Reliability of Usability Evaluation Methods

Vignesh Krubai

Bentley University

Introduction

Trochim(2006) describes reliability as “In research, the term reliability means "repeatability" or "consistency". A measure is considered reliable if it would give us the same result over and over again (assuming that what we are measuring isn't changing)”

Gray & Salzman (1998) defined Usability Evaluation Methods (UEM) from a Human-Computer Interaction perspective as “methods used to evaluate the interaction of the human with the computer for the purpose of identifying aspects of this interaction that can be improved to increase usability.” Some UEMs can also be applied to other systems beyond the computer.

The following are a couple of most commonly used usability evaluation methods:

Usability Testing

Usability Testing is the method in which a participant is made to interact with the system either on specific tasks or an overall product depending on the goal of the study. Participants can be asked to explain what is going through their mind while performing the task using the ‘thinking out loud’ method. They can also debrief with the moderator of the test at the end of the session orally or through predefined questionnaires.

Expert Reviews

Expert reviews are done by specialists in the usability field. They evaluate the system based on the guidelines they have developed through personal experience or they may use the standard heuristic guidelines as described by Nielsen & Molich(1990), ISO 9241 standards, persona-based review etc.

This paper will discuss the reliability issues of the methods used to assess the usability of a system and the reasons for this. From looking at all the existing research done on the topic, the takeaway for practitioners in the field will be discussed.

Comparative Usability Evaluation

The Comparative Usability Evaluations were performed in order to verify the consistency of usability studies. Different evaluators use evaluation methods and the lack of standard classification of results can cause unique interpretation of results. This difference in identification of problem and assignment of severity ratings is termed as evaluator effect.(Hertzum & Jacobsen, 2001) Initially the evaluator effect was only tested for heuristic evaluations but later studies done to test the evaluator effect in usability testing showed that only a fifth of the problems were detected by all users and about half of the problems were identified by single evaluators.(Jacobsen, Hertzum, & John, 1998)

Molich et al. (1998) attempted to determine the level of consistency across different evaluators testing the same system. In the first Comparative Usability Evaluation (CUE) test, four labs separately tested a Windows calendar management application and between them, 141 usability problems were identified. A striking revelation from this test was that only one problem was reported by all four labs. In all, 91% of the results were uniquely identified by individual labs. The method of obtaining usability test results also differed with some labs preferring a qualitative approach whereas others focused on qualitative aspects such as product satisfaction and approval etc.

Molich, Ede, Kaasgaard, & Karyukin(2004) conducted the second comparative usability evaluation to reaffirm the findings from CUE-1 as well as correct the shortcomings of CUE-1. In this study nine teams performed usability evaluation tests on www.hotmail.com. The teams were selected so that there was a healthy mix of experience levels and cultural variations. The teams were also given the goals of the usability test but were not restricted in the method that they used to conduct the test. Having goals gave the usability teams an opportunity to interact with the development team in order to understand and discuss the goals. The results showed that not even a signal problem was commonly identified by all nine teams and 71% of the problems were uniquely identified. At the end of the CUE – 2 tests the authors acknowledged that since the usability testing goals were very broad, the teams had time to test only a specific set of tasks which they felt were important. This resulted in different teams testing different aspects and hence identified fewer problems in common. Also since the websites are content-heavy there was potential scope for identifying more problems than if it was a simple website. Another study by Kessner, Wood, Dillon, & West (2001) illustrates this example since they studied a simple dialog box and found that only 44% of the problems were identified uniquely. Molich et al.(2004) also refuted that the use of inexperienced evaluators had an effect on the result of the study.

The CUE- 3 study was done by eleven usability experts individually who assessed the www.avis.com website(Hertzum, Jacobsen, & Molich, 2002). They performed different studies on using different forms of heuristic evaluation and afterwards met in groups to discuss the usability problems. In compiling the results, Hertzum et al.(2002) found that only one error was found by all eleven evaluators with 79% of the problems reported by one or two evaluators. An interesting outcome of the CUE-3 tests was that when the evaluators sat in groups and discussed the problems they had identified, even if the problems were different, this seemed to reinforce that they had made the correct evaluation. Hertzum et al.(2002) attributed this to the evaluators categorizing the problems into groups and if two evaluators had errors in the same category, they considered their evaluations to be accurate. This reduced their perceived impact of evaluator effect and they may incorrectly assume that a single evaluation is sufficient**.**(Hertzum et al., 2002)

The CUE – 4 studies used a mix of nine teams conducting usability tests and eight teams conducting expert reviews(Molich & Dumas, 2008). These teams performed evaluations on the Hotel Pennsylvania website. They were free to choose any evaluation method of their choice; however they had to classify all usability issues as per the categories specified by the authors of the study. The results of the CUE-4 study validated the conclusion from the previous CUE tests that there is a lot of variability in identification of problems since 60% of the problems was reported by single teams only. No problem was reported by all the teams and the average percentage of overlap was 11.5%. Molich & Dumas (2008) also identified that there were occasions when key issues went unnoticed. There were also instances when contradictory assessments were given to the same feature that was tested.

From the four Comparative Usability Evaluation studies, the main takeaways that were observed by Molich & Dumas (2008) was that usability tests have a wide range of variability across evaluators and once we start looking for problems we can find an enormous number of problems. Fixing all these problems is not practically feasible for the team that is redesigning the system based on the results. The studies also emphasized the importance of working in conjunction with the developers so that there is higher productivity in the usability analysis done. They also identified that there is lot of scope for improvement of the classification schemes.

Further Research based on CUE studies

Since there were a high number of uniquely identified problems in the CUE studies, Lindgaard & Chattratichart (2007) hypothesized that if there were more users then more problems could be found. The results that they analyzed from the CUE-4 studies did not reveal any correlation between the number of users and number of problems found. They also attempted to determine relation between the number of user tasks and the proportion of problems found. They were able to find evidence in the CUE – 4 results to show that a higher number of user tasks produced better results. Based on these observations it can be concluded that if we had to choose between the number of users and the number of tasks it would be more beneficial to have more tasks with fewer users. However the number of users must also be balanced such that there is positive return on investment.(Lindgaard & Chattratichart, 2007)

Gray & Salzman (1998) argued that there was very less importance given to the design of evaluations. They looked at different UEM studies and identified key weakness in the validity of those studies. Hartson, Andre, & Williges(2001) felt that the reason for this irregularity in results was due to the lack of a standard metrics, measures and criteria for comparison. They were mindful of the fact that creating an ‘ultimate criterion’ was very difficult since it is not possible to have a one size fits all evaluation method in the real world. However they set the ball rolling in the direction of more standardized evaluation methods by identifying measures such as thoroughness, validity, effectiveness, reliability downstream utility and cost effectiveness.(Hartson et al., 2001) These indicators were aimed at getting at the root of the differences in measures. They would also be useful in helping the evaluators put in perspective the degree of variability in user evaluation methods and they may then modify their evaluation methods to provide more credible results.

Practitioner's Takeaway

Now that it has been emphatically proven that there is a lot of variability in usability evaluation methods, Wilson(2006) proposed the approach of triangulation to convert the variability into an asset in usability evaluations. He proposed that a set of different evaluations can be performed on the same system and the results from all the studies can be combined to identify common occurrences of issues. The number of occurrences of an issue across different evaluation methods can also be used to categorize its severity. The two kinds of triangulation that can be used are: between-methods triangulations where different tools like focus groups, usability testing and questionnaires are used within the same test and within-method triangulation where the facilitator, observer, user-group or geographic location is varied across different sessions of the test so that a more holistic result is obtained(Wilson, 2006). The disadvantage of this method is that it could take additional time and resources to perform and this is not a luxury that most organizations have.

The CUE studies have also proven that there are benefits in speaking to the developer. Similarly it would also be beneficial in keeping in mind the business goals of the system that is being tested. This would also give a new perspective to designing the usability evaluation method and the final report. Hence a combination of business and development goals can be used in setting the context for the usability evaluation. This can be effectively used along with Wilson's (2006) method of triangulation to create and implement a fruitful usability evaluation of the system or product that is being studied.

References

Gray, W., & Salzman, M. (1998). Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction*, *13*(3), 203-261.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria For Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, *13*(4), 373-410.

Hertzum, Morten, & Jacobsen, N. E. (2001). The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, *13*(4), 421-443.

Hertzum, M., Jacobsen, N. E., & Molich, R. (2002). Usability inspections by groups of specialists: perceived agreement in spite of disparate observations. *CHI'02 extended abstracts on Human factors in computing systems* (pp. 662–663). ACM.

Jacobsen, Niels Ebbe, Hertzum, M., & John, B. E. (1998). The evaluator effect in usability tests. *CHI 98 conference summary on Human factors in computing systems - CHI '98*, (APRIL), 255-256. New York, New York, USA: ACM Press.

Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the reliability of usability testing. *CHI'01 extended abstracts on Human factors in computing systems* (pp. 97–98). ACM.

Lindgaard, G., & Chattratichart, J. (2007). Usability testing: What have we overlooked? *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1415–1424). ACM.

Molich, R., Bevan, N., Butler, S., Curson, I., Kindlund, E., Kirakowski, J., & Miller, D. (1998). Comparitive Evaluation of Usability Tests. *Usability Professionals Association 1998 Conference* (pp. 189-200).

Molich, Rolf, & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*(3), 263-281.

Molich, Rolf, Ede, M., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology*, *23*(1), 65-74.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *ChI 90 Proceedings* (Vol. 17, pp. 249-256). ACM.

Trochim, W. (2006). Research Methods - Knowledge Base. *Web Center for Social Research Methods*. Retrieved October 30, 2011, from http://www.socialresearchmethods.net/kb/reliable.php

Wilson, C. E. (2006). Triangulation: the explicit use of multiple methods, measures, and approaches for determining core issues in product development. *interactions*, *13*(6), 46–63. ACM.