies. Therefore the actual educational level is higher then the measured value.

4.3. Settings

The research has been conducted at two locations. The realistic environment is located in one of the national offices of ABN AMRO Bank N.V. located at the Foppingadreef. The other environment is located on the university terrain of the Technical University Eindhoven in the game lab.

The usability lab at ABN AMRO consists of two rooms. The test lab is a room looking like a regular room/office as in which one would use the products that are being tested. The second room is the observation room. Between these rooms is a wall of one-way mirrors, allowing researchers in the observation room to study the behavior of users in the test lab without being noticed. This way, direct observations can be done without influencing the user. Also camera's, screen duplicators and sometimes eye-trackers are available in order to maximally observe the user.

Figure 1, usability Lab setup ABN AMRO

Within ABN AMRO the usability lab (see figure 1) is used to test web applications and websites. Most tests are run by a usability expert using the CTA method. Different tasks/scenarios are created by the usability expert. Using these tasks/scenarios, the usability expert guides the test person though the product. The designer and project leader are present in the observation room making observations.

The game lab on the university location is a little different. The lab also exists of two rooms. One looks like a living room / office. This is where the user makes the assignments. The other room is located next to the first room but is not separated by a one-way mirror, but by a wall.

4.4. Materials

The tests analyzed in this research are conducted in two different environments. In both environments the setup is quite similar, and both locations use a Tobii eye tracker device.

In figure 2 is shown the schematic setup of the ABN AMRO Usability lab. Figure 5 shows the setup of the TU/e usability lab.

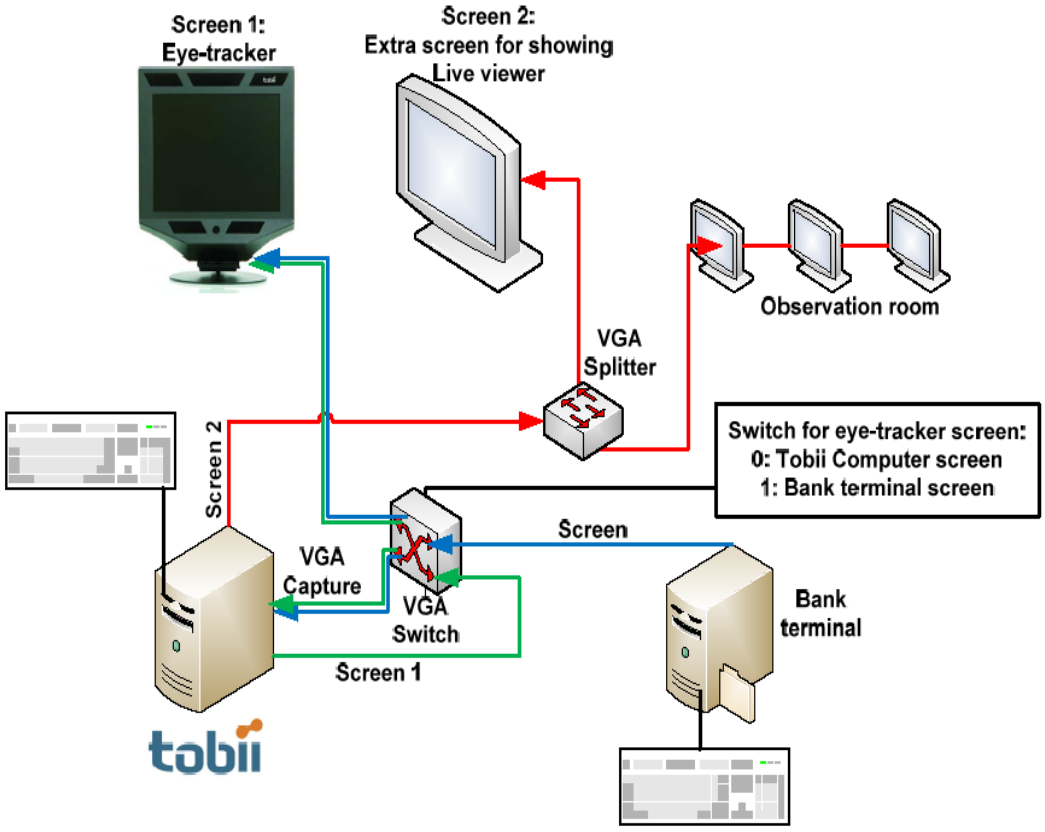


Figure 2, The schematic setup of the ABN AMRO Usability lab

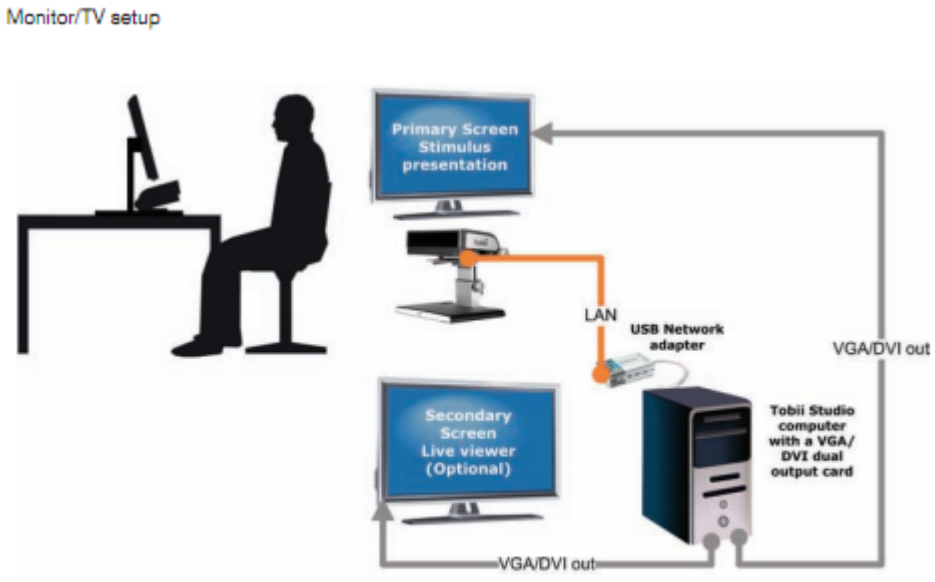


Figure 3, The schematic setup of the Technical University Eindhoven Usability lab

The secondary screen in the university setup was split in two. One screen was located in the test lab, and one in the observation lab. This way eye tracking data could be observed in both test methods.

4.5. Measurements

Normally in usability test research an ideal task model is made. This model defines per task a maximum amount of time and an ideal path. Every time a user deviates from this model, either in path or duration, this is an indication that a problem occurred. In this research this is not possible. The ABN AMRO website contains many navigational possibilities. An ideal task model is impossible to create for there are multiple paths that are all equally efficient.

In some literature the amount of time is measured to detect usability problems (Haak et al, 2006; Haak, et al 2007; Haak, et al 2009; Riel, 2007; Burg, 2008). When a user needs more time to complete a task, this indicates a problem. In this research the CTA method is compared to the RTE method. One of the differences of these methods is the representativeness of time (Nielsen & Pernice 2010). In the CTA method users take time to explain what they see and what they are doing, where in the RTE method users just keep on working. Therefore in this study we decided not to use time as a problem indicating tool.

In this research problems are observed during the test or while reviewing the test recordings. To make sure the observed problems are reliable a inter-rater reliability test is conducted. More about this test can be read in the observation part below.

Both test methods are recorded, not only by the Tobii studio software, but also by a screen recording tool. This means that in the RTE method not only the initial task is recorded, but also the review phase. This is important because this is the part where the verbalized data are generated and gathered.

In total 19 problems where detected (see appendix 4). These 19 problems are divided into three main categories: Content(8), Navigation(10) and Layout(1). Main categories Content and Navigation are split in sub categories. The content category is split up into subcategories; Relevance (0), Comprehensibility (4) and Comprehensiveness (4). The Navigation category is split up into subcategories; Ease of use (4), Structure (4), Hyperlinks (2), Speed (0) and Search options (0). Per subcategory the number of observations are added. The added number of problems are used in the calculations. The number of problem observations per subcategory are discussed in the result section.

The Hassenzahl questionnaire is used (designed by Hassenzahl, 2007) to measure the users attitude towards the different qualities of the website, before and after the test. In this research either four or seven questions where asked on a 7-point scale. In total 20 questions are used in the questionnaire. Using a factor analysis the measurements that have a negative influence on the main effect are removed (appendix 1). In total 17 questions are used to measure the different qualities. The qualities measured are; PQ (6), HQI (4), HQS (5), Beauty (1), Goodness (1). The average score of each quality is used in the calculations. These scores where rated from 1 to 7.

The web evaluation questionnaire (WEQ) as designed by Elling, et al (2007) is also conducted after the user finished the tasks. In this questionnaire a total of three to four questions where asked per sub category to measure users attitude toward the website quality on these subjects.. In the questionnaire exists 28 questions measure main and sub qualities. Main categories: Content (9), Navigation (16) and Layout (3). Main categories content and navigation are split in sub categories. The content category is split up into subcategories: Relevance (3), Comprehensibility (3) and Comprehensiveness (3). The Navigation category is split up into subcategories; Ease of use (3), Structure (4), Hyperlinks (4), Speed (2) and Search options (3). Questions are rated on a 5-point scale. The questions formulated in a negative way are recoded so that all questions are rated as 1 to be most negative, and 5 to be most positive. Averages per category and per subcategory are calculated.

4.5. Procedures

In both environments only predefined comments are used on predefined moments depending on the method used.
In the CTA and RTE method there are different protocols for the behaviour of the experimenter. In the CTA method his task is to ensure the user keeps on thinking aloud and sometimes show him what to do if the task takes too long. In the RTE method the experimenter leaves the room when the assignments are conducted. He observes the progress in a different room. After all the assignments have been completed, the tasks are discussed using a screen and the eye movement recording.

In this paragraph the procedure is explained step by step. Every step in the process has its own number. Since the procedure changes for every condition the steps are split in two. These steps also have a letter.

1a. Users arrive at the ABN AMRO building by car or public transportation. They have to ring a bell and a receptionist invites them in. The user goes through a

revolving door up a stairway and arrives in the "Dialogue house". Here they have to report to the reception. From here they will be escorted to the waiting area. When it is time to begin the test the experimenter welcomes the user and brings him / her to the usability lab.
1b. Users arrive at the specific University building and wait in a study/break area where they will meet the experimenter. When it is time to begin the test the experimenter welcomes the user and brings him / her to the game lab.
2. After the user enters, he/she is explained what will happen the next one, to one and a half hour. Haak et al (2003) used a Dutch explanation to prepare the users for the usability test. This instruction was used in a slightly altered Dutch introduction. The instructions can be found in appendix 3.
3. The user is shown the ABN-AMRO website homepage for 40 seconds.
4. Now they are asked to fill in two questionnaires. The first one is about their mood influenced by the first impression of the website. This way the mood is measured at the moment the first questionnaire is conducted. The second questionnaire measures the first impression of the website per quality as described by Hassenzahl.
5a. (CTA condition) The user is given a few tasks to fulfil on the website. While concentrating on these tasks the experimenter sits next to the users stimulating them to keep thinking out loud. A set of sentences have been listed to stimulate the user to keep thinking out loud and to further explain their comments. The questions are derived from several usability tests the researcher conducted and observed. The questions do not influence the users reply towards a specific answer and the responses give a good insight into the usability problem. The comments were used if the user was not reading and did not say anything for about 5 seconds. They can be found in appendix 3.
5bI. (RTE condition) The user is given a document with several tasks, and a document to fill in the answers of these tasks. The users first individually fulfil the tasks in silence, while the experimenter is observing him/her from the observation room.
5bII. (RTE condition) Using a recording of the task process, including the mouse movements and eye movements, the process is reviewed together with experimenter and users. Generating verbalized feedback on the process and encountered problems. The user is asked to describe what he was doing and what problems he encountered. The exact explanation can be found in appendix 3.
6. After doing the assignments two questionnaires have to be completed. The first one is the same questionnaire asked before doing the tasks. The second questionnaire is the WEQ questionnaire.
7a. (ABN location) The user is thanked for his/her participation and is escorted to the exit.
7b. (TU/e location) the user has to sign a payment paper, receives €15,-. The user is thanked for his/her participation and is escorted to the exit.

4.6. Analysis

Three types of data are analyzed. First a document is formed with all the usability problems discovered. The problems that where uncovered only once, are usually based on users individual preferences. If only one of the 48 users has difficulties understanding some interaction, this does not necessarily mean there is a problem. In usability testing a problem is not defined by one user making a mistake, but by users systematically making the same mistake. For that reason the definition of a usability problem is: problems that occur more than once. All the problems that were encountered only once in the 48 usability tests where removed. Burg (2008) reports the same procedure. By following this procedure the results are comparable with other researches. The remaining problems (see appendix 4) are used to compare the outcome of the different tests methods, focussing on differences in location and / or method.

The problems found are categorised into three main categories. Elling, Lentz & De Jong (2007) defined these three different main categories in which a website can be judged. These categories are layout, content and navigation. In the theoretical frame the categories are described more thoroughly. By using these categories more precise comparisons can be made between the results of this research and the results of previous research.

In order to categorize the usability problems three people categorized the 19 usability problems individually. All problems where allocated into the same main categories by the three persons. Allocating the usability problems to the subcategory classification lead to a discussion. One usability problem was divided to subcategory 1.3 but could also be divided into in subcategory 1.2. Two problems where divided into subcategory 2.3 that could also be put in subcategory 2.5.

4.7. Observations

In order to test the reliability of the observations an inter-rater reliability test was conducted. In this test the 19 usability problems as discussed in appendix 4 were rated by a second observer. The problems were described really thorough. The problems could be observed on only one occasion in the test, and were either observed or not observed. The recordings were numbered. The problems observed were compared for the same users. The second observer looked at four of the same recordings as the experimenter (first observer). Two of the recordings were from a CTA method, and two where from the RTE method. In this test only the problems as observed by observer 1 where used. The results of two independent observers are shown in table 4. Using this table Cohen’s Kappa could be calculated. In this calculation the chance that users observe the same

problem by chance is weighted and cancelled out. From these results Cohen's Kappa is derived and is found to be 0.83. This is a very respectable result leading to the conclusion that the observation method is reliable.
Hereby we note that one of the usability problems could be observed during multiple stages of the test. In the four test compared this problem did not occur, and therefore does not influence the outcome.

Table 4, Inter-rater reliabilty test results

		Observer 1		
		Problem detected	No problem detected	Total
Observer 2	Problem detected	27	5	32
	No problem detected	3	41	44
	Total	30	46	76

5. Results

The result section is divided into three parts. In the first part the influence of method and location on the detection of usability problem categories are described. In the second part the influence of method and location on the WEQ results are described. In the third part the influence of method and location on the different perceived qualities are described.

5.1. Influence of usability method & location on the number of usability problems found

The usability problems are observed during the usability tests and during the recorded observation. In total 19 unique usability problems are observed and listed. These usability problems have been divided into three main categories and six subcategories. The data are prepared as discussed in the measurement section. In table 5, the number and average number per user, of usability problem observations per category and subcategory are shown.

Table 5, Average number of usability problems found per category per method

Category	CTA Method		RTE Method	
	Number of problems (Average per user)	St.dev	Number of problems (Average per user)	St.dev
1, Content	76(3,167)	1,37261	87(3,625)	1,58286
1.2, Comprehensibility	29(1,208)	,58823	35(1,458)	,83297
1.3, Comprehensiveness	47(1,958)	1,12208	52(2,167)	1,09014
2, Navigation	117(4,875)	1,48361	105(4,375)	1,20911
2.1, User friendliness	56(2,333)	1,00722	35(1,500)	,93250
2.2, Structure	38(1,583)	,97431	43(1,792)	,83297
2.3, Hyperlinks	23(,958)	,20412	26(1,083)	,40825
3, Layout	6(,250)	,44233	2(,083)	,28233

In order to measure whether method and/or location have an effect on the user ratings, a two-way ANOVA on each usability problem category and subcategory, with two between subject factors (method and location), is conducted. The results are shown in Tables 6, 7 and 8.

Table 6, Tests of Between-Subjects Effects (method)

Usability problem category and subcategory	F	Sig.
1, Content	,049	,827
2, Navigation	,321	,574
3, Layout	,489	,488
1.2, Comprehensibility	,291	,592
1.3, Comprehensiveness	,421	,520
2.1, User friendliness	5,526	,023
2.2 Structure	1,430	,238
2.3 Hyperlinks	1,974	,167

Table 7, Tests of Between-Subjects Effects (location)

Usability problem category and subcategory	F	Sig.
1, Content	,701	,407
2, Navigation	,722	,400
3, Layout	1,956	,169
1.2, Comprehensibility	2,068	,158
1.3, Comprehensiveness	,031	,860
2.1, User friendliness	,273	,604
2.2 Structure	,045	,834
2.3 Hyperlinks	1,974	,167

Table 8, Tests of Between-Subjects Effects (Method * location)

Usability problem category and subcategory	F	Sig.
1, Content	2,380	,130
2, Navigation	,321	,574
3, Layout	,489	,488
1.2, Comprehensibility	,291	,592
1.3, Comprehensiveness	2,923	,094
2.1, User friendliness	,068	,795
2.2 Structure	,836	,365
2.3 Hyperlinks	,363	,550

One significant main effect is found (table 6,7) and no interaction affects are found (table 8). A significant effect between method and the number of usability problems found in subcategory 2.1 is found ($F(1,44)=5.53$, $p=.023$). The result shows that the CTA method is more successful in uncovering usability problems in subcategory 2.1, compared to the RTE method. The effect is shown in figure 4.

Table 9, Average number of usability problem found per method accounted for the effect of location.

Usability problem subcategory	Method	Mean	Std. Error
2.1, User friendliness	CTA	2,300	,271
	RTE	1,400	,271

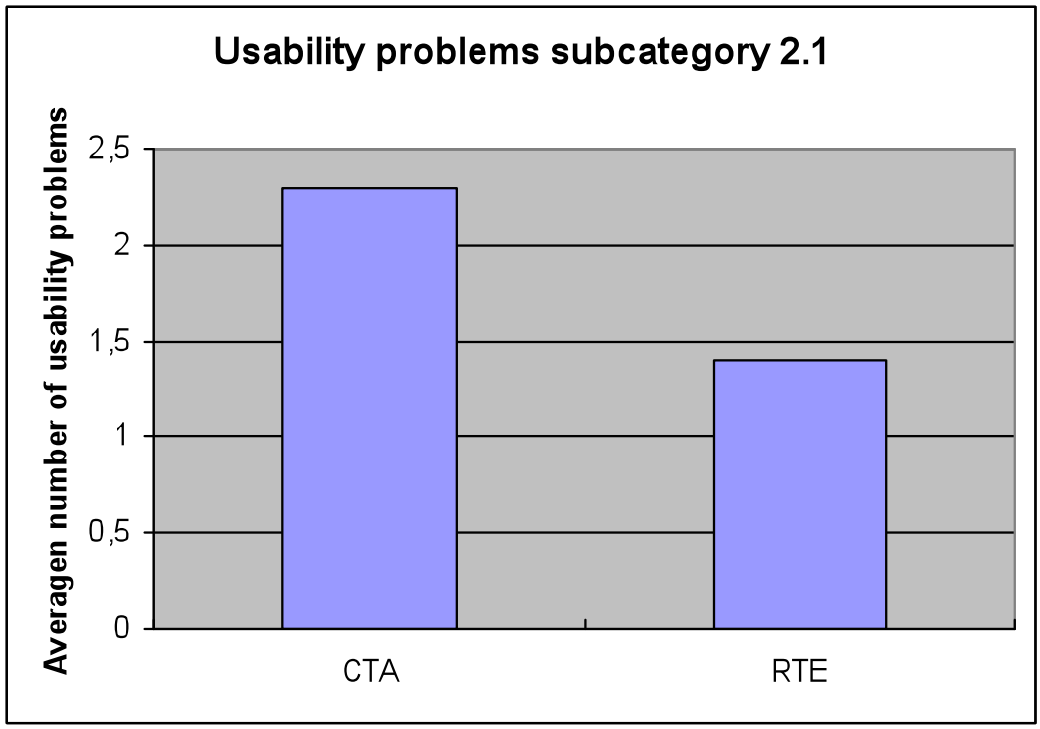


Figure 4, Average number of usability problem found per method accounted for the effect of location.

5.1.1. Usability problems in category 2.1. "User friendliness"

The CTA method is better in detecting usability problems in sub category 2.1. "User friendliness". This is the only category where the number of problem detections is significantly influenced by method. In order to get a better understanding on the influence of the method on usability problem detection the usability problems in this category are researched. Usability problem category 2.1. "user friendliness", contains four individual usability problems. Using Chi square tests differences in usability problem detection per method are tested.

5.1.1.1. Usability problem 1: It is difficult to interact with the main menu

Problem description: Users have problems using the main navigation. They click on items when they should not, and they expect a mouse over function when they should click.

With a chi-square test no significant difference between the problem detection in both methods is found $X^2(1, N=48)=1.231$, p=.267.

Table 9, Crosstabulation of method and Usability problem 6

		Usability problem 1		
		No	Yes	Total
Method	CTA	18	6	24
	RTE	21	3	24
	Total	39	9	48

5.1.1.2. Usability problem 5: A hyperlink on the website is hard to find.

Problem description: There is a hyperlink that links to a page with an overview of the maximum amount of Euro's the particular insurance will cover. For some users this link is hard to find. They either don't see it, or don't understand what it links to.

With a chi-square test a significant difference is found between the problem detection in both methods $X^2(1, N=48)=4.269$, p=.039. The results show that in the CTA method this problem is uncovered more often then in the RTE method.

Table 10, Crosstabulation of method and Usability problem 5

		Usability problem 5		
		No	Yes	Total
Method	CTA	6	18	24
	RTE	13	11	24
	Total	19	29	48

5.1.1.3. Usability problem 8: not finding the comparison tool.

Problem description: When a user wants to compare two travel insurances they can use the travel insurance comparison tool. Most users do not know that such a tool exist. Others search for it, but are unable to find its location. In one of the assignments the user has to compare two travel insurances. In a previous assignment they already visited the travel insurances separate pages. Especially for this task they ideally should look for a comparison tool. Only about 10 percent found the comparison page.

The pearson chi square data ($\chi^2(1, N=48)=5.581$, p=.018) is not representative because there is an expected value in the chi square test is below 5. In this case

Fisher's exact test becomes important. With an exact alpha of p=.050 we conclude the CTA method uncovers UP8 more often then the RTE method.

Table 11, Crosstabulation of method and Usability problem 8

		Usability problem 8		
		No	Yes	Total
Method	CTA	0	24	24
	RTE	5	19	24
	Total	5	43	48

5.1.1.4. Usability problem 15: missing the assistance to calculate the cubic meters of a house.

Problem description: When a user wants to find out the estimated cost of a "House Contents Insurance", they have to fill in a form. One of the questions in this form is: how many cubic meters is your home? Users find it hard to estimate the cubic meters of their home and miss help/guidance in filling this question.

The data suggest that this problem is more often uncovered in the CTA method compared to the RTE method. The statistics however show there is no significant difference between Up15 and the two research methods. Although the alpha comes close to significance ($\chi^2(1, N=48)=2.948$, p=.086) we cannot conclude an effect.

Table 12, Crosstabulation of method and Usability problem 15

		Usability problem 15		
		No	Yes	Total
Method	CTA	16	8	24
	RTE	21	3	24
	Total	37	11	48

5.2. Influence of usability method & location on WEQ ratings

User ratings are obtained with the WEQ questionnaire. In this questionnaire users were asked 28 questions, specifically designed to measure the perceived quality of each usability problem category and sub-category. Data were prepared as discussed in the measurement section. In order to measure whether the usability tests have an influence on the WEQ results a two-way ANOVA on each WEQ category mean and sub category mean, with two between subject factors (method and location), is conducted.

Table 13, The main effect of method on WEQ categories and subcategories

WEQ Usability problem category and subcategory	F	Sig.
1, Content	,000	,989
2, Navigation	,780	,382
3, Layout	,233	,632
1.1, Relevance	,016	,898
1.2, Comprehensibility	,065	,800
1.3, Comprehensiveness	,075	,785
2.1, User friendliness	2,591	,115
2.2, Structure	,316	,577
2.3, Hyperlinks	,039	,843
2.4, Speed	,078	,781

Table 14, The main effect of location on WEQ categories and subcategories

WEQ Usability problem category and subcategory	F	Sig.
1, Content	,388	,537
2, Navigation	2,573	,116
3, Layout	,001	,980
1.1, Relevance	,530	,471
1.2, Comprehensibility	,016	,899
1.3, Comprehensiveness	,365	,549
2.1, User friendliness	1,288	,263
2.2, Structure	2,703	,107
2.3, Hyperlinks	3,043	,088
2.4, Speed	,240	,626

Table 15, The interaction effect of location and method on WEQ categories and subcategories

WEQ Usability problem category and subcategory	F	Sig.
1, Content	,025	,874
2, Navigation	2,309	,136
3, Layout	3,255	,078
1.1, Relevance	,090	,766
1.2, Comprehensibility	,794	,378
1.3, Comprehensiveness	,075	,785
2.1, User friendliness	5,516	,023
2.2, Structure	,906	,346
2.3, Hyperlinks	,215	,645
2.4, Speed	,594	,445

No main effects are found (table 13,14). In subcategory 2.1 and only in subcategory 2.1 a significant interaction effect (F(1,44)=5,516; P=,023) between the method and location is found.

In the two-way ANOVA on the WEQ subcategory 2.1 mean, with two between subject factors (method and location), the two main effects are not significant (F(1,44)=2.59, p=.115) (F(1,44)=1.28, p=.263). The interaction effect of method and location is significant (F(1,44)=5,52, p=.023). In figure 5, the effect is shown. Users rated the user friendliness significantly higher participating in the ABN AMRO location with the CTA method, then in the ABN AMRO location with the RTE Method. A simple effect analysis shows that the individual difference of CTA and RTE method in the ABN AMRO location is significant with an alpha of p=.036. In the university environment method does not have a significant influence on the user friendliness rating. The individual difference of CTA and RTE method in the university location is not nearly significant with an alpha of p=.370.

Table 16, The interaction effect between the method and location in WEQ subcategory 2.1

Method	Location	Mean	Std. Error
CTA	ABN AMRO	3,833	,408
	University Eindhoven	3,450	,182
RTE	ABN AMRO	2,583	,408
	University Eindhoven	3,683	,182

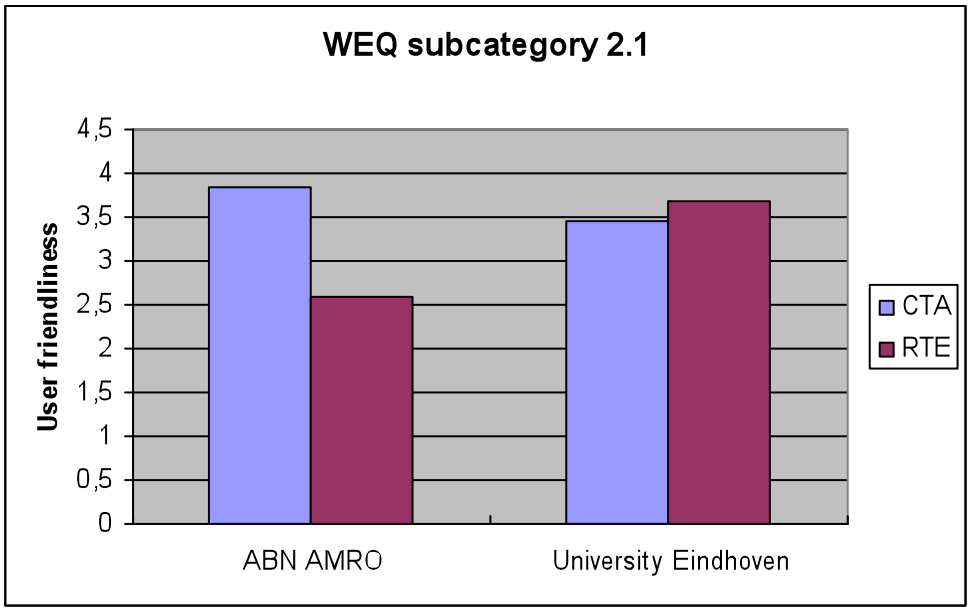


Figure 5, The interaction effect between the method and location in WEQ subcategory 2.1

5.3. Influence of usability method & location on qualities, beauty and goodness user ratings

In this research a questionnaire is conducted concerning: Pragmatic quality(PQ), Hedonic qualities; stimulation(HQS) and identification(HQI), beauty and goodness. The questionnaire used was originally designed and used by Hassenzahl (2004). The questionnaire is answered before and after the user did the assignments. By using the data from these questionnaires we can compare the results, and look at the influence of the location and method on the perceived goodness of the tested website.

The scores from the first questionnaire (Pre) are given after the user looked at the website (homepage of www.abnamro.nl , July 2011) for 40 seconds without interacting with it. The scores from the second questionnaire (Post) are given after users completed six assignments, in a total of 30 to 60 minutes. Questions are rated on a 7-point scale. Values vary from minus three to plus three.

When measuring PQ, HQS, HQI, goodness and beauty, the method and location where the questionnaire is conducted are very important. Crowne & Marlowe (1960) and Ross & Mirowsky (1984) suggest that people tended to give socially desirable responses. Giving a good rating for the ABN AMRO website can be considered to be a socially desirable responses. This would lead to the assumption that Pre test scores at the ABN AMRO location generally would be more positive then the Pre test scores at the university location. Looking at the data this trend is clearly visible. In the second test, the Post test, this effect seems to have been disappeared. This could mean that the effect of answering the questions in a socially desirable way is lost after conducting several tasks in an usability test.

A repeated measurements analysis of variance with the within-subject factor ”time of measurement” (pre usage, post usage) and between-subject factors “Method and location”, is conducted for each type of measurements)(PQ, HQS, HQI, Beauty and Goodness).

The result section below is divided into five parts. In every part one of the measured items by Hassenzahl’s questionnaire is analyzed. The data are divided into two sections. The within subject effect section discusses the within subject effect and the between subject effects between the dependent variables (method and Location) and the within subject effect of the concerning measurement. In the between subject effects section the between subject effects of method and location on of the average Pre/Post test values if the concerning measurement are discussed.

5.3.1. Pragmatic qualities

A repeated measurement ANOVA was conducted to test the influence of method and location on PQ rating differences in Pre and Post test scores. Pre and Post test scores of PQ are used as within subject variables. Method and location are added as between subject factors. In this paragraph first the within subject results are discussed and then the between subject factors. When a relation is significant or near significance it is discussed in the text below. At the beginning of every part a table is shown with F values and alphas of all relevant relations.

Within subject effects:

One within subject effect is discussed. As can be seen in table 17, none of the effects is significant and one is near significant. The interaction effect between method and (Pre/Post)test PQ is discussed for it is near significant.

Table 17, Tests of Within-Subject effect of PQ

Effect	F	Sig.
(Pre/Post)PQ	2,407	,128
(Pre/Post)PQ * Location	,000	,996
(Pre/Post)PQ * Method	3,056	,088
(Pre/Post)PQ * Location * Method	1,110	,298

Although the effect is not significant, the data indicate a possible interaction effect between the Pre/Post test ratings of PQ and the method used ($F(1,43)=3.06$, $P=.088$). In the CTA method the PQ rating decreased, where in the RTE Method the PQ rating stays the same. A simple effect analysis shows no significant difference within the Pre test ratings for PQ values ($p=.135$).

This result is in line with the hypothesis that suggests that when participating in the CTA method, users reflect on their own progression and they will experience the usability problems more conscious. Therefore CTA method will have a negative effect on the PQ rating.

Table 18, The interaction effect of Location and Method on Pragmatic qualities

Method	PQ	Mean	Std. Error
CTA	Pre test	4,292	,152
	Post test	3,850	,174
RTE	Pre test	3,963	,153
	Post test	3,989	,175

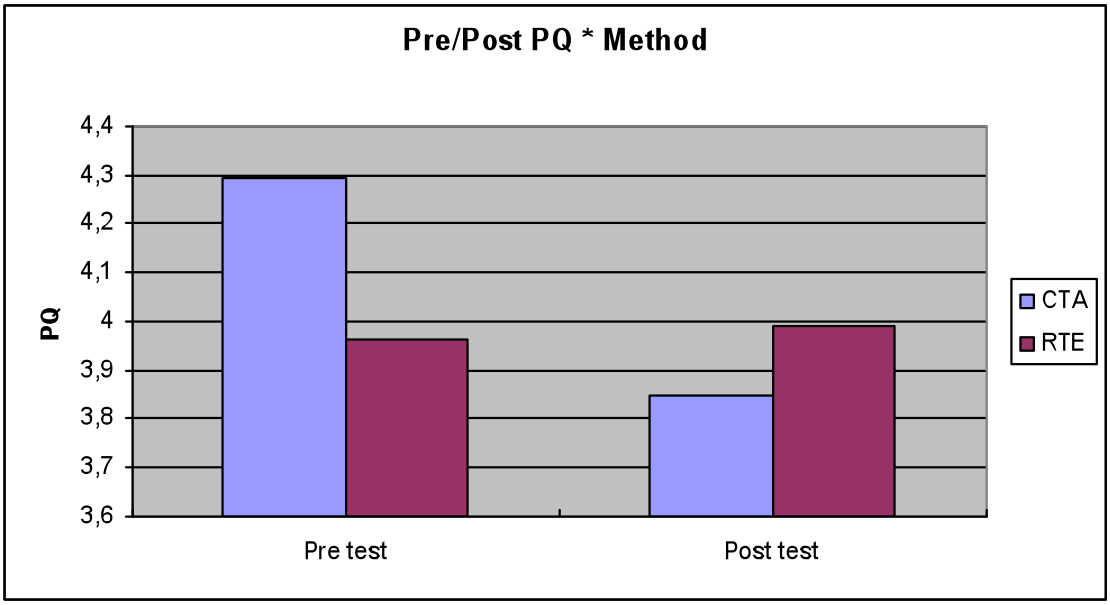


Figure 6, The interaction effect of method on Pragmatic qualities

Between subject effects:

No between subject effects are discussed. As can be seen in table 19, no effect is significant or near significant.

Table 19, Tests of Between-Subjects effects of PQ

Effect	F	Sig.
Location	1,158	,288
Method	,253	,618
Location * Method	,674	,416

5.3.2. Hedonic qualities – stimulation

A repeated measurement ANOVA was conducted to test the influence of method and location on HQS rating differences in Pre and Post test scores. Pre and Post test scores of HQS are used as within subject variables. Method and location are added as between subject factors. In this paragraph first the within subject results are discussed and then the between subject factors. If a relation is not significant or near significance it is not discussed in the text. At the Beginning of every part a table is shown with F values and alphas of all relevant relations.

Within subject effects:

One within subject effect is discussed. As can be seen in table 20, one of the interaction effects is significant. The effect of location on HQS is discussed for it is significant.

Table 20, Tests of Within-Subject effect of HQS

Effect	F	Sig.
(Pre/Post)HQS	,694	,409
(Pre/Post)HQS * Location	4,296	,044
(Pre/Post)HQS * Method	,108	,743
(Pre/Post)HQS * Location * Method	,409	,526

The hedonic quality stimulation, measures how the design stimulates the user to keep using the website. In the research a significant interaction between the Pre test, Post test scores of HQS and the location was found ($F(1, 44)=4.30$, $p=.044$). In order to measure the effect of the individual locations on Post test scores of HQS, a simple effect analysis is conducted. The Pre and Post test scores of HQS are compared per location. As can be seen in Table 21, in the ABN AMRO location the post test scores decreased, not significant ($p=.119$), compared to the Pre test scores. Where in the university location the Post test score increased, also not significant ($p=.136$), compared to the Pre test score.

Table 21, Simple effect analysis of Pre/Post HQS and Location

Location	(Post-Pre)Mean	Std.Error	Significance
ABN AMRO	-,375	,236	,119
University Eindhoven	,160	,105	,136

Table 22, effect of Location and HQS

Location	Test moment	Mean	Std. Error
ABN AMRO	Pre test	-,425	,363
	Post test	-,800	,325
University Eindhoven	Pre test	-,790	,162
	Post test	-,630	,145

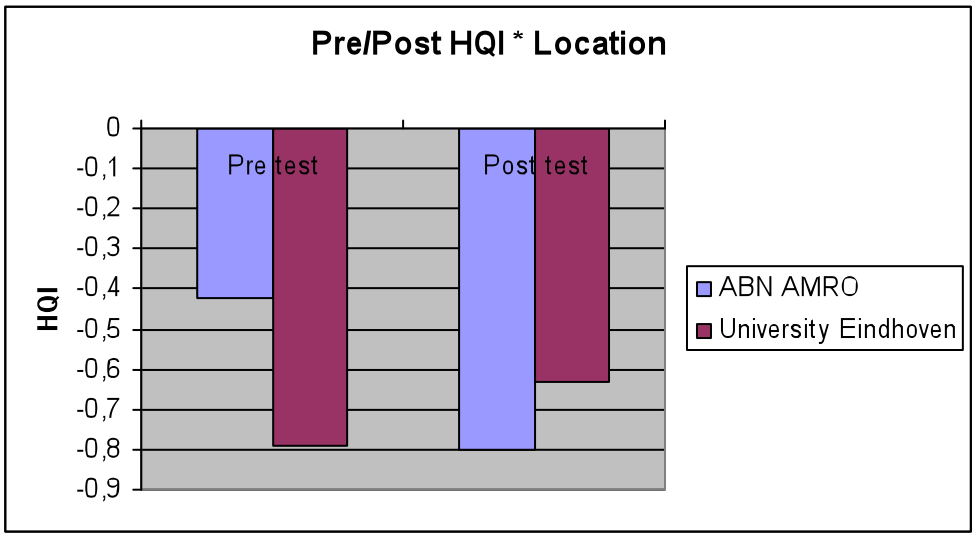


Figure 7, effect of location on HQS

Between subject effects:

No between subject effects are discussed. As can be seen in table 23, no effect is significant or near significant.

Table 23, Tests of Between-Subjects effects of HQS

Effect	F	Sig.
Location	,076	,785
Method	,393	,534
Location * Method	,076	,785

5.3.3. Hedonic qualities – Identification

A repeated measurement ANOVA was conducted to test the influence of method and location on HQI rating differences in Pre and Post test scores. Pre and Post test scores of HQI are used as within subject variables. Method and location are added as between subject factors. In this paragraph first the within subject results are discussed and then the between subject factors. If a relation is not significant or near significance it is not discussed in the text. At the beginning of every parts a table is shown with F values and alphas of all relevant relations.

Within subject effects:

One within subject effects is discussed. As can be seen in table 24, one of the effects is significant. The within subject effects HQI is discussed for it is significant.

Table 24, Tests of Within-Subject effect of HQI

Effect	F	Sig.
(Pre/Post)HQI	8,115	,007
(Pre/Post)HQI * Location	,001	,975
(Pre/Post)HQI * Method	,370	,546
(Pre/Post)HQI * Location * Method	,124	,726

Despite the fact that the general appreciation decreases for almost every aspect, when people perform the usability test, the HQI rating is increasing after usage. There is a significant effect between the Pre and Post test variable ($F(1,44)=8.12$, $p=.007$). Users tent to increasingly identify themselves with the ABN AMRO website after using it for a short period of time.

Table 25, the influence of test moment on HQI

Test moment	Mean	Std. Error
Pre test	-1,150	,139
Post test	-,872	,125

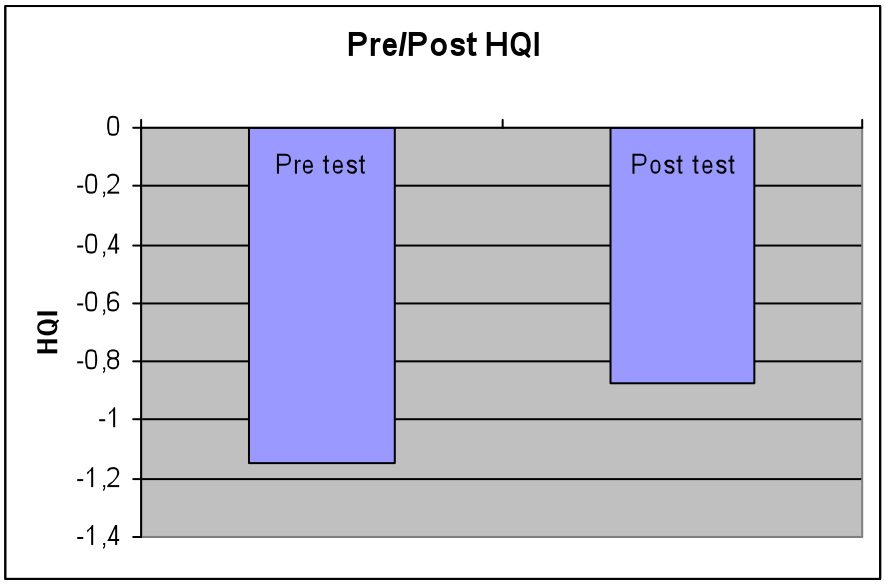


Figure 10, difference in HQI.

There is no indication that either location ($F(1, 44)<0.01$, $p=.975$) or method ($F(1, 44)=3.70$, $p=.546$) have any influence on this matter. Where the general HQI rating was very negative in the Pre test, it is still negative but more positive in the Post test.

Between subject effects:

No between subject effects are discussed. As can be seen in table 26, no effect is significant or near significant.

Table 26, Tests of Between-Subjects effects of HQI

Effect	F	Sig.
Location	,959	,333
Method	1,169	,286
Location * Method	1,784	,189

5.3.4. Beauty

A repeated measurement ANOVA was conducted to test the influence of method and location on beauty rating differences in Pre and Post test scores. Pre and Post test scores of beauty are used as within subject variables. Method and location are added as between subject factors. In this paragraph first the within subject results are discussed and then the between subject factors. If a relation is not significant or near significance it is not discussed in the text. At the beginning of every parts a table is shown with F values and alphas of all relevant relations.

Within subject effects:

Three within subject effects are discussed. As can be seen in table 27, two of the effects are significant and one effect is near significant. The within subject effects beauty and interaction effects of location on beauty and method on beauty are discussed.

Table 27, Tests of Within-Subject effect of Beauty

Effect	F	Sig.
(Pre/Post)Beauty	14,872	,000
(Pre/Post)Beauty * Location	13,984	,001
(Pre/Post)Beauty * Method	3,947	,053
(Pre/Post)Beauty * Location * Method	,874	,355

By participating in the usability test, user ratings of beauty of the ABN AMRO website decreases significantly ($F(44,1)=14.872$, $p<.001$). In the Post test questionnaire user rating were far below the Pre test scores. On a scale from -3 to 3, the average rating decreased by 0.825. The within subject effect of Beauty alone, accounted for 25,3% of all the variance. The effect is shown in Figure 28.

Table 28, Within subject effect of Beauty

Beauty	Mean	Std. Error
Pre test	-,237	,230
Post test	-1,062	,206

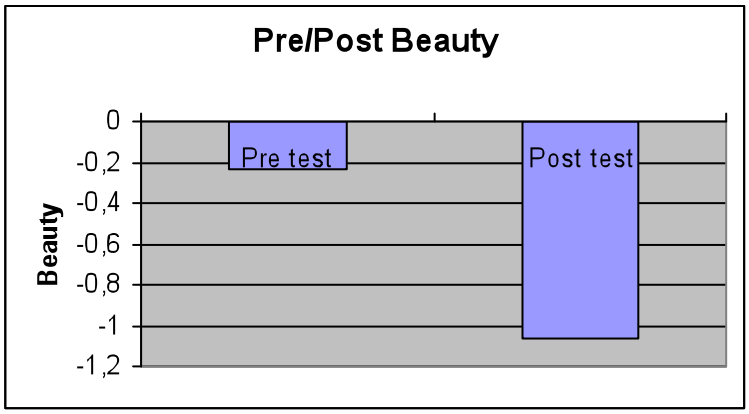


Figure 11, Within subject effect of Beauty

The data shows there is an interaction effect (figure 12) between location and beauty (F44,1)=13.984, p=.001). In order to measure the effect of the individual locations on post test scores of beauty, a simple effect analysis is conducted. The Pre and Post test scores of beauty are compared per location. As can be seen in Table 29, in the ABN AMRO location the post test scores decreased significantly (p<.001), compared to the Pre test scores. Where in the university location the Post test score hardly differs from the Pre test score.

Table 29, Simple effect analysis of Pre/Post Beauty and Location

Location	(Post-Pre) Mean	Std.Error	Significance
ABN AMRO	-1,625	,391	,000
University Eindhoven	-,025	,175	,887

Table 30, effect of Location on Beauty

Location	Beauty	Mean	Std. Error
ABN AMRO	Pre test	,375	,420
	Post test	-1,250	,376
University Eindhoven	Pre test	-,850	,188
	Post test	-,875	,168

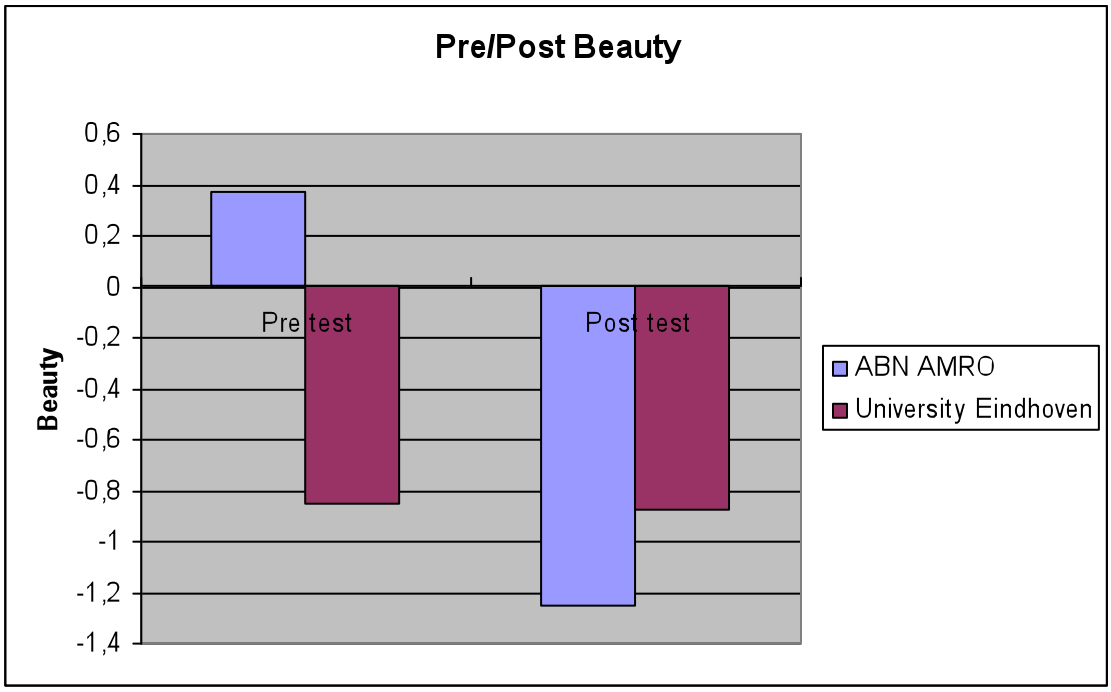


Figure 12, Effect of Location and Method on mean Beauty

The data indicate there is an interaction effect (figure 13) of method and Pre/Post test scores of beauty ($F(44,1)=3.947$, $p=.058$). In order to measure the effect of the individual method on post test scores of beauty, a simple effect analysis is conducted. The Pre and Post test scores of beauty are compared per method. As can be seen in Table 31, in the RTE Method the Post test scores decreased significantly ($p<.001$), compared to the Pre test scores. Where in the CTA Method the Post test score hardly differ from the Pre test score.

Table 31, Simple effect analysis of Pre/Post Beauty and Method

Method	Mean difference (Post-Pre)	Std.Error	Significance
CTA	-,400	,391	,193
RTE	-1,250	,175	,000

Table 32, Effect of Method on Pre/Post test Beauty

Method	Beauty	Mean	Std. Error
CTA	Pre test	-,600	,325
	Post test	-1,000	,291
RTE	Pre test	,125	,325
	Post test	-1,125	,291

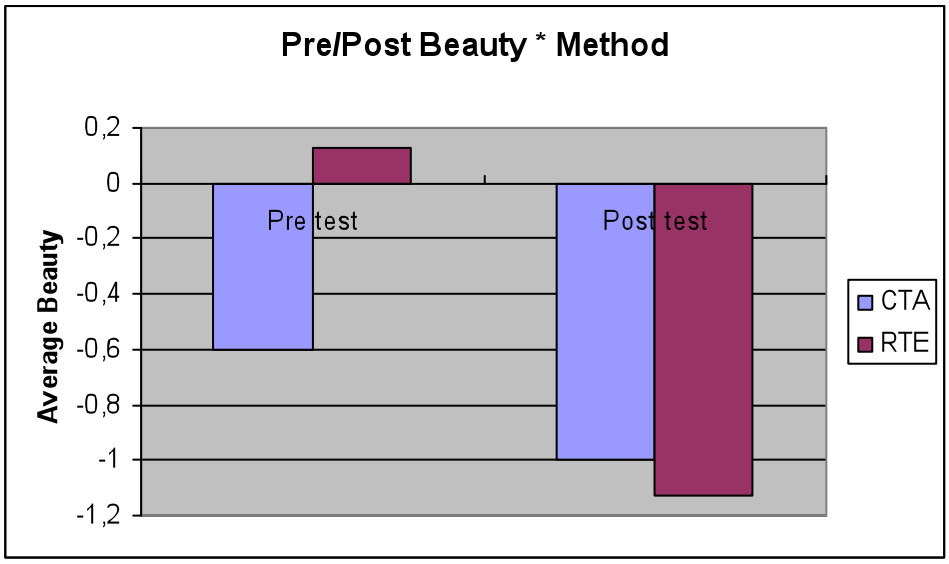


Figure 13, Interaction effect between Beauty and Location.

Between subject effects

One between subject effects is discussed. As can be seen in table 33, one effect is significant. The interaction effect of method and location on the average beauty ratings is discussed.

Table 33, Tests of Between-Subjects effects of Beauty

Effect	F	Sig.
Location	1,249	,270
Method	,622	,434
Location * Method	4,706	,036

The combined factors method and location have an significant interaction effect on the beauty ratings ($F(1,44)=4.71$, $p=.036$). In figure 14, the interaction effect is shown. The data indicate that within the ABN AMRO Location the RTE method users rate the website to be more beautiful then users in the CTA condition. A simple effect analysis shows the data is not significantly different within the ABN AMRO location ($p=0.112$). The data also indicate that within the university location the users in the CTA method rate the website to be more beautiful then users in the RTE condition. A simple effect analysis shows the data is not significantly different within the university location ($p=0.098$).

Table 34, Interaction effect of Method and Location on the Beauty mean

Location	Method	Mean	Std. Error
ABN AMRO	CTA	-1,000	,491
	RTE	,125	,491
University Eindhoven	CTA	-,600	,220
	RTE	-1,125	,220

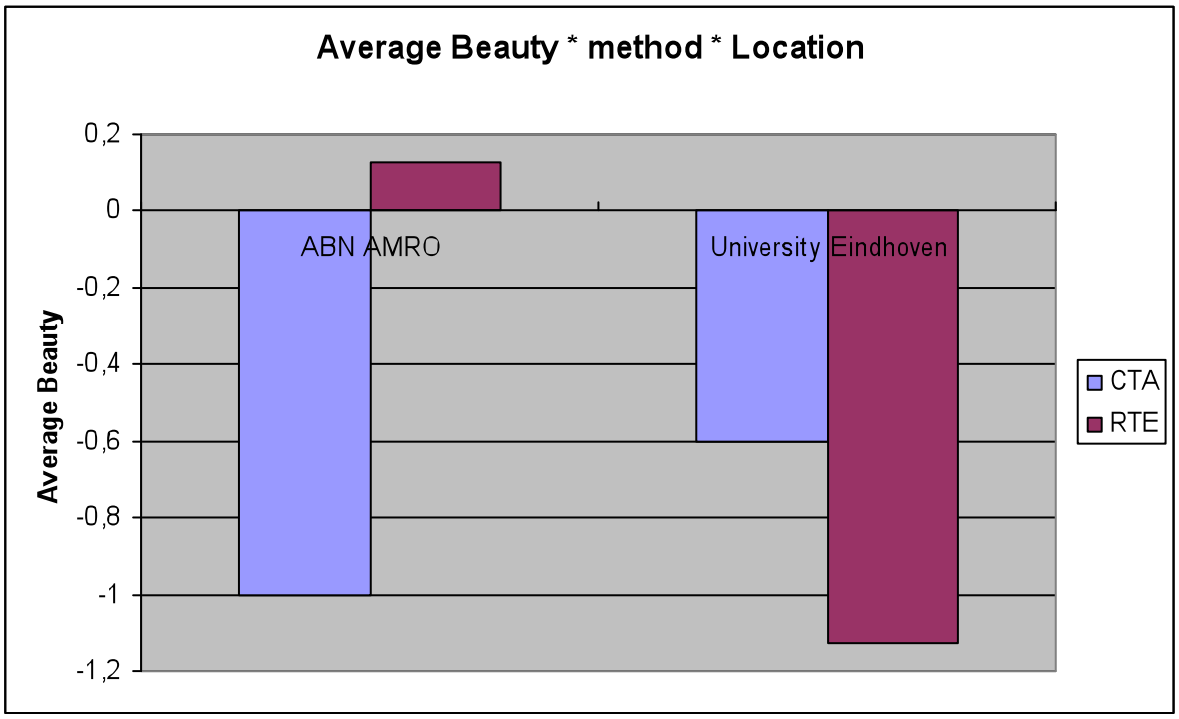


Figure 14, Interaction effect between average Beauty, Method and Location.

5.3.5. Goodness

A repeated measurement ANOVA was conducted to test the influence of method and location on the goodness rating differences in Pre and Post test scores. Pre and Post test scores of goodness are used as within subject variables. Method and location are added as between subject factors. In this paragraph first the within subject results are discussed and then the between subject factors. If a relation is not significant or near significance it is not discussed in the text. At the beginning of every part a table is shown with F values and alphas of all relevant relations.

Within subject effects:

Two within subject effects are discussed. As can be seen in table 35, one effect is significant, and one effect is near significant. The within subject effect is discussed for the effect is nearly significant. The interaction effect between location on goodness is discussed for it is significant.

Table 35, Tests of Within-Subject effect of Goodness

Effect	F	Sig.
(Pre/Post)goodness	3,218	,080
(Pre/Post)goodness * Location	5,093	,029
(Pre/Post)goodness * Method	,968	,331
(Pre/Post)goodness * Location * Method	,164	,687

The data indicates that Pre test goodness is rated higher then Post test goodness. The Pre/Post test scores of goodness shown in table 36, do not differ significantly. The data shows no significant difference between the Pre test and Post test scores ($F(1,44)=3.22$, $p=.080$).

Table 36, Within subject effect of Goodness

goodness	Mean	Std. Error
Pre test	1,225	,163
Post test	,838	,200

The variable location has a significant effect on the Pre/Post test scores of goodness (F(1,44)=5,093, p=.029). This means that Pre/Post test scores of goodness are interacting with location. In order to measure the effect of the individual locations on post test scores of goodness, a simple effect analysis is conducted. The Pre and Post test scores of goodness are compared per location. As can be seen in Table 37, in the ABN AMRO location the post test scores decrease significantly (p=.032), compared to the Pre test scores. Where in the university location the Post test score hardly differs from the Pre test score.

Table 37, Simple effect analysis of Pre/Post goodness and Location

Location	Mean difference (Post-Pre)	Std.Error	Significance
ABN AMRO	-.875	.394	.032
University Eindhoven	.100	.176	.574

Table 38, Effect of location on the pre/post test scores of goodness

Location	goodness	Mean	Std. Error
ABN AMRO	Pre test	1,500	,297
	Post test	,625	,366
University Eindhoven	Pre test	,950	,133
	Post test	1,050	,164

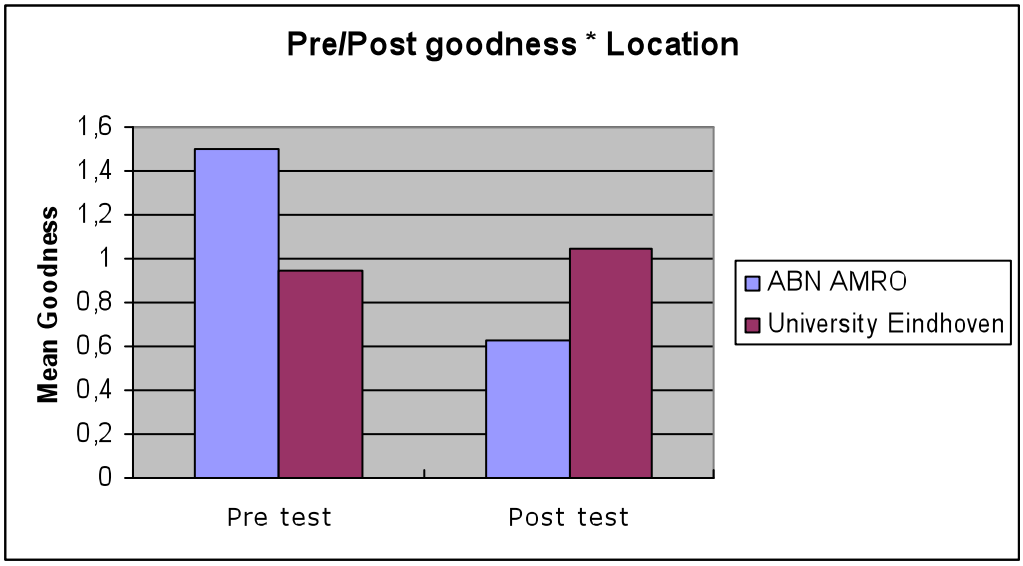


Figure 15, Interaction affect between Goodness and Location.

Between subject effects:

Two between subject effects are discussed. As can be seen in table 39, two effects are significant. The effect of method on the average goodness and the interaction effect of method and location on the average goodness is discussed.

Table 39, Tests of Between-Subjects effects of Goodness

Effect	F	Sig.
Location	,045	,833
Method	4,705	,036
Location * Method	6,297	,016

The variable method has a significant effect on the mean goodness ($F(1,44)=4.70$, $p=.036$). We will not go into this effect further, since also a significant interaction of location and method on the mean goodness is found ($F(1,44)=6.30$, $p=.016$). In figure I, the interaction affect is shown. A simple effect analysis shows that the individual ABN AMRO location scores differ significantly ($p=.014$). Within the ABN AMRO location, the average goodness scores, are significantly higher then the average goodness scores in the RTE method. In the university location, this effect does not hold.

Table 40, Effect of Location and Method on the within subject effect of Goodness

Location	Method	Mean	Std. Error
ABN AMRO	CTA	1,500	,297
	RTE	,375	,379
University Eindhoven	CTA	,950	,170
	RTE	1,050	,170

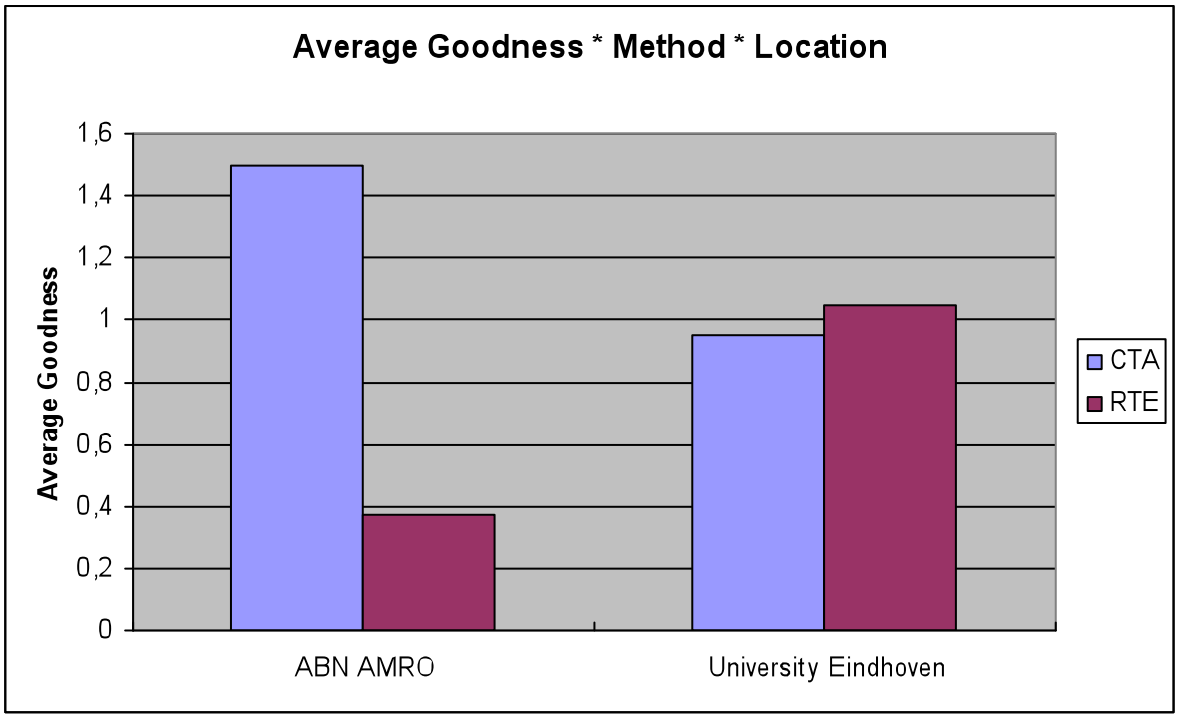


Figure 16, Effect of Location and Method on the average Goodness per method.

6. Discussion

The two methods used in this research, CTA and RTE, do not differ in the detection of usability problems in the main categories as discussed in the WEQ (Elling et al, 2007). The CTA method however was significantly better in detecting problems concerning user friendliness. A deeper differentiation in subcategories could not be made. Three raters separately allocated the detected 19 problems among the main and sub categories. Although all the problems where allocated into the same main category, not all the usability problems where divided into the same sub category. Obviously allocating problems among (sub) categories is a subjective process. In order to make the (sub) category differentiation more reliable a specific set of rules should be created to allocate problems into categories.

The two locations used in this research, ABN AMRO usability lab and University Eindhoven game experience lab, do not differ in the detection of usability problems in the main categories as discussed in the WEQ (Elling et al, 2007). The conclusion is that detected usability problems in an academic environment are equal to the problems detected within a setting in a big company. This is in contrast to the statement of Nielson & Pernice (2010) that academic papers are irrelevant for commercial design.

The CTA test method has a positive effect on user ratings. A divers range of questions was rated higher after users where tested in the CTA method compared to users that where tested with the RTE method. It seems that users feel better about a website if they where able to comment on it. When questionnaires are used after the usability test, experimenters should take into account that listening to a users comment makes them rate the website higher.

The influence of the test location on the results out of questionnaires is big. Especially the effect on Pre/Post test differences was measurable. In general when using Pre/Post questionnaires in a realistic setting as the ABN AMRO usability lab, the pre test scores are biased. A small research on first impression measurements in two different locations, one university environment and at a building of the website companies, could give better insights in the influence of these environments on pre test measurements.

The mood questionnaire (PANAS) was conducted. Due to some operational problems the answers are not used in this research. The purpose of the questionnaire was to measure differences in users attitude. Unfortunately the effect of users mood could not be measured. The possible differences in

questionnaire results depending on location might be better evaluated using the PANAS data. In further research it would be interesting to use these to evaluate the location dependant influences.

The users differed in several aspects per location. In both locations the users where gathered in the regular way for that location. The ABN AMRO location users where gathered by an external bureau that supplies users for marketing research. Each user received 40 euro's compensation. The users in the University environment where gathered via a mail database for students that regularly perform in university research. These users received 15 euro's compensation. These differences where unavoidable but could in some way have influenced the results. The age level and educational level of the ABN AMRO location users was manipulated in order to match the University environment participants. Normally users participating in the ABN AMRO environment in average have a higher age and a lower educational level.

Due to software complications on data migration, no representational heatmaps could be generated. Nielson & Pernice (2010) state that about 30 users with representational eye tracking data give valid data. In this research the heatmaps could maximally be compared with 6 users, making the comparison irrelevant. Therefore the data were left out. An interesting question for further research is: does the usability test method influence eye movements?

The most recent questionnaire of Hassenzahl is used to measure different qualities, beauty and goodness. This questionnaire was translated into Dutch, from English and German versions. When analysing the data, we found three questions not correlating with the rest of the questions. In the translation these three questions lost their descriptive meaning. For this reason those questions where excluded, and not used in the data analysis.

One point of discussion is the amount of users. In this research 48 users were tested; 8 of them in the ABN AMRO location, the other 40 in the University location. More users in the ABN AMRO location would increase the generalisability of the test. When calculating interaction effects the number of users in some groups is reduced to 4 users. This number of users in one group makes the parametric methods, as used in this research to calculate the alpha values, less/not reliable. This causes a serious problem for interpreting the results. Additional research with a higher number of users in each group would lead to more generalisable conclusions.

7. Conclusion

After presenting the results now the research question will be answered.

What is the influence of usability tests method and location in usability testing?

In order to answer this question three sub questions are used.

1. What is the influence of usability test method on the number of usability problems uncovered per category, depending on usability test method and location?
2. What is the influence of usability test method and location on the WEQ results per category?
3. What is the influence of usability test method and location on the Quality questionnaire results?

These three sub questions help detecting influences of method and location.

1a. What is the effect of location on the number of uncovered usability problems per categories?

There where no differences found in the usability problems categories per location. This result implies that a usability test conducted in a laboratory setting within a big company as ABN AMRO will uncover the same number of problems per category as in a university environment. This is in contradiction with the statement of Nielsen & Pernice (2010), that academic papers are irrelevant for commercial design. The amount of problems detected per category in academic papers is in this research setup comparable to the problems found in a setting within a big company.

1b. What is the effect of the method on the number of uncovered usability problems per categories?

The CTA method is better in detecting usability problems in sub category 2.1 "User friendliness". Main category 2 "Navigation" does not show any difference in the number of usability problems uncovered. In the hypothesis a difference in main category 2 was presumed. The RTE method would uncover more problems because we expected that in the CTA method users were expected to be pushed to work in a more structured way. Contrary to the hypothesis, the CTA method uncovers significantly more problems. Looking at the individual problems within subcategory 2.1 no common feature can be detected. Therefore a reason for this difference is not found. The experimenter expects that the constant questions on self report and thinking aloud makes the user more aware of the "test" situation. This could affect the users and thereby stimulating them to consciously look for

more issues to comment on. The issues they comment on in these situations belong to the category of user friendliness.

2. What is the effect of the method and location on categorical user ratings?

The research shows no main effect. This means that nor method, nor location alone have a significant effect on the WEQ ratings. However, the combination of method and location does have a significant interaction affect on the WEQ ratings of subcategory 2.1. Results show that in the ABN AMRO location the CTA method has a significant more positive influence on user ratings then the RTE method.
This result is in contrast with the outcome of sub question 1, where the CTA problem found more usability problems in this same subcategory. The reason for this difference is not clear. A suggestion is that talking more about user friendliness, decreases user frustration and thereby leading to a higher user friendliness rating. Another reason could be that in the CTA method the user creates a personal affiliation with the experimenter. The users could hereby unconsciously be influenced to rate the product higher when they feel this is socially desirable.

3. What is the effect of usability testing on hedonic and pragmatic quality ratings?

Despite the fact that the general opinion decreases for almost every aspect, the HQI rating is increasing after usage. There is a significant effect between the pre and post test variable. This means that the rated HQI increases when users work with the insurance part of the ABN AMRO website.

3a. What is the effect of the usability test method on hedonic and pragmatic quality ratings?

Within the ABN AMRO Location, the CTA Method results in higher average Goodness scores, compared to the RTE Method. Goodness is measured as a sort of overall rating of the website. Goodness ratings show the same effect as described in sub question 2. It seems that talking about problems on the website results in a higher rating of the website.

3b. What is the effect of the location on hedonic and pragmatic qualities?

The location seems to have an influence on the Pre test scores of the measured qualities. Possible explanation for the differences of the location on the answers of the Pre test is that the users enter the test location with a different set of expectations. The users that Participate in the ABN AMRO location generate a higher level of expectation of the hedonic and pragmatic qualities of the website.

Hedonic qualities – stimulation (HQS) measures how the design stimulates the user to keep using the website. In the research a significant interaction between

the pre test, post test scores of HQS and the location was found. This means that depending on the location a usability test is conducted, the usability test either increases or decreases the stimulation. When measuring HQS one should take into account the influence of the research location. In this research the big company environment has a more positive effect on the Pre test scores, and a more negative effect on the Post test scores, compared to the university environment.

The ABN AMRO location has a positive influence on the Pre test Beauty rating. This is exactly what was hypothesised. We expect that this effect is caused by the fact that, users are impressed by the ABN AMRO dialogue house environment. Besides that, users feel more social pressure to give socially answers in the ABN AMRO location. After usage of the website the ABN AMRO ratings on Beauty dropped to the level of the University ratings. This means that the initial effect of location on Beauty ratings only holds for the Pre test, it does not influence user ratings on Beauty in the Post test.

8. References

- Burg, J.V.D. (2008). Zien zonder te luisteren. *Masterscriptie Communicatiestudies*, Universiteit van Utrecht.

- Cooke, L. and Cuddihy, E. (2005). Using Eye Tracking to Test the Validity and Accuracy of Think-aloud Protocol, *Proceedings of the IEEE International Professional Communication Conference*, July, Limerick, Ireland.

- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349-354

- Eger, N., Ball, L.J., Stevens, R. & Dodd, J. (2005). Validating the use of eye-movement replay to cue retrospective verbal protocols in online usability testing. *ACM Press*. Lancaster University, Lancaster.

- Eger, N., L. Ball, L.J., Stevens, R. & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eyemovement replay*. People and Computers XXI – HCI … but not as we know it: Proceedings of HCI* 2007.

- Elling, S., Lentz, L. & De Jong, M. (2007). Website Evaluation Questionnaire: *Development of a research-based tool for evaluating informational websites.* In: Wimmer, M.A., H.J. Scholl & A. Grönlund (Eds.): EGOV 2007, LNCS 4656 (293-304). Berlin Heidelberg: Springer-Verlag.

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

- Guan, Z., Lee, S., Cuddihy, E., Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. *CHI 2006 Proceedings.* Usability Methods, Montréal, Québec, Canada.

- Gombert, J. (2009). Je eigen ogen als afleidingsmanoeuvre??!! Een kwalitatief onderzoek naar het effect van het terugzien van de eigen oogbewegingen bij participanten tijden Retrospective Thinking-Aloud. *Masterscriptie Communicatiestudies,* Universiteit Utrecht.

- Haak, M., Jong, M. & Schellens, P.J. (2006). Hardopdenkprotocollen en gebruikersonderzoek: Volledigheid en reactiviteit van de synchrone hardopdenkmethode. *Tijdschrift voor Taalbeheersing*, 28 (3), 185-197.

- Haak, M., Jong, M. & Schellens, P.J. (2007). Evaluation of an informational Web site: Three variants of the think-aloud method compared. *Technical Communication*, 54(1), 58 – 71.

• Haak, M., De Jong, M. & Schellens, P.J. (2009). Evaluating municipal websites: A methodological comparison of three think-aloud variants. In: *Government Information Quarterly*, 26, 193 – 202.

• Haas, E. (2009). 'Ik moet even wennen aan die stippeltjes'. Een onderzoek naar de toegevoegde waarde van het terugzien van de oogbewegingen op verbalisaties bij retrospectief hardopdenken, uitgevoerd op een gemeentelijke website. *Masterscriptie Communicatiestudies*, Universiteit Utrecht.

• Hassenzahl, M. (2004). The Interplay of Beauty, Goodness, and Usability in Interactive products. HUMAN-COMPUTER INTERACTION. Volume 19, pp. 319–349

• Kaikkonen, A., Kallio, T., Aki Kekäläinen, A., Anu Kankainen, A., Cankar, M., (2005) Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing*. Journal of usability studies*. Issue 1, Vol. 1, November 2005, pp. 4-16

• Mark, R. (1994). Prototyping for tiny fingers. *Communications of the ACM*. Volume 37, Issue 4, 22-27

• Mueller, F., & Lockerd, a. (2001). Cheese: Tracking Mouse Movements on Websites, A Tool for User Modeling, *proceedings of 2001 conference on Human Factors ans Computing Systems (CHI 2001),* ACM Press.

• Nell, L. (2009). Ik verbaliseer dus ik evalueer. Een studie naar de invloed van de gebruikersgerichte webevaluatiemethoden synchroon hardopdenken en retrospectief hardopdenken en demografische kenmerken op de taakuitvoering en verbalisaties van proefpersonen. *Masterscripte Communicatiestudies*, Universiteit Utrecht.

• Nielsen, J. (1994). *Usability Engineering*, Academic Press Inc, p 165

• Nielsen, J., & Pernice, k. (2010). Eyetracking web usability. Berkeley, CA: Nielsen Norman Group.

• Potappel, E. (2007). Kun jij gedachten lezen? Een onderzoek naar de toegevoegde waarde van oogbewegingen bij retrospectieve hardopdenkprotocollen in probleemopsporend usabilityonderzoek. *Masterscriptie Communicatiestudies*, Universiteit van Utrecht.

• Just, M. A. & Carpenter, P. A. (1976). Eye fixations and cognitive processes, Cognitive Psychology 8, 441-480.

• Rayner, k. (1998) Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, Vol. 124, No. 3, 372-422

• Riel, E. V. (2010). Weet u nog wat u hier dacht? Een onderzoek naar de bijdrage van observaties en verbalisaties aan het detecteren van

problemen en de mate waarin persoonskenmerken en de conditie RTE en RTA daar invloed op hebben. *Masterscriptie Communicatiestudies*, Universiteit van Utrecht

- Ross, c., Mirowsky, J. (1984) Socially-desirable response and acquiescence in a cross-cultural survey of mental health*. Journal of health and social behavior,* vol. 25 (juni):189-197

- Watson, D., Clark , L. A. (1988). Development and validation of brief measurer of positive and negative affect: The PANAS scales. *Journal of personality and social psychology*, vol. 54, no 6 1063 – 1070.

- Zhu, W., Vu, K.-P. L., & Proctor, R. W. (2005). Evaluating Web Usability. In R. W. Proctor & K.-P. L. Vu (Eds.), *Handbook of Human Factors in Web Design* (pp. 321-337). Mahwah, NJ: Lawrence Erlbaum Associates, Publisher

Appendix

Appendix 1

Questionnaire 1:

	Neutraal							
Technisch	0	0	0	0	0	0	0	Menselijk
Typisch	0	0	0	0	0	0	0	Origineel
Isolerend	0	0	0	0	0	0	0	Integrerend
Standaard	0	0	0	0	0	0	0	Creatief
Moedig	0	0	0	0	0	0	0	Voorzichtig
Moeilijk	0	0	0	0	0	0	0	Gemakkelijk
Behoudend	0	0	0	0	0	0	0	Vernieuwend
Saai	0	0	0	0	0	0	0	Aangrijpend
Praktisch	0	0	0	0	0	0	0	Onpraktisch
Uitdagend	0	0	0	0	0	0	0	Simpel
Professioneel	0	0	0	0	0	0	0	Amateuristisch
Oud	0	0	0	0	0	0	0	Nieuw
Verwarrend	0	0	0	0	0	0	0	Helder
Elegant	0	0	0	0	0	0	0	Opzichtig
Voorspelbaar	0	0	0	0	0	0	0	Onvoorspelbaar
Duur	0	0	0	0	0	0	0	Goedkoop
Onhandelbaar	0	0	0	0	0	0	0	Handelbaar
Presenteerbaar	0	0	0	0	0	0	0	Niet presenteerbaar
Mooi	0	0	0	0	0	0	0	Lelijk
Slecht	0	0	0	0	0	0	0	Goed

Questionnaire is translated from Hasenzahl (2007)

Questions marked with yellow are removed for the analysis.

Factor analysis with 7 Hedonic Quality – Stimulation questions. This table indicates that question 3 and 6 do not measure the same effect, all questions are rated as 1 being the worse score and 7 being the best score.

Question	Factor	
	1	2
1	,671	,441
2	,804	-,011
3	-,413	,712
4	,588	-,391
5	,874	,079
6	-,512	-,610
7	,786	-,184
Extraction Method: Principal Component Analysis.		

Factor analysis with 5 Hedonic Quality – Identification questions. This table indicates that question 1 does not measure the same effect, all questions are rated as 1 being the worse score and 7 being the best score.

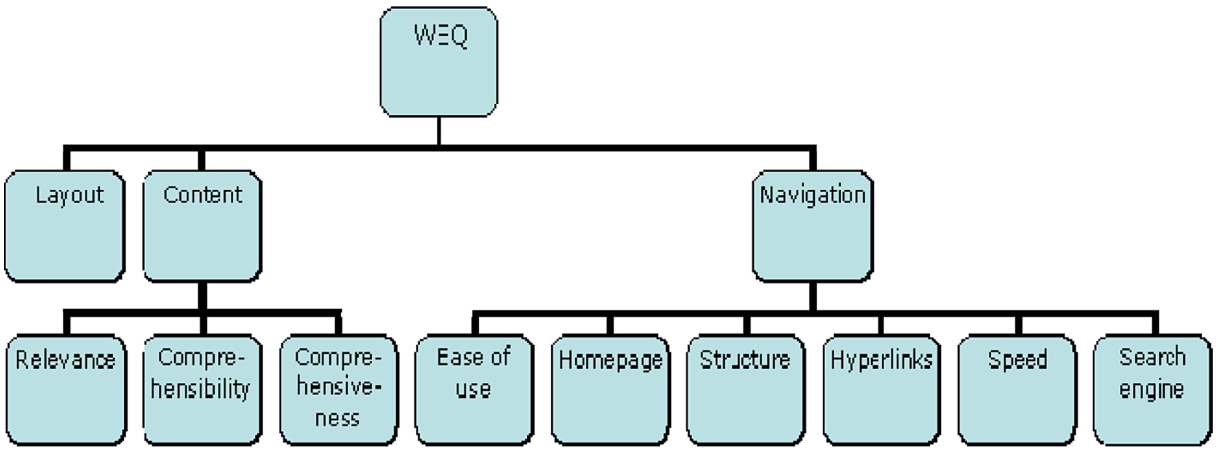
Question	Factor	
	1	2
1	-,069	,877
2	,879	-,013
3	,709	,409
4	,589	-,445
5	,734	,060
Extraction Method: Principal Component Analysis.		

Questionnaire 2:

Source: Elling, S., Lentz, L. & De Jong, M. (2007)

Dimension	Number of items	Reliability
WEQ total	**32**	**.97**
1 Content	**10**	**.88**
Relevance	3	.72*
Comprehensibility	4	.75*
Comprehensiveness	3	.69
2 Navigation	**19**	**.96**
Ease of use	3	.90*
Structure	5	.80
Hyperlinks	6	.81*
Speed	2	.76
Search engine	3	.86
3 Layout	**3**	**.88**

* = one question removed



1.1 Relevance
*I find the information in this website helpful.
The information in this website is of little use to me.
This website offers information that I find useful.

1.2 Comprehensibility
*I think the information in this website is described clearly.
The language used in this website is easy to me.
I find the information in this website easy to understand.
I find many words in this website difficult to understand.

1.3 Comprehensiveness
Certain information I was looking for was missing in this website.
The website provides me with sufficient information.
I find the information in this website precise.

2.1 User friendliness
I find this website easy to use.
*I had difficulty using this website.
I consider this website user friendly.

2.2 Structure
I know where to find the information I need on this website.
I was constantly being redirected on this website while I was looking for information.
I always know where I am on this website.
I find the structure of this website clear.
The convenient set-up of the website helps me find the information I am looking for.

2.3 Hyperlinks (including Homepage)
The homepage clearly directs me towards the information I need.
The homepage immediately points me to the information I need.
*I find the homepage confusing.
*I think it is difficult to spot the hyperlinks on this website.
It is clear which hyperlink will lead to the information I am looking for.
Under the hyperlinks, I found the information I expected to find there.

2.4 Speed
I think it takes a long time to download a new web page from this site.
I think this is a fast website.

2.5 Search Option
The search option on this website helps me to find the right information quickly.
The search option on this website gives me useful results.
The search option on this website gives me too many irrelevant results.

3.0 Layout
I think this website looks unattractive.
I like the way this website looks.
I find the design of this website appealing.

Appendix 2

Chi – Square test between method and usability problem 5.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,269	1	,039		
Continuity Correction[b]	3,136	1	,077		
Likelihood Ratio	4,347	1	,037		
Fisher's Exact Test				,075	,038
Linear-by-Linear Association	4,180	1	,041		
N of Valid Cases	48				

Chi – Square test between method and usability problem 7.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8,084[a]	1	,004		
Continuity Correction[b]	6,189	1	,013		
Likelihood Ratio	9,058	1	,003		
Fisher's Exact Test				,010	,005
Linear-by-Linear Association	7,916	1	,005		
N of Valid Cases	48				

Chi – Square test between method and usability problem 8.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,581[a]	1	,018		
Continuity Correction[b]	3,572	1	,059		
Likelihood Ratio	7,514	1	,006		
Fisher's Exact Test				,050	,025
Linear-by-Linear Association	5,465	1	,019		
N of Valid Cases	48				

Chi – Square test between method and usability problem 15.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2,948[a]	1	,086		
Continuity Correction[b]	1,887	1	,170		
Likelihood Ratio	3,036	1	,081		
Fisher's Exact Test				,168	,084
Linear-by-Linear Association	2,887	1	,089		
N of Valid Cases	48				

Appendix 3, instructions

For CTA this is:

> *Denk hardop terwijl je de taken uitvoert. Je hoeft niet extra goed je best te doen, doe gewoon alsof ik niet aanwezig ben. Je kunt mij niet om hulp vragen, maar ik zal je wel eraan herinner hardop te blijven denken als je een tijdje stilvalt. Vergeet verder niet dat we niet jou, maar de website testen. Als er iets misgaat, is de website niet gemakkelijk genoeg, en ligt dat niet aan jou.'*

For RTE this is:

> *Ik zal zo hiernaast gaan zitten. Jij zult aan de hand van deze opdrachten door de website gaan. Je hoeft niet extra goed je best te doen, doe de opdrachten alsof je thuis bent. Je kunt mij niet om hulp vragen. Achteraf zullen we de vragen bespreken aan de hand van een schermopname, probeer je commentaar te onthouden tot dat moment. Vergeet verder niet dat we niet jou, maar de website testen. Als er iets misgaat, is de website niet gemakkelijk genoeg, en ligt dat niet aan jou.'*

Questions used during the usability tests.

> *"Wat gebeurde er nu?, Wat ziet u hier?, Wat vindt u hiervan?, Wat vindt u van de stappen die u moet ondernemen?, Was dit duidelijk voor u?, Waar verwachtte u deze informatie?, Vindt u deze informatie belangrijk?, Hoe komt dat?/ waar ligt dat aan?, Weet u nog wat u hier dacht?, Wat gebeurde er hier?, Waar was u nu naar op zoek?"*

Appendix 4, Uncovered usability problems, that where used in the comparison.

up.nr	Description	Category
1	Main Navigation, users find it hard to use the main navigation, they click on items when they should not, and they expect mouse over functions when they need to click.	2.1
2	Cost overview screen, users find it hard to understand what the final cost will be.	1.2
3	Cost overview screen, It is hard for users to understand that the "poliskosten" do not include "assurantie belasting", while it says: "inc. assurantie belasting".	1.2
4	Maximum insurance payment screen, There are no headers in the lower tables showing where the information belongs to.	1.3
5	Travel insurance product information screen, users find it hard to find the "Maximum insurance payment screen".	2.1
6	Maximum insurance payment screen, user are searching for the page header, they want information about which travel insurance the screen is for.	1.3
7	Maximum insurance payment screen, Users get confused about where they are on the website. The URL of the "doorlopende reisverzekering" says the information is from another travel insurance (kortlopende reisverzekering).	1.2
8	Calculation tool page, users find it hard (or are unable) to find the travel insurance comparison tool.	2.1
9	Travel insurance calculation page, The hyperlinks (on the right) to other travel insurance calculation tools is unclear.	2.3
10	Travel insurance product information screen, users are searching for the minimal contract duration, they find it hard to find or are unable to find this problem.	1.3
11	House Contents Insurance page, the tab that links to the calculation tool is normally called "berekenen" in this page it is called "aanvragen"	2.3
12	House Contents Insurance calculation tool, the question "huurwoning koopwoning" is shown in a gray colour. Users mistake this text colour for the indication the field is inactive.	3.0
13	When the user is looking for an insurance for jewellery, but could not find the information he was looking for.	2.2

14	House Contents Insurance calculation tool, the number of questions is to high, the user finds this inconvenient and prefers another path.	2.2
15	House Contents Insurance calculation tool, users find it hard to estimate the Cubic meters of their home and miss help/guidance in filling this question.	2.1
16	Contact page, the user is unable to find what the cost are for the emergency help service.	1.3
17	Contact page, The user does not know the meaning of "collect call"	1.2
18	"lijfrente garantie polis" product information page, the breadcrumbs do not lead trough the path that was taken to get to the information. This confuses the perception of the user about their whereabouts on the website.	2.2
19	Insurance overview page, it is not clear for the user where to find the overview page. The user expects the product overview to be in the main menu.	2.2