Applying User Testing Data to UEM Performance Metrics

Jarinee Chattratichart
Department of Computing, Communications Technology
and Mathematics
London Metropolitan University
London, UK
J.Chattratichart@londonmet.ac.uk

Jaqueline Brodie
Department of Information Systems and Computing
Brunel University
Uxbridge, UK
Jacqueline.Brodie@brunel.ac.uk

Abstract

The lack of standard assessment criteria for reliably comparing usability evaluation methods (UEMs) is an important gap in HCI knowledge. Recently, metrics for assessing thoroughness, validity, and effectiveness of UEMs, based on user data, have been proposed to bridge this gap. This paper reports our findings of applying these proposed metrics in a study that compared heuristic evaluation (HE) to HE-Plus (an extended version of HE). Our experiment showed better overlap among the HE-Plus evaluators than the HE evaluators, demonstrating greater reliability of the method. When evaluation data, from testing the usability of the same website, was used in calculating the UEM performance metrics, HE-Plus was found to be a superior method to HE in all assessment criteria with a 17%, 39%, and 67% improvement in the aspects of thoroughness, validity, and effectiveness, respectively. The paper concludes with a discussion concerning the limitations of the effectiveness of the UEM from which the real users' data was obtained.

Categories & Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces — evaluation/methodology

General Terms: Performance, Reliability

Keywords: Heuristic evaluation

INTRODUCTION

Comparative studies have been conducted over the last two decades to provide guidance to usability practitioners on choosing UEMs. To date, however, these studies have not been able to provide usability practitioners with meaningful comparisons of UEMs. First, the 'evaluator effect' is a common problem, i.e. different evaluators evaluating the same product with the same UEM report different sets of problems [5]. Secondly, there has been a lack of scientific rigour in the methodologies employed by some UEM studies. And, thirdly, no common or standardized set of appropriate metrics for comparing UEMs exist to allow a meaningful interpretation of results within and across studies [3].

Recently, attempts have been made to find and define appropriate assessment criteria for comparing UEMs which are based on the 'realness' of data, i.e. actual usability problems experienced by real users [4]. The criteria that have been proposed for this are thoroughness, validity, and effectiveness. Since these criteria are based on 'realness' of data, meaningful comparisons of UEMs can be achieved.

We have proposed a UEM called HE-Plus, which is an extension to heuristic evaluation [1], and have conducted two comparative studies of these UEMs. The first study demonstrated a significant superiority of HE-Plus over heuristic evaluation (HE) in terms of reliability but not in ease of use of the method. HE-Plus procedure was subsequently refined and a second comparative study between HE and the revised version of HE-Plus was carried out. Two user testing experiments of the same product were also conducted at the same time. In this paper data from the user testing have been used to compute the three UEM performance metrics as defined in [4]. The main objective of this paper is to relate the challenges of doing this and to share with other UEM researchers the lessons we have learned from attempting to incorporate 'realness' of data into UEM comparisons.

UEMs EVALUATED

The two UEMs compared in this study were Nielsen's [7] heuristic evaluation (HE) and HE-Plus. Briefly, HE-Plus is an extended version of HE. The difference is that in HE-Plus evaluators are given a 'usability problems profile', which consists of problem areas commonly identified for the type of products or interfaces being evaluated, to be taken into consideration while going through the list of heuristics used [1].

UEM PERFORMANCE METRICS

Three assessment criteria for UEMs defined by Hartson, Andre, and Williges [4] are *Thoroughness*, *Validity*, and *Effectiveness.* Thoroughness measures the proportion of real problems identified by a UEM. Validity measures the proportion of the problems identified by a UEM that are real problems. None of these metrics on their own addresses errors arising from false alarms and misses. Effectiveness accounts for these errors and is defined as the product of Thoroughness and Validity. All the metrics have a value from 0 to 1 and are computed as follows:

$$\textit{Thoroughness} = \frac{\text{Number of real problems identified}}{\text{Number of real problems that exist}}$$

$$\textit{Validity} = \frac{\text{Number of real problems identified}}{\text{Number of problems identified}}$$

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

Effectiveness = Thoroughness x Validity

THE COMPARATIVE STUDY

This section describes the methodology employed in our second comparative study of HE vs HE-Plus.

Hypothesis

We hypothesised that HE-Plus would outperform HE and that HE-Plus would achieve higher ratings from evaluators on the issues of usability and evaluators' confidence in the method.

Method

Participants

Ten evaluators participated in this study. They were MSc students at London Metropolitan University.

Design

The experiment was a between-subjects design. The independent variable was the UEM (2 levels: HE and HE-Plus). The dependent variables were the UEM performance metrics, which were to be calculated from usability problems reported by the evaluators.

Materials and procedure

Two groups of five participants each were randomly assigned to either the HE group or the HE-Plus–group. The website to be evaluated was the Meadow Hall shopping centre (http://www.meadowhall.co.uk/home.cfm). Before the experiment, participants rated their own expertise in doing heuristic evaluation on a scale of 1 (novice) to 5 (expert) in a pre-test questionnaire. Then, they were given a UEM training pack that contained exactly the same information for both groups, except for the information about their respective UEM. Participants read through the information, during which time, they were free to ask any questions.

Both groups were given the same set of Nielsen's [7] heuristics but the HE-Plus group was also given a 'usability problem profile' for websites which was obtained from our previous study [1]. Evaluation time was limited to 90 minutes. Participants recorded their evaluation in a Word document while exploring the website. After submitting their evaluation report, they were asked to complete a 5-point scale post-test questionnaire (1 for lowest and 5 for highest) to rate the website, the evaluation method they used, and their confidence in their own evaluation results.

USER TESTING

Two user testing experiments were conducted. Both experiments compared two shopping centre websites. One of the websites evaluated was Meadow Hall shopping centre (http://www.meadowhall.co.uk/home.cfm). The other was Merry Hill shopping centre website (http://www.merryhill.co.uk/home.html). A brief description of these two experiments is given below.

Method

The two experiments were of a different design. One of the experiments (Experiment 1) was a within-subject design and the other (Experiment 2) was a between-subjects design. In both experiments, the independent variable was the website (2 levels: Meadow Hall and Merry Hill). The dependent variables were task completion time and correctness of the answers given by participants. Qualitative data from interview with users and observations were also collected. There were 14 and 12 participants in Experiment 1 and 2, respectively. Both participants and experimenters were MSc students at London Metropolitan University. Random assignment and counterbalancing were employed where appropriate. All participants were asked to perform the same set of tasks and answer the same questions in both experiments. The same pre-test and post-test questionnaires were used in the two experiments. The latter asked participants to rate various aspects of the websites. A short and unstructured interview was also conducted at the end.

ANALYSIS OF RESULTS

Metrics

Reliability metric

Kessner et al. [6] used the mean number of evaluators finding a problem to compare reliability of usability evaluation results by different teams. Based on this metric, we devised a reliability metric, called *OLP* (the mean number of evaluators finding a problem per unit group expertise), to correct for the variations in the evaluators' expertise in heuristic evaluation. This was because the wide variation of expertise level among our novice evaluators affected the number of problems identified within the limited time imposed upon them. For example, we found that one HE-Plus evaluator identified about a third of the problems reported by the group. Upon further investigation, we discovered that the student was an expert usability engineer who did heuristic evaluations regularly on the job (her report was thus not used in our analysis).

Performance metrics

The UEM performance metrics used were *Thoroughness*, *Validity*, and *Effectiveness*. In order to compute the values for these metrics, data from the evaluators' reports and user testing experiments were prepared as described below.

Data handling

UEM comparison data

Two master lists of 50 and 49 problems were generated from the HE and HE-Plus groups, respectively. The two authors of this paper firstly independently examined the lists to remove non-usability problems and then categorised similar problems based upon problem categories obtained in our first comparative study in which evaluators evaluated a shopping centre website using HE or HE-Plus [1]. For the problems that did not fit into any of the original categories, new categories were added. We then went through our categories together to resolve any disagreements in our results. There were 38 and 30 problem categories identified by the HE group and HE-Plus group, respectively. The number of evaluators reporting the problems in the same category and average group expertise were then obtained for the *OLP* calculation.

User testing data

Time and accuracy data of 18 tasks for the Meadow Hall site from Experiments 1 and 2 were investigated. Both completion time and accuracy performance of each task were carefully considered together and a list of problems was identified from the task performance. Additional usability problems were also obtained from the interview and observational data. This formed a list of problems that real users experienced when using the Meadow Hall website. These problems were then categorised the same way as that for the UEM data, using the same problem category list. There were 17 problem categories identified. All established from those in the list.

Hit, Miss, and False Alarm

Following the definitions given by [2] for a hit, miss, and false alarm, the total numbers of hits, misses, and false alarms were obtained by matching the problems in the two UEM lists with those identified from the user testing experiments. With reference to the formulae for the three UEM performance metrics previously given,

- Real problems identified consist of hits.
- Problems identified consist of hits and false alarms.
- Real problems that exist consist of hits and misses.

Results

Thoroughness, *Validity*, and *Effectiveness* were computed using the formulae given previously. A t-test revealed that HE-Plus evaluator's performance was marginal better than that of the HE evaluators on *Validity* (t = 1.995, df = 7, p = 0.086) and *Effectiveness*, (t = 2.11, df = 7, p = 0.072). In terms of reliability, Kolmogorov-Smirnov test revealed a significant difference of *OLP* between HE and HE-Plus groups, Z =1.703, p < 0.01. Overall group performance on these metrics is tabulated in Table 1 together with the evaluators' subjective ratings on the UEMs they used.

DISCUSSION

Our hypothesis was supported. The result showed that HE-Plus outperformed HE. Furthermore, the HE-Plus evaluators rated the method higher than the HE evaluators.

Reflecting on our experience with the application of real user testing data to compute the UEM performance metrics, we found firstly, that we gained more confidence in the HE-Plus method. In our previous studies, only the reliability of the two UEMs could be compared. Despite the positive outcome of our findings, errors from false alarms and problems missed were not accounted for in those studies. Secondly, having clear definitions for terminologies [2] and assumptions for the formulae for the performance metrics [4] to work with helped us to more confidently prepare the data collected for the analysis. However, the formulae given by the latter authors were easier to use when the terms in the numerator and the denominator were expressed in terms of hits, misses, and false alarms instead. The use of these three vocabularies has made the formulae more meaningful.

Table 1. Overall UEM performance of HE and HE-Plus

Metrics and Ratings	HE	HE-Plus	% Improved
Thoroughness	0.65	0.76	17
Validity	0.28	0.39	39
Effectiveness	0.18	0.30	67
Reliability, *OLP*	0.70	1.02	46
UEM usability	2.8	4.0	43
Evaluators' confidence	3.3	4.0	21

One major concern that we have is the effectiveness of the UEM from which the 'realness' data is obtained. Due to evaluator effect, it is unlikely that an evaluation by any UEM would reveal all problems that exist in a product. Therefore, the values of these metrics are best used in relative terms and should not be taken as absolutely correct.

In an imperfect scenario where 'realness' data are obtained from a user testing that has *Effectiveness* value less than 1, some of the problems predicted by a UEM that should have been hits could be mistaken as false alarms just because they were not found by user testing. We call these 'false negatives'.

Figure 1 illustrates one such scenario. The oval represents the set of problems identified by a user testing, the circle represents all the problems that exist, and the rectangle represents problems predicted by a UEM.

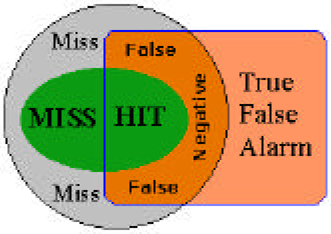

Figure 1. A scenario when user testing effectiveness is low

Due to 'false negatives', the UEM effectiveness calculated (E_{CAL}) could be under-estimated or over-estimated. Based on the definitions and formulae used here, we derived an equation:

$$E_{UEM}/\ E_{CAL}\ =\ E_U\ (1+F/H)^2,$$

where E_{UEM} and E_U are actual UEM effectiveness and the *Effectiveness* of the user testing, respectively. F is the number of 'false negatives' and H is the number of the problems identified by *both* the user testing and the UEM. From this, the relationship between E_U and %F (defined as 100 x F/(F+H)) was then plotted for $E_{UEM}/\ E_{CAL}$ ranging from 1 to 5 in Figure 2. At $E_{UEM}/\ E_{CAL} = 1$, actual UEM effectiveness is equal to the UEM effectiveness calculated.

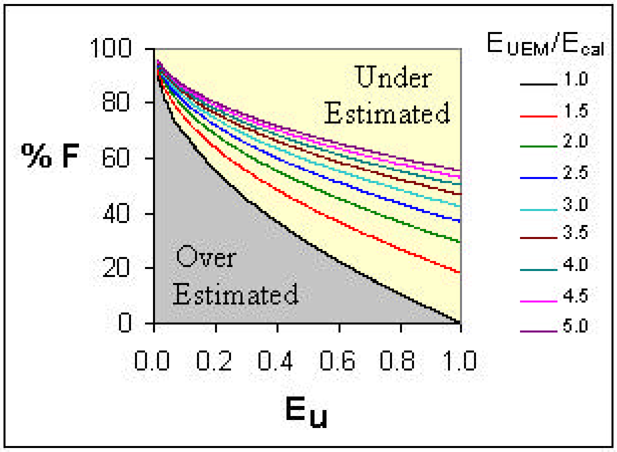

Figure 2. Effect of user testing effectiveness on the value of UEM effectiveness

To illustrate, the light coloured area in Figure 2 represents under-estimated cases and the darker area, over-estimated cases. For example, if $E_U = 0.8$ and %F = 20 (if we knew), the actual E_{UEM} would have been under-estimated because $E_{UEM}/\ E_{CAL}\ > 1$.

CONCLUSION

We have described a comparative study of two heuristic inspection methods, heuristic evaluation (HE) and HE-Plus, that are compared using UEM performance criteria based on real users' data obtained from two user testing experiments. The performance metrics show that HE-Plus is a more effective method than HE. While 'realness' of data used in the computation of the metrics addressed the issue of false alarms and problem missed, 'false negatives', defined as hits that are mistaken as false alarms, becomes another issue to be addressed by future UEM research.

ACKNOWLEDGEMENTS

We thank the MSc students in IT Usability (ITU), IT Consultancy (ITC), Multimedia Systems (MMS), and Information Systems Development (ISD) at London Metropolitan University for their participation in the comparative study and the user testing experiments.

REFERENCES

1. Chattratichart, J. and Brodie, J. Extending the heuristic evaluation method through contextualisation. *Proc. HFES2002*, HFES (2002), 641-645.
2. Cockton, G. and Woolrych. Sales must end: Should discount methods be cleared off HCI's shelves? *Interactions*, September and October issue (2002), 14-18.
3. Gray, W. D. and Salzman, M. C. Damaged Merchandise? A review of experiments that compare usability evaluation methods. Human-Computer Interaction, 13 (1998), 203-261.
4. Hartson, H. R., Andre, T. S. and Williges, R. C. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, 15, 1 (2003), 145-181.
5. Hertzum, M. and Jacobsen, N. E. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 15, 1 (2003), 183-204.
6. Kessner, M., Wood, J., Dillon, R. F., and West, R. L. On the reliability of usability testing. *Proc. CHI2001 Extended Abstracts*, Washington, DC: ACM Press (2001), 97-98.
7. Nielsen, J. Heuristic evaluation. J. Nielsen and R. L. Mack. (Eds.), *Usability Inspection Methods*. New York: John Wiley & Sons (1994), 25-62.