Usability Study

of the Negometrix web application, supplier side

Govert-Jan Slob
3219615
g.j.slob@students.uu.nl
30-11-2013

CONTENTS

1. About this Study 3
2. Introduction 3
3. Usability 3
   3.1. Measuring Usability 4
4. Method and Results 4
   4.1. Heuristic Evaluation 4
      4.1.1. The Negometrix heuristic evaluation 5
   4.2. Usability Testing 7
      4.2.1. The Negometrix usability test 8
   4.3. SUS Questionnaire 9
      4.3.1. The Negometrix SUS 9
5. Evaluation and Discussion 10
6. Conclusion 12
7. References 13
8. List of Appendices 15

1. About this Study

This document describes a usability study that was executed for Negometrix B.V. by Govert-Jan Slob from September to November 2013. The contact at Negometrix was Matthieu Hoogerwerf. This study and all its components are the property of the author, and distribution other than within Negometrix is prohibited without the consent of the author.

2. Introduction

Negometrix is a company located in De Meern that facilitates online digital tendering for companies in The Netherlands. Negometrix runs a web application that covers the whole process of tendering, from publication of the tender, to making offers to win a tender. By placing a tender on Negometrix' website, companies looking for offers can easily manage the tender in one place, and because Negometrix runs in the cloud, they can access their data anywhere. For companies looking to make an offer for a tender, the Negometrix web application provides a centralized location for finding tenders.

The Negometrix web application is the company's main source of income, and that income increases with the number of companies using the service. Consequently, it is of great importance that the usability of the application is high: a highly usable product helps to prevent people leaving the website prematurely. In the context of usability, the component of learnability is specifically critical. A characteristic of a publicly accessible web application is that it is being used by a highly diverse group of people with different backgrounds and properties. Consequently, computing skills and experience with similar systems will vary highly among the user group, making learnability paramount.

The aim of this study is to investigate the current usability of the web application, find bottlenecks and suggest improvement for the application. The first step in this process is a heuristic evaluation, carried out by a small number of expert evaluators. The usability issues found in the heuristic evaluation will be the input for the subsequent usability test, in which a group of potential users will carry out a task using the system. Finally, a qualitative appraisal of the application's usability will be collected via a questionnaire that users in the usability test are asked to fill in. Quantitative data is collected during the usability test so that baseline usability scores are produced, to which possible future test scores can be compared.

3. Usability

This study revolves around the notion of usability, and its consequences for the user experience of the Negometrix application. Usability is a term that is interpreted in different ways, even by people who are part of the science community. The most widely used, and industry accepted, definition of the concept of usability is that of the ISO 9142-11 standard, which defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO/IEC, 1998). More recently other authors have also attempted to define usability, of which Jacob Nielsen's take on the subject is probably the best known alternative. Nielsen simply defines usability as a quality attribute that characterizes how easy to use a product is. He

assigns five components to the construct of usability: learnability, efficiency, memorability, errors and satisfaction (Nielsen, 2012). An advantage of Nielsen's approach is that it is a bit more specific than the ISO version, possibly making it easier to use when trying to measure usability. In practice, the same metrics will probably be used to collect the data for both approaches. For this study I have used the ISO 9142-11 standard because it has been determined by international consensus and is also the common industry standard (ANSI, 2001). Also, it would have been too labor intensive to try and measure the memorability of the application because two usability test would be required, which does not fit the nature of the assignment.

3.1. Measuring Usability

Measuring usability according to the ISO standard means measuring efficiency, effectiveness and satisfaction. Efficiency is about being able to quickly complete tasks using the system, and is generally measured by recording the speed with which the application is used. During a usability test, recording the time it takes to complete a specific task is the best way to measure efficiency. Effectiveness concerns the ability to achieve goals as complete and accurate as possible, and is usually measured by counting completion rates and errors. Satisfaction can be documented by getting subjective assessments from users, using a post-test questionnaire such as The System Usability Scale (SUS) developed by Brooke (1996). It is also possible to generate a single usability score by combining the scores for efficiency, effectiveness and satisfaction. The 'Single Usability Metric' or SUM (Sauro & Kindlund, 2005) is an example of a single measure to score usability. Such a method can be helpful when comparing usability of a single product before and after changes have been made. However, producing a single score for the usability of the Negometrix application is not within the scope of this study. How efficiency, effectiveness and satisfaction have been measured for the Negometrix study can be found in the Test Plan Usability Test (Appendix C).

4. Method and Results

4.1. Heuristic Evaluation

The first step in this usability study is the execution of a heuristic evaluation (HE), which is a type of usability inspection method (UIM). Usability inspection methods are analytical evaluation methods that are conducted without the help of users, but with that of trained analysts. Because no end users have to be involved, no equipment has to be used and because only a few analysts are required, these methods are considered 'discount methods' (Cockton, Lavery & Woolrych, 2003b). UIMs are aimed at finding usability problems, helping to understand them and to support fixing them. The same authors explain that a UIM consists of three distinct phases, that are applied in fixed order. The first phase is analyst preparation, which is aimed at making sure the analysts understand the UIM, the system they will analyze and it's usage context. The second phase is candidate problem discovery, in which the UIM is applied to find possible problems. In the final phase, confirmation and elimination of candidate problems, the primary goal is to confirm issues as probable or improbable. The outcome of the three steps is a list of possible usability problems. Even though user-based methods are more reliable, UIMs like heuristic evaluation can still be of added value. It's possible to "anticipate user difficulties that may not emerge during testing", Woolrych, Cockton and Hindmarch (2004, p. 140) conclude. A weakness of

UIMs is that they are prone to deliver many false positives, and sometimes miss important usability problems (Cockton et al., 2003b).

The UIM of heuristic evaluation was developed by Jacob Nielsen in the nineties (Nielsen & Molich, 1990; Nielsen, 1992; Nielsen, 1994). In a heuristic evaluation an evaluator systematically examines the user interface to find usability problems by judging the user interface's compliance with a small set of 'rules of thumb'. These rules of thumb are called 'heuristics' and describe common properties a usable interface should have. The most used heuristics are the ten chosen by Nielsen (1994). These were picked from a larger set of heuristics because they provided the best explanation of the usability problem and because they detected the most issues. Using these guidelines, the evaluators compose a list of usability problems with references to the heuristics that are violated by those problems. The problems that are found also receive a severity rating on a scale of 1 (minor problem) to 4 (highly problematic). These severity ratings indicate how much impact a particular issue has on the usability of the system.

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

Figure 1: Nielsen's Heuristics

When it comes to detecting problems, a higher number of people involved as analysts usually brings about more detected problems. This is true for detecting usability problems by means of a heuristic evaluation too, but only to a certain extent. Nielsen and Landauer (1993) found that the increase in number of problems detected rapidly slows down when using more than five evaluators. When usability is critical, using more than five evaluators will pay off, but when budget is tight a number of five evaluators makes for the best cost-benefit ratio.

A major weakness of heuristic evaluation is its proneness to false positives. Consequently, to ensure maximum accuracy of the heuristic evaluation, the possible problems should be analyzed to either confirm them as probable problems or improbable problems. This analysis is made done using the usability test, which can confirm or refuse the issues from the heuristic evaluation. Still, a heuristic evaluation can produce valuable results: usability testing and expert reviews conducted with experienced analysts produce largely the same list of problems, several authors claim (Molich & Dumas 2008; Molich, Ede, Kaasgaard & Karyukin, 2004).

4.1.1. The Negometrix heuristic evaluation

A successful heuristic evaluation requires that a few prerequisites are in place. The most important aspect of a heuristic evaluation are the people performing it, the expert evaluators. To conduct an effective heuristic evaluation it is important that the evaluators are familiar with the method of HE so that the method is properly applied. Disparity in understandings of HE can lead to flawed predictions of usability issues. A good manual or tutorial can help mitigate these effects. Secondly, the evaluators must understand the system and its application domain. This knowledge is important to be able to make solid judgements about the impact of a possible usability problem, and a lack of such knowledge can lead to incorrect claims. Last but not least is the possession of a general knowledge of human-computer

interaction, which helps evaluators understand the heuristics and the possible influence of human characteristics.

For the Negometrix heuristic evaluation five evaluators were recruited from a group of post-graduate computer science students. The evaluators were required to have at least finished their Bachelor's degree and gained some experience with usability engineering during their education. To properly prepare the analysts, they were instructed and informed as follows. Before the analysts started the evaluation, they were briefed about the system, its context and its users. Additionally, they were given a brief explanation of the HE-method to complement tutorial material on HE they had received by e-mail earlier. The evaluators were also given a task to complete by using the system. This ensured a structured approach to the discovery of problems, which improves chances of avoiding false positives compared to an unstructured approach (Cockton, Woolrych, Hall, et al., 2003).

The heuristic evaluation resulted in a list of predicted problems detected by the analysts, including a severity rating for each problem, based on the rating system of Nielsen (1992). The results of the individual evaluators were aggregated into a master problem set that counted 75 issues. Finally, a final list of 30 issues that were given a severity rating of two or higher and issues that had been detected by more than one evaluator, was created. Differences in reporting style and problem interpretation within the group of analysts made this a challenging process. After analysis of the aggregated problem set, it became apparent that the majority of the detected problems could be explained with the help of just two heuristics:

1. *Recognition rather than recall.* According to this guideline, the user's memory load should be minimized by making objects, actions and options visible. The user should not have to remember information, but it should be easily inferred from the user interface itself. Many problems the evaluators identified where caused by the fact that the functions of many user interface objects, actions and options were not easily identifiable and could therefore not be recognized easily.

2. *Aesthetic and minimalist design.* This guideline propagates leaving out information that is irrelevant or rarely needed. Cause for the relatively high amount of issues connected to this guideline are the screens that are unclear because of the presence of unnecessary elements.

Other issues were connected to two other heuristics.

3. *Consistency and standards.* A smaller amount of issues were caused by ignoring standards and inconsistency across the application. For example, buttons were not always placed below or on the right side of content.

4. *Visibility of system status.* This guideline suggests always keeping the user informed about what is going on. Issues that defied this guideline were mostly caused by the absence of confirmation messages and other feedback regarding the status of the system.

The list of predicted problems that a HE produces is useful to get an overview of what problematic areas a system contains, but because the list is only of a predictive nature, the issues need to be tested during a test session with end users. Consequently, the contents of the final list of issues from the heuristic evaluation were taken into the usability test to be confirmed or to be rejected as a usability problem.

4.2. USABILITY TESTING

Usability testing is one of the most used usability methods today, and it is said by usability specialists that it has the largest impact on product improvement (Rosenbaum, Rohn & Humburg, 2000). During usability testing people who are representative of the target end user population are employed as participants and use the system to complete a realistic task. While performing the task, the participants are observed to collect empirical data about the usability of the product (Rubin & Chisnell, 2008). For a valid usability test it is important that the participants are end users or potential end users. If a test is conducted with other populations the results cannot be generalized to the target population. To find potential candidates for a usability test, a user profile with relevant characteristics that potential participants must match is often used. Another essential aspect of a good usability test is selecting the right tasks for the participants to perform. The tasks should include tasks that are performed frequently by end users, or those that are critical to accomplish the job. The tasks should also use areas of the system in which usability problems are expected (Cockton, Lavery & Woolrych, 2003a). The location of the test should also be considered because it may influence participant behaviour during testing. This is especially true for complex operational environments such as airplanes and hospital operating rooms. Testers of consumer software usually simulate the characteristics of the use environment, but there is no proof that it significantly influences the validity of the test (Dumas & Fox, 2007).

The number of users with the best cost-benefits ratio to use in a usability test is five (Virzi, 1992; Nielsen & Landauer, 1993; Faulkner, 2003). Testing with five users will help you find 85% of the usability issues present in the system. Adding to up to fifteen users will help to discover more usability issues, but each new issue will have relatively cost a lot and therefore Nielsen recommends testing multiple times instead of testing with more users. Because of time limitations, it is often impossible to develop tasks to test all components of a design. Components that are not included in the tasks are not being evaluated. This limitation can somewhat be mitigated by combining a usability test with a thorough evaluation technique such as heuristic evaluation (Cockton, Lavery & Woolrych, 2003a).

During usability tests, the 'think-aloud protocol' is often used to detect problems by gaining insight in the cognitive process of the user. Using the think-aloud protocol, users are asked to say out loud what they are thinking, looking at, clicking on, feeling etcetera. The think-aloud protocol was developed as a theoretical framework by Ericsson and Simon (1982), while Boren and Ramey (2000) proposed a new framework more specifically targeted at usability testing. The Boren and Ramey framework is the best documented and makes for the most natural interaction with the participant, while Ericsson and Simon's framework has the least influence on performance times (Krahmer & Ummelen, 2004). Benefits of using the think-aloud protocol are that it is cheap, easy to learn and can be applied flexibly. A possible downside of the think aloud method in usability tests, are the longer performance times that it may cause.

Data collection during a usability test is done on a quantitative and a qualitative level. Quantitative data collected are errors, task times and success rates. Qualitative data are user utterances like statements of frustration, and user opinions. These data are the basis for detecting usability problems, and multiple data types are often used to support a finding.

4.2.1. The Negometrix usability test

In order to find out which usability issues are really experienced by end users, a usability test was conducted. At the beginning of the test, five participants with characteristics matching those of the end users were recruited from a group of relatives and acquaintances. Special attention was paid to their computer skills and experience with similar systems, which had to be low. The users had never used the system before, so learning effects would not be present. In order for the test situation to resemble the real use environment as much as possible, users had to complete a task that is similar to the task that end users use the system for. Before commencing the tasks the participants were asked to read a scenario that explained the role they were assuming. While the users performed the tasks they were observed by the facilitator, who used a second monitor to view the user's progress and collect data. The idea behind this approach was to make the user feel as comfortable as possible and to cause as little hindrance as possible. Other approaches place the facilitator right behind the user, possibly causing the user to feel uncomfortable. The tests took place at the home location of the facilitator, because this was the most convenient place for both the users and the facilitator.

The purpose of the usability test was to discover usability problems and their cause, and to write recommendations to fix them. Moreover, the test would provide answers to the research questions stated in the test plan, and produce baseline usability scores that can be used to compare future usability tests with. The data necessary to achieve these goals was collected by having users think aloud and by recording the screen output of the sessions. The measures used for the data collection are stated in the Test Plan Usability Test, along with the research questions (Appendix C).

Image 1: The Negometrix usability test in progress

The quantitative and qualitative data that the usability test resulted in, was analyzed for usability issues and a list of sixteen problems was drafted. Each problem was analyzed for its cause, and a recommendation to solve the problem was proposed. Problems were also given a severity rating to depict the impact that a particular issue can have on the usability of the system, using the rating system of Rubin and Chisnell (2008) which uses a scale of 1 to 4. Because most analysts miss the larger part of the problems in video data (Jacobsen, Hertzum, & John, 1998) this list cannot be seen as a complete listing of all usability problems present in the system. The quantitative data, the list of usability problems and recommendations, as well as the answers to the research questions can be found in the Results Usability Test document (Appendix D).

A relatively low percentage of issues from the heuristic evaluation could be confirmed as usability problems during the usability test. Only 11 of the 30 issues were experienced by the participants in the usability test. This however does not mean that the other 19 issues are all false positives, they were just not detected in this particular test. A possible explanation for the discrepancy can be found in the explorative nature of the HE, in which more parts of the user interface are used than during a usability test. Secondly, the fact that an expert evaluator is a different type of user with different goals, and is

specifically looking for problems, may have caused the evaluators to detect much more issues than the facilitator in the usability test. And last but not least, the use of a different type of computer technology (laptops) by the evaluators could have created the discrepancies.

	Task completion rate (1,2)	Task completion time (1,2)	Critical errors	Non critical errors
User 1	100%,100%	20,23, 26;55	1	8
User 2	0%,100%	41:40, 45:00	1	9
User 3	0%,100%	22:50, 27:20	1	4
User 4	100%, 100%	20:30, 22:46	0	6
User 5	100%,100%	26:20, 29:45	0	8
Average	60%, 100%	26:00, 33:00	1,4	7

Table 1: Quantitative data from the usability test.

4.3. SUS QUESTIONNAIRE

The final stage of this usability study is the execution of a questionnaire to collect subjective data about the usability of the application. These data can provide metrics to rate the satisfaction part of the usability construct. The questionnaire that was used in this phase is the System Usability Scale (SUS). The SUS scale gives a subjective assessment of usability: how do users perceive the usability of a system after having used it? In the SUS, a five point Likert scale is used to measure satisfaction for ten different items. The scores for these items are then combined into a final score that displays the system's usability. The questionnaire was developed by John Brooke (1996), in order to provide an answer to the need for subjective opinions about a system's usability in a time where the focus lay on objective measures of efficiency and effectiveness (Brooke, 2013). The main reason to develop the SUS scale was the need to compare the usability of systems. Efficiency and effectiveness are so much influenced by the context of use that these cannot be compared between systems. Satisfaction, according to Brooke, did not suffer so much from sensitiveness to context of use and can be compared between systems. After having used the SUS in 500 (Sauro, 2011) and 1000 (Bangor, Kortum & Miller, 2009) studies, several authors found that the SUS scale was a reliable and valid measurement tool. The average SUS score for systems was 68 points (Sauro, 2011).

4.3.1. The Negometrix SUS

For the Negometrix usability study, a SUS questionnaire was presented to the participants after they had completed the tasks for the usability testing phase. The result is an average score of 58,5 points, which is a below average score. The scores were in the range of what was expected based on the usability tests that had been conducted just before the questionnaire was used. The results reflect the general view that the system's usability is of a below average nature. The system can be picked up and used by new users with novice computer skills, but there is also much room for improvement.

The ten items were not meant to be used individually, but it is worth taking a look at them to see if something stands out. The most significant results come from items 4, 7 and 10. The participants think the system can be used without the help of a technical person and they think that the system can be learned quickly by most people. Because of the small sample size in this usability study – only 5 participants were used - the SUS scores cannot be generalized to the population. Furthermore, the results of the individual items should be given a limited value, because they were not originally meant to

be counted as single constructs. The results do however give an indication of the satisfaction that the system would score in a larger sample with the same characteristics. Another reason to pay attention to the obtained scores, is their proximity to the results from the heuristic evaluation and the usability test.

Item	Average score (1-5)
1. I think that I would like to use this system frequently	3,2
2. I found the system unnecessarily complex	3
3. I thought the system was easy to use	3
4. I think that I would need the support of a technical person to be able to use this system	1,8
5. I found the various functions in this system were well integrated	3
6. I thought there was too much inconsistency in this system	2,8
7. I would imagine that most people would learn to use this system very quickly	3,4
8. I found the system very cumbersome to use	2,8
9. I felt very confident using the system	3
10. I needed to learn a lot of things before I could get going with this system	1,8

Table 2: Question averages for the Negometrix SUS

5. Evaluation and Discussion

To assess the value of this study, the quality of the heuristic evaluation and the usability test are good measures. The quality of a heuristic evaluation or a usability test is measured by their effectiveness, which in turn is determined by the thoroughness and the validity of the evaluation or test (Sears, 1997). According Sears, validity is measured by: number of real usability problems found by UIM/number of problems predicted by UIM. Thoroughness is: Number of real usability problems found by UIM/number of real problems that exist.

Ideally one wants a heuristic evaluation to predict all, but only genuine usability problems, and not generate many issues that do not actually exist. Unfortunately it is hard to measure the effectiveness and say anything useful about these constructs, because thoroughness and validity are usually determined in a subjective manner, which is also the case with the Negometrix HE. The subjectivity lies in the choice of abstraction level with which the usability issues have been reported (Cockton & Lavery, 1999). It is hard to determine how many issues are true positives if there is a big difference between the granularity in the reports from the different evaluators. For example, a particular area in a UI can contain more usability problems in a specific reporting style than in an abstract reporting style. Furthermore, measuring thoroughness seems impossible because one will only be able to find a small portion of all the usability flaws that exist, never knowing the total amount of issues present. (Molich et al., 2004). It has been reported (Woolrych, Cockton & Hindmarch, 2004) that a structured problem

report format (SPRF) can improve validity and appropriateness, with a smaller improvement in thoroughness.

The quality of a usability test can be measured in a similar way as the quality of a heuristic evaluation; using effectiveness and validity as identifiers. Again, the choice of abstraction levels makes measuring these constructs problematic. Furthermore, assessing the validity of a usability test is also hindered by the lack of agreement among usability professionals about what the goals of a usability test are. In the case of usability tests there is another key characteristic that has received a lot of attention: reliability, i.e. the ability to repeatedly find the same problems. Many experiments and analysis papers have been written about the reliability of usability testing, and many came to the same conclusion; that a poor overlap exists among the usability problems detected by different usability teams. Hertzum and Jacobsen (2001), who call this phenomenon the 'evaluator effect', mention that evaluating and testing a user interface is a cognitive activity which is very subjective. Other causes for the evaluator effect are the use of different methods, tasks, problem criteria and reporting styles by usability teams (Molich et al., 2004).

Because of the difficulties mentioned and because it was not within the scope of this study, validity and thoroughness have not been measured for the usability test and the heuristic evaluation. At this point it is also impossible to say anything useful about the reliability of the Negometrix usability test because it has not been performed a second time, using the same product, tasks, location and environment.

Each study has its strong and weak aspects, and these elements can contribute to harm the quality of the study. The ability to perform a high quality HE was influenced by the fact that heuristic evaluation is a poorly documented usability inspection method. There is hardly any material that contains instructions or guidelines on how a heuristic evaluation should be performed. This made it hard to use a standardized approach for all the evaluators. Likewise, the skills of the analysts will probably have had an impact on the thoroughness of the HE. The analysts had little experience with heuristic evaluations, and this may have influenced the quality of the HE, as they could have missed important issues or reported more false positives due to a lack of experience.

Elements that could have had a negative influence on the quality of the usability test are the skills of the evaluator. The usability test in this study was the first usability test the facilitator had conducted. Furthermore, the facilitator intervened a few times during the test when participants struggled, and these 'rescue' actions may have prevented understanding why participants struggle or how they recover. Finally, the test took place in a different type of environment than the end use environment.

The Negometrix study also has its strong points, that can positively influence the quality of the study. First in line is the fact that this study combined a heuristic evaluation and a usability study. This helps detect more usability issues and improves the ability to detect true positives. Consequently, the 11 issues that surfaced in both the HE and the usability test can be quite valuable. A second strength is the amount of planning that went into this study. Finally, the gathering of both quantitative and qualitative data is a strength.

For a future usability study, the use of a structured report format (SPRF) can assist in maintaining a certain granularity level, and improve the heuristic evaluation. Also recommended, is the use of real usability experts as evaluators.

6. CONCLUSION

The goal of this usability study was to discover usability problems in the supplier section of the Negometrix application, and to provide recommendations for fixing those problems. Preferably, the study had to be executed according to scientific literature on the subject. The study started with a heuristic evaluation, in which five experts judged the user interface on compliance with ten guidelines. After selecting the issues that were rated two and higher or that were detected by at least two analysts, 30 issues remained. The majority of the issues could be connected to two of Nielsen's guidelines: 'recognition rather than recall' and 'aesthetic and minimalist design'. Consequently, Negometrix should strive to help users identify user interface objects' functions and try to remove any unnecessary elements from the UI. Most of the other issues had to do with 'consistency and standards', and 'visibility of system status', meaning that the user's natural work process should be supported more, and that more system feedback should be introduced. Secondly, a usability test using five participants with characteristics similar to those of the end users was conducted. Users were observed by the facilitator on a second monitor while the participants performed a realistic task. The data collected consisted of task times, errors, completion rates and user utterances. With the help of the usability test, 11 of the 30 issues found in the heuristic evaluation were confirmed as real issues. The usability test itself brought to light 16 issues that had a severity rating of two and higher or that were experienced by multiple participants, and 11 other issues. The issues the participants experienced in the usability test had many different causes, among which were visual clutter, ignoring the user's natural task flow, deviating from standards, and terminology problems. Finally, the SUS questionnaire provided a satisfaction based usability score of 58.5 points, which is slightly below the average.

This usability study, has provided Negometrix with valuable information about how to improve the usability, and thus the user experience. The issues that have been found by both the analysts in the HE and the participants in the usability test are the best candidates to be fixed, because they received the most verification. Finding the usability issues, was just the first step towards improving the user experience of the application. The next step is applying the improvements, which will most likely improve the usability of the system, and also reduce the help desk costs because users will need less assistance. The author has tried to provide recommendations for each of the most severe usability issues to help Negometrix take the appropriate action. These recommendations can be found in Appendix D. It is recommended that the improvements that are made based on this study are also tested during a future usability test to assess their effect. The quantitative data collected during this study can then be helpful for comparing two versions of the system.

7. REFERENCES

1. ANSI. (2001). *Common industry format for usability test reports* (ANSI-NCITS 354-2001). Washington, DC: American National Standards Institute.

2. Bangor, A., Kortum, P.T., & Miller, J.T. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies, 4(3),* pp. 114-123.

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

4. Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In Jordan, P.W., Thomas, B., Weerdmeester, B. A. & McClelland, A. L. (Eds.) *Usability Evaluation in Industry*. London, England: Taylor and Francis.

5. Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies 8 (2),* pp. 29-40.

6. Cockton, G., Lavery, D., & Woolrych, A. (2003a). Usability testing: current practice and future directions. In Jacko, J. A. & Sears, A. (Eds.), *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications archive.* Lawrence Erlbaum Associates, Inc., pp. 1118-1138.

7. Cockton, G., Lavery, D., & Woolrych, A. (2003b). Inspection-based evaluations. In Jacko, J. A. & Sears, A. (Eds.), *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications archive,* pp. 1118-1138. Lawrence Erlbaum Associates, Inc.

8. Cockton, G., & Lavery, D. (1999). A Framework for Usability Problem Extraction. In Sasse, A., M. & Johnson, C. (Eds.) *Proceedings of Interact 99*, Amsterdam, Netherlands: IOS press.

9. Cockton, G., Woolrych, A., Hall, L. & Hindmarch, M. (2003). Chaning Analysts' Tunes: The Surprising Impact of a New Instrument for Usability Inspection Method Assessment*,* In Palanque, P., Johnson, P. & O'Neill, E. (Eds.), *People and Computers 17: Designing for Society* pp. 145-162. Germany: Springer-Verlag.

10. Dumas, J. S., & Fox, J. E. (2007). Usability Testing: Current practice and future directions. In Jacko, J. A. & Sears, A. (Eds.), *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications archive.* Lawrence Erlbaum Associates, Inc.

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

12. Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments & Computers, 35*(3), 379-383.

13. ISO/IEC. (1998). *9241-11 Ergonomic requirements for office work with visual display terminals (VDT)s - Part 11 Guidance on usability*.

14. Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability tests. In Karat, C.M., & Lund, A. (Eds.), *Proceedings of the Conference of Human factors in Computing Systems* (CHI`98), (pp. 255-256), New York, NY: ACM Press.

15. Krahmer, E., & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing, IEEE Transactions on Professional Communication, 47 (2), 105-117.

16. Molich, R. & Dumas, J. S. (2008). Comparative Usability Evaluation (CUE-4). *Behaviour & Information Technology,* 27(3).

17. Molich, R., Ede, M. R., Kaasgaard, K. & Karyukin, B. (2004). Comparative Usability Evaluation. *Behaviour & Information Technology*, *23(1),* pp. 65-74.

18. Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In Chew, J.C. & Whiteside, J. (Eds.), *Proceedings of the CHI '99 Conference on Human Factors in Computing Systems* (pp.249-256). New York, NY: ACM Press.

19. Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In Bauersfeld, P., Bennet, J., & Lynch, G. (Eds.), *Proceedings of the CHI'92 Conference of Human factors in Computing Systems* (pp. 373-380). New York, NY:ACM Press

20. Nielsen, J., & Landauer,T., K. (1993). A mathematical model of the finding of usability problems. In Ashlund, Mullet, Henderson, Hollnagel & White (Eds.), *Proceedings of the CHI'93 Conference of Human factors in Computing Systems* (pp. 206-213). New York, NY: ACM Press

21. Nielsen, J. (1994b). Heuristic Evaluation. In Nielsen, J. & Mack, R. L. (Eds.) *Usability Inspection Methods,* (pp. 25-62). Hoboken, NJ: John Wiley & Sons,.

22. Nielsen, J. (2012). Usability 101: *Introduction to Usability.* Retrieved from http://www.nngroup.com/articles/usability-101-introduction-to-usability/

23. Rosenbaum, S., Rohn, J., & Humburg, J. (2000). A toolkit for strategic usability: results from workshops, panels, and surveys. In Turner, Szwillus, Czerwinski, Peterno & Pemberton (Eds.), *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems* (pp. 337-344). New York, NY: ACM Press.

24. Rubin, J. & Chisnell, D. (2008) *Handbook of Usability Testing* Indianapolis. Hoboken, NJ: Wiley publishing Inc.

25. Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In Kellog, W. & Zhai, S. (Eds.), *Proceedings of the CHI '05 Conference of Human factors in Computing Systems,* (pp. 401-409). New York, NY: AMC Press.

26. Sauro, J. (2011). A practical guide to the System Usability Scale: Background, benchmarks, & best practices. Denver, CO: Measuring Usability LLC.

27. Sears, A. (1997). Heuristic Walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction, 9*(3), pp. 213-234.

28. Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34,* pp. 457-468.

29. Woolrych, A., Cockton, G. & Hindmarch, M. (2004). Falsification testing for usability inspection method assessment. In Dearden, A., & Watts, L. (Eds.), *Proceedings of the CHI '04 Conference of Human factors in Computing Systems,* (pp. 137-140). New York, NY: ACM Press.

8. List of Appendices

Appendix A – Manual HE

This appendix contains the manual that was given to the expert evaluators in the heuristic evaluation.

Appendix B – Results HE

This appendix contains the results of the heuristic evaluation: a list of issues with severity ratings and accompanying heuristics.

Appendix C – Test Plan Usability Test

This appendix contains the test plan for the Negometrix usability test. Among others, it contains the participant characteristics, usability goals, measures and research questions.

Appendix D - Results Usability Test

This appendix contains a list of usability issues that were detected by the users during the usability test. It also contains severity ratings, and proposed solutions for the issues.

Appendix E – SUS Sheet

This appendix contains the System Usability Scale questionnaire that was used to gather satisfaction data.