CptS 543 Assignment #1

Critical Review of The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods

Jessamyn Dahmen

2/9/2016

Summary. This article analyzes three established usability evaluation methods (UEMs) in terms of "the evaluator effect." The authors define the evaluator effect as the general differences in how usability evaluators detect and rate usability problems within a system when using one of the three UEMs. The three UEMs evaluated were cognitive walkthrough (CW), heuristic evaluation (HE), and thinking-aloud study (TA). To analyze the evaluator effect the authors examined the results of several previous studies using these UEMs that specifically explored the evaluator effect as well as some that provided the necessary results but not explicitly look at evaluator effect. For their basic measure of the evaluator effect the authors preferred the *any-two-agreement* measures over the *detection rate* measure.

The authors found that there is no one "best" UEM and that each method demonstrates different strengths and weaknesses in terms of the evaluator effect. The main contributions of this paper to the literature are to show how each UEM is affected by the evaluator effect and also demonstrate that using more than one evaluator is preferable to using more users and only one evaluator. The authors identified three aspects of the UEMs analyzed that contribute to the evaluator effect: vague goal analysis, vague evaluation procedures, and vague problem criteria. When trying to identify the severity and number of usability problems they found that multiple evaluators are beneficial due to the variability that consistently occurs between evaluators in the results they examined. It was also concluded that the evaluator effect will need to be mitigated but is likely not possible to eliminate. Finally, the researchers concluded that while each UEM has weaknesses in terms of evaluator effect, they were still among some of the best usability evaluation techniques available at the time.

Critical Review. There are two listed authors for this paper: Morten Hertzum and Niels Ebbe Jacobsen. The former was associated with the Center for Human-Machine Interaction in Riso National Laboratory and the latter worked with Nokia Mobile Phones, both in Denmark. Morten Hertzhum has earned both a Master's and PhD in Computer Science from the University of Copenhagen and has extensive experience in both industry and academia. It appears that Hertzhum did not start looking at the evaluator effect specifically until about 1998, around 3 years prior to this paper's publication, although he did have other publications related to Human Computer Interaction (HCI). Niels Ebbe Jacobsen has high levels degrees in Computer Science, Psychology, and Business. Jacobsen also has experience in industry and academia, with an emphasis on industry and HCI. It appears that Jacobsen also started looking at the evaluator effect specifically around 1998, often co-publishing papers with Hertzum.

Both authors appear to be well established with extensive experience in this area of research even prior to this paper's publication. In their citations the authors draw on research from a variety of perspectives and countries, although these perspectives seem to be primarily limited to people based out of Europe and the United States. This may bias the authors in terms of the differences in how UEMs are administered across different countries. Potentially they could be missing information on how evaluator effect can be mitigated, or effective ways to address the three problems related to vagueness. It may be the case however, that at the time of publication this area of research was new enough to not have an extensive body of literature on which to draw upon from people other than several experts based in Europe and the United States.

One weakness of the paper that is in a way acknowledged by the authors is that the evaluator effect is a measure of reliability only. This implies that studying the evaluator effect alone only addresses the problems associated with studying the extent to which independent evaluations produce similar results. It does not deal with validity, specifically the extent to which the problems identified during a usability study show up in real world-use. To strengthen the contributions of this paper it would have been ideal if the authors looked at both a measure of reliability and validity.

Another weakness is related to the sample sizes of the studies that were used by the authors. As Table 1 indicates several of the studies examined by the authors had a very small sample of evaluators, with the highest number being 77 individuals and two studies using 6 and 3 laboratories respectively with an unspecified number of individuals. The low number of evaluators in many of the studies is an issue that the authors point out several times throughout the paper. However, it may have biased the extent to which evaluators differ, especially their calculated average range of difference between two evaluators (5% to 65%). Furthermore, to measure evaluator effect the authors examine two different measures, *detection rate* and *any-two agreement*. Each measure has its own set of drawbacks but the authors state a preference toward the latter measure. Unfortunately, some of the studies they used did not have sufficient data to calculate *any-two agreement.* Also, there is no discussion if there are other ways to quantitatively measure evaluator effect beyond those two methods.

Based on the description of the studies evaluated the authors tried to compare a variety of different studies using different evaluation methods and make general cohesive conclusions based on their findings. One strength of the paper is that the authors do examine how differences between the procedures for each UEM could affect outcome. However, the studies they examined very greatly in terms of number and type of evaluators, evaluated system, and UEM used. There does not seem to be any same system that was evaluated using all three techniques and similar evaluators. Trying to compare these different studies without a more balanced representation of different techniques and types of evaluators may result in many variables affecting outcome in addition to UEM procedures that the authors did not discuss as thoroughly.

Integration with Related Work. In terms of the work preceding this work this paper offers several novel insights. In comparison to Jacobsen and Hertzum's (1998) previous publication on the subject of evaluator effect this paper offered a much more in depth analysis and discussion drawing on a larger sample size of evaluators and different UEMs. In a larger context this paper also seems to have been one of the first studies to specifically analyze evaluator affect across all the most popular UEMs of the time. As demonstrated by a review comparing UEMs written by Gray and Salzman (1998) evaluator effect was often indirectly studied but explicitly addressed until the time of this paper's publication.

After this paper was published it has been cited a large amount of times by several studies examining UEMs and other aspects of usability in general. In some studies such as the one conducted by Hvannberg, Law, and Larusdottir (2007) the researchers are skeptical that differences in usability evaluations, specifically Heuristic Evaluations, are completely due to the evaluator effect. This skepticism is based on the small sample of evaluators used in this study. Another subsequent study conducted by Vatrapu and Perez-Quinones (2006) explored how differences in cultures can impact the outcomes of usability studies. The results of this study seemed to support the findings of Jacobsen and Hertzum, especially when considering evaluators interviewing users from the same culture compared to evaluators interviewing users from different cultures. However these findings were not entirely supportive of Jacobsen's and Hertzum's paper as the evaluators did not make any judgment decisions about usability problems.

It would seem that the small sample size used in this study may bring into question the actual significance of evaluator effect in terms of affecting usability testing outcome. Nevertheless, it would appear that much of the subsequent literature that cites Jacobsen and Hertzman's study does acknowledge that the evaluator effect exists and can impact the reliability of UEMs.

Implications for HCI. A major implication for HCI researchers that is brought up by this study and related work is that the validity of the UEMs has not been extensively studied enough. Even if researchers can establish methods that reduce differences between evaluators, does doing this help each UEM detect problems that are useful to addresses in real world settings? Evaluators may be able to more consistently detect similar problems but they may not be the ones that matter to actual users. Another implication mentioned by the paper for researchers deals with the lack of studies that examine whether evaluators are consistent across evaluations. It may be the case that the differences measured by the evaluator effect are not based fully on "true" disagreement between evaluators, but inconsistencies in an individual evaluator’s performance and abilities. Finally, another implication that this paper touches upon is how reliable the measures of evaluator effect are. It seems at the time of publication there were only two measures *detection rate* and *any-two agreement*. Each measure has different strengths but it may be the case that there are other measures that capture aspects of the evaluator effect that are not fully captured by these two measures alone.

One main implication for HCI practitioners is that for any of the three UEMs analyzed in this paper, it is not a good idea to have only one evaluator. If using only one evaluator practitioners not only miss major usability problems, they may also fail to assign the correct severity levels for each detected problems. The paper also implied that adding another evaluator may be more beneficial than adding more users for some of the UEMs. Another implication for practitioners is that it is better to specific about goals and task analysis for a subset of critical problems than be vague and try to capture problems for all aspects of a system. According to the authors usability evaluations will always involve evaluator judgment to some degree so there is no way to completely eliminate differences between evaluators. Rather than trying to eliminate these differences it is more effective to leverage them and include more evaluator perspectives to get better usability problem coverage. It is also important to clearly define goals and procedure and focus on only the most important aspects of a system if possible.

An implication for users of technology is that many of the systems deployed, even several years after this paper’s publication, can profoundly be affected by the evaluator effect and the vagueness that is often used to guide evaluators when performing UEMs. These systems may suffer from more usability issues than would be present if evaluators had included even one additional evaluator in their analysis. This paper also implies that without further analysis of validity in addition to reliability, users may have to interact with systems that may have produced consistent results when being evaluated but still may not be useful in terms of real world problems.

References Cited

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

Hvannberg, E. T., Law, E. L. C., & Lérusdóttir, M. K. (2007). Heuristic evaluation: Comparing ways of finding and reporting usability problems.*Interacting with computers*, *19*(2), 225-240.

Jacobsen, N. E., Hertzum, M., & John, B. E. (1998). The evaluator effect in usability tests. *CHI 98 Conference Summary on Human Factors in Computing Systems - CHI '98*.

Vatrapu, R., & Pérez-Quiñones, M. A. (2006). Culture and usability evaluation: The effects of culture in structured interviews. *Journal of usability studies*, *1*(4), 156-170.