

Towards Increasing Reliability of Heuristic Evaluation

Introduction

To date comparative studies have not convinced the HCI community about the reliability of results from usability evaluations. User testing of the same products by different usability teams found different problems (Molich et al., 1999). Discount methods also have the same problem. With heuristic evaluation method, for example, overlaps between evaluators' results are often as low as 10% (Molich & Robin, 2003). However, the strength of the heuristic evaluation method is in its cost-effectiveness. In a comparative study (Cuomo & Bowen, 1994) of heuristic evaluation, cognitive walkthrough, and interface guidelines, heuristic evaluation found the highest percentage of true problems per hour of inspection (9.3% as compared to 6.8% by cognitive walkthrough and 2.3% by guidelines). Furthermore, there is a pressing need for a discount evaluation method like heuristic evaluation. As a 'quick and easy to perform' method, heuristic evaluation was rated as the most commonly used evaluation method by 134 HCI practitioners at CHI 98, CHI 99, and UPA 99 Conferences (Rosenbaum et al., 2000). However, to best benefit from the method, we need to find ways to improve the reliability of its results.

Recently, Chattrachart (2003) identified a 'profile', which consists of usability problem areas, associated with the complex application that she evaluated employing a naturalistic inquiry approach. In this research, she pioneered the concept of 'usability problems profile', which refers to the usability problem areas that are commonly identified for products or interfaces of the same type. This concept was then used to extend and improve the reliability of HE results and ease of use of the method. She and a colleague (Chattrachart & Brodie, 2002) introduced the HE-Plus method, in which evaluators were given a list of problem areas that constitute a '*usability problems profile*' to be taken into consideration when applying the list of heuristics used.

Two experiments that compared heuristic evaluation (HE) and HE-Plus method have been carried out with positive outcomes. The main objective in this paper is to describe the first experiment carried out to assess a possible extension to the heuristic evaluation method in order to help address some of its limitations as highlighted above.

Empirical Study

To investigate the benefit of using a 'usability problems profile', we compared the performance of two groups of novices evaluating a web site. The HE group employed Nielsen's heuristic evaluation method (Nielsen, 1994) and the HE-Plus group employed the HE-Plus method.

Hypothesis

We hypothesized that more overlapping results would be found among HE-Plus evaluators than HE evaluators in the experiment.

Method

Participants. Ten research students at Brunel University participated in this experiment. All were experienced Internet users.

Design. The participants were equally divided into two groups and randomly assigned to either the HE group or the HE-Plus-group.

Materials. The participants were given a training pack a few days before the URL was given to them. The pack described the procedure for the evaluation method assigned to the owner of the pack. All packs were identical except for the information concerning the evaluation method to be used. For HE method, a list of Nielsen's (1994) ten heuristics was given. For the HE-Plus method, however, a list of 'usability problems profile' was given in addition to the ten heuristics. The problem areas which constitute the profile for this study was taken from Lindgaard (1994).

Procedure. Participants were asked to use the method assigned to them and described in their training pack to evaluate the usability of an online shopping web site (<http://www.lakesideonline.uk.com>). They were instructed to carry out the evaluation individually at their own pace. They were advised to spend between one and three hours exploring the site however they wished. Both groups were to examine the web site based on, but not limited to, the set of heuristics given. Each participant in the HE-Plus group was to also look for problems in the areas given in the 'usability problems profile'. All participants were asked to submit list of usability problems they found at the end of the evaluation. After the submission, a one-page questionnaire was given to participants. The questionnaire asked participants to rate the web site, the evaluation method they used, and the confidence in their own evaluation results.

Results

Reliability Metrics. Kessner et al. (2001) in their recent research compared the reliability of the usability testing results of six usability teams with those discussed in Molich et al.'s (1999) study. Following the metrics used by Kessner et al. (2001), the mean number of evaluators finding a problem (based on unique problem categories found) was used to compare the reliability between the two methods in our study. A higher mean indicates more overlap, hence, more consistent results. In addition, the percentage of problems found by 1, 2, 3, 4, 5 evaluators was used as an indicator of the overlap in the evaluators' findings.

Problem grouping. A list of problems was obtained from the participants' reports. There were 145 problems reported in total. One hundred and thirty-two usability problems were identified and similar usability problems were categorized independently by the two authors. A final set of 36 unique usability problem categories (despite an 18% initial disagreement) was agreed upon.

Findings. Table 1 presents a summary of statistical findings of the two groups. The HE group spent an average of three hours on the evaluation while the HE-Plus group spent on average only two hours. The former found 49 usability problems while the latter, 83 problems. Of the 36 problem categories, 22 were reported by both groups, five by HE group alone, and nine by HE-Plus group alone.

None of the 36 problem categories was found by all five evaluators in either group. Problems found by 1, 2, 3, and 4 evaluators were 67%, 15%, 11%, and 7 % for the HE group respectively. These figures were 26%, 29%, 26%, and 19% for the HE-Plus group. The mean numbers of evaluators finding a problem (M) were 1.19 ($SD = 1.09$) and 2.06 ($SD = 1.31$) for HE and HE-Plus groups, respectively. Mann-Whitney test showed a significant difference between the two means, $z = 2.91$; $p < 0.01$.

Questionnaire results. Average ratings, on a scale of 1 to 5 (see Table 1) revealed that the participants found the original method easier to use and learn than the new method. They were also more confident in their own evaluations than the HE-Plus group.

	HE	HE-Plus
GENERAL STATISTICS:		
Average time taken (hr)	3	2
Number of problems found	49	83
Number of problem categories	27	31
OVERLAP:		
Mean number of evaluators finding a problem	1.19	2.06
Percentage of problems found by:		
2 or more evaluators	33	74
3 or more evaluators	18	45
4 or more evaluators	7	19
SUBJECTIVE RATINGS:		
Web site experience	2.8	2.8
Usability of the method used	4.6	3.1
Confidence in own evaluation	4.8	4.1

Table 1. Summary of statistical data of both groups.

Discussion

Our hypothesis was supported. The HE-Plus group spent on average one hour less than the HE group on the evaluation and yielded more reliable results than the HE group. HE-Plus yielded significantly better overlapping results than HE as shown in Table 1. However, subjective ratings indicated that participants found the original method easier and therefore had higher confidence in their evaluation results. This might have been due to the additional information in the instruction pack regarding the problem areas that had to be considered, making the recommended procedure more complex for the HE-Plus group than for the HE group. In our second study (not reported here), the procedure was simplified and HE-Plus was reported easier than HE (Chattrachart & Brodie, 2003).

Conclusion and Future Research

We have proposed and tested HE-Plus for improving the reliability of the heuristic evaluation method. It appeared that the *'usability problems profile'* given to the HE-Plus group helped the evaluators focus their evaluations and hence resulted in more reliable results.

To further test our method we suggest that more comparative studies similar to this study be carried out with both novice and expert usability professionals evaluating different types of applications. Another area for future research that could prove beneficial to practitioners

is compiling a 'profile bank', that is, a database of problem areas for different types of applications, so that evaluators can, in the future, choose an appropriate profile for what needs to be evaluated.

References

- Chattratchart, J. (2003). *Usability Issues and Design Principles for Visual Programming Languages*. Unpublished Ph.D. Thesis, Brunel University, UK.
- Chattratchart, J. & Brodie, J. (2003). HE-Plus: Toward usage-centered expert review for website design. In L. L. Constantine (Ed.), *Proceedings of forUSE 2003, Second International Conference on Usage-Centered Design* (pp. 155-169). Massachusetts: Ampersand Press.
- Chattratchart, J. & Brodie, J. (2002). Extending the heuristic evaluation method through contextualisation. In *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society* (pp.641-645). HFES.
- Cuomo, D. L. & Bowen, C. D. (1994). Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers*, 6(1), 86-108.
- Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the reliability of usability testing. *Proceedings of CHI'2001 Extended Abstracts*, 97-98. Washington, DC: ACM Press.
- Lindgaard, G. (1994). *Usability Testing and System Evaluation: A Guide for Designing Useful Computing Systems*. Chapman & Hall.
- Molich, R. & Robin, J. (2003). Comparative Expert Reviews. *Proceedings of CHI'2003 Extended Abstracts*, 1060 – 1061. Washington, DC: ACM Press.
- Molich, R., Thomsen, A. D., Karyukina, B., Schmid, L., Ede, M., van Oel, W., & Arcuri, M. (1999). Comparative evaluation of usability tests. *Proceedings of CHI'99 Extended Abstracts*, 83-84, Washington, DC: ACM Press.
- Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack. (Eds.), *Usability Inspection Methods*, pp. 25-62. New York: John Wiley & Sons.
- Rosenbaum, S., Rohn, J. A., & Humburg, J. (2000). A toolkit for strategic usability: Results from workshops, panels, and surveys. *Proceedings of CHI'2000*, 337-344. Washington, DC: ACM Press.