

Learning from experienced users in real life

Liu, Y and Osvalder, L-O

Division Design, Dept. Product & Production Development
Chalmers University of Technology, SE-412 96, Göteborg, Sweden
e-mail: yuanhua@hfe.chalmers.se

Test subjects' performance in usability evaluations show end users' demands, expectations and limitations. In this study, heuristic evaluation, thinking-aloud method and co-discovery method were employed to evaluate the experienced users' performance on the interface of a Whirlpool microwave oven at a working-place. The purpose was to investigate experienced users' performance on casual tasks. The results showed that all the test subjects failed in completing several task scenarios, which promotes designers to reconsider the users' expectations on interface design and the practical efficiency of multi-functionality issue. In addition, some interesting findings and differences among the three evaluation methods were also implied.

Keywords: experienced user, interface design, heuristic evaluation, thinking-aloud method, co-discovery method.

1. Introduction

Nowadays, usability evaluations play an important role in the product design process. The aim is to improve the quality of the product design in the design process cycle, which refers to the determination of the effectiveness of an interface either in use or in prototype, and providing suggestions on improvements for the next generation of the product design.

Although many comparisons between different usability evaluation methods have been conducted in recent years, the relevant knowledge is still new and incomplete as a research topic from a historical viewpoint (Hartson et al., 2003). From many literature studies on comparisons between different usability evaluation methods, Hartson et al. (2003) found two aspects related to the confusion and ambiguities, i.e. adopting different evaluation criteria, and neglecting the characteristics and limitations of different methods. Furthermore, confounding the situation in comparative studies is indicated as a miscomprehension of the limitations of usability evaluation methods and to what conditions those limitations apply (Hartson et al., 2003). Gray and Salzman (1998) pointed out that the lack of understanding the evaluation on usability evaluation methods makes the state of usability evaluation method issues most disturbing to many people. Therefore, Hartson et al. (2003) appealed for more carefully designed comparison studies – both to contribute to the science of usability and to provide practitioners with more reliable information about the relative performance of various usability evaluation methods used for various purposes.

Besides the methodology aspect, users' influence in the evaluation can never be neglected. As test subjects in empirical evaluations, how the users' expertise or previous experience influence their final decision making in performance is always an interesting and practical topic that is worthy of further investigations.

2. Objectives

The objective of this pilot study was to obtain proposals for future redesign issues from a usability evaluation, not only from the users' experience aspect, but also from an evaluation methodology aspect. Three evaluation methods were used and analyzed and compared: heuristic evaluation, the thinking-aloud method and the co-discovery method.

3. Methods

In this study, the interface of a Whirlpool microwave oven that is located in the coffee room at an academic department was selected for the evaluation. Seven casual task scenarios, which reflect typical main functions in practical life, were selected for the evaluation.

The problem severity was judged and analyzed based on four degrees: (A) No problem, which means the state that nothing wrong happens in the task completion period; (B) Minor errors, which indicate the errors that can be noticed and corrected either instantly or quickly by the test subject; (C) Major errors, which indicate the errors on which the test subject has a wrong mental model of task completion, but they can spot and rectified with a greater cost in terms of time and annoyance; and (D) Failure, which indicates the errors that can prevent the test subject from completing the task or lead to completely catastrophes.

3.1 Heuristic evaluation

Two usability experts were used as inspectors in the heuristic evaluation (Nielsen, 1994). The heuristic list employed in the evaluation was a combination of usability principles proposed by Nielsen (1994) and Jordan (1998). These heuristics were: visibility of system status, match between system and the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, help users recognize/diagnose/recover from errors, physical ergonomics/consideration of user resources, prioritization of functionality and information, and explicitness.

Thinking-aloud method

The thinking-aloud method was conducted according to the procedure proposed by Nielsen (1993). Eight staff members from the department participated as test subjects in the test. A pre-session interview regarding the test subject's general experience on microwave oven usage was made before the evaluation started.

During the evaluation, task completion time and error rates were collected. Each task scenario was given three minutes for the test subjects to complete.

After each session, interview was also made regarding test subjects' comments on the tasks and the interface. Finally, each test subject gave ranks regarding satisfaction on the 'System Usability Scale' questionnaire (Brooke, 1996).

Co-discovery method

The co-discovery method is also called 'constructive interaction' method, which is actually a usability testing method with participants working in pairs in the task performance, while being observed by the usability expert during evaluation (Dumas and Redish, 1993).

Eight other staff members from the department participated as test subjects in this test. A pre-session interview was also made before the evaluation started. Task completion time and error rates were collected.

4. Results

4.1 Overall judgment on problem severity

The results showed differences between the thinking-aloud method and the co-discovery discovery method. The problem severity was lower on three out of seven tasks with the co-discovery method. When comparing the thinking-aloud method and heuristic evaluation, the results on problem severity were agreed on two out of seven tasks; no clear tendency could be found on the other five tasks. In the same way, although the results were in agreement on three out of seven tasks between the co-discovery method and the heuristic evaluation, no clear conclusion could be drawn on the other four tasks.

4.2 Analysis on heuristics

The results showed that all three usability methods detected problems regarding visibility, match between the system and the real world, consistency, and explicitness. Additionally, the heuristic evaluation method detected usability problems related to error prevention and helping recognize/diagnose/discover from errors.

4.3 Thinking-aloud vs. Co-discovery methods

Statistical analysis was made on task completion time between the thinking-aloud method and the co-discovery method. The t-test results implied that there was no difference on task completion time between the two evaluation methods for all task scenarios. Based on the Mann-Whitney analysis, the statistical results also implied that there was no significant difference on satisfaction about the interface design between the two evaluation methods.

4.4 Subjective assessment on interface design

In terms of comments on the interface design, 44% of the test subjects regarded the button layout and the icons on the buttons as good features, while 13% of the test subjects regarded the size of the large adjusting knob was a good feature. 63%

of the test subjects thought the explicitness of some functionality was very bad, for instance, defrost and fan. 31% of the test subjects thought consistency in button activation and priority of functionality were bad. 19% of the test subjects complained about that certain functionality display did not match between the system and the real world, for instance, the defrost level was only displayed in numbers rather than in corresponding resemblance icons.

5. Discussion

5.1 *Findings from the study*

Based on the analysis on problem severity between heuristic evaluation and the thinking-aloud method, as well as between the heuristic evaluation and the co-discovery method, it was shown that the predictions made in the heuristic evaluation did not generally agree with the practical results made by test subjects when analyzing with the thinking-aloud and co-discovery methods.

From the analysis, the problem severity was found lower on three out of seven tasks by the co-discovery method. This might probably happen due to the strength of the co-discovery method, since the test subjects (in pairs) had better performance on task completion with the help of cooperation and discussion. This could be seen when analyzing task completion time. However, from the statistical analysis, the difference in task completion time and subjective satisfaction was not significant between the thinking-aloud and co-discovery methods.

In the CHI'92 workshop on usability inspection methods, a question was raised regarding whether heuristic evaluation method described problems end users would have in reality, or whether the evaluation results in some way were related to ultimate end users satisfaction (Dutt et al., 1994). In the present study, two 'false' usability problems related to error prevention and helping recognize /diagnose/recover from errors were found only by heuristic evaluation method, i.e. not by thinking-aloud method or co-discovery method. This implied the agreement with the comments by Dutt et al. (1994), which indicated the weak point of heuristic evaluation method, that is, heuristic evaluation might identify 'false' problems irrelevant to usability thus leading the designer in the wrong direction.

5.2 *Evaluator effect in the comparison*

Evaluator effect has been found as a limitation in many studies. Desurvire et al. (1991, 1992a, 1992b) used three levels of evaluator expertise in their study. Their result showed that only the usability experts found nearly as many problems as usability testing. Just as Doubleday et al. (1997) pointed out, heuristic evaluation relies heavily on the expertise of the evaluator, both in the area of usability and in domain knowledge. In Hertzum and Jacobsen's (2001) literature review on eleven usability studies, it was found that the evaluator effect exists for both novice and experienced evaluators, for both cosmetic and severe problems, for both problem detection and severity assessment, and for evaluations of both simple and

complex systems. The reason for these effects was assumed and explained by that usability evaluation is a cognitive activity, which requires that the evaluators make judgments. In this study, although two usability experts conducted the heuristic evaluation together, disagreements on the problem severity and relevant heuristics of usability problem were still attained. However, an interesting finding was that no clear or obvious correlation could be found between the two experts' professional levels and their prediction powers.

5.3 User's experience vs. product design

In this study, the test subjects were actually experienced end users of microwave ovens for at least 10 years, and they had at least 6 months experience with the tested microwave oven. But still, all the test subjects failed in several task scenarios. Although they had used various types of microwave ovens in the past, their previous experienced did not help them much in their performance on the present microwave oven. Both the interview and questionnaire analysis revealed possible reasons for this – users' normally only focus on the most basic functionalities, rather than all the functionalities provided by the interface design. Therefore, a question about multi-functionality on product design can be raised: How much does the multi-functionality on product design benefit the end users? In other words, does multi-functionality really benefit rather than confuse the users in reality?

From the test results in this study, it was obvious that the experienced users were skilful for most basic and frequently-used functionalities, but acted probably in the same way as novice users for other functionalities. Therefore, a recommendation was that designers need to consider carefully when choosing test subjects and select tasks for the evaluations of user interfaces.

6. Conclusion

Based on the results and analysis of the present study, the following conclusions can be drawn:

- (1) Although heuristic evaluation predicted correctly on problem severity in several task scenarios, it was still hard to conclude that the method had strength in predicting problem severity;
- (2) The strength of the co-discovery method was a lower problem severity in task performance than the thinking-aloud method showed;
- (3) No significant difference was found between the thinking-aloud and the co-discovery methods regarding task completion time and subjective satisfaction;
- (4) The evaluator effect should be considered carefully in heuristic evaluation;
- (5) Experienced users' expertise was helpful in most basic and frequently used functionalities of the interface;
- (6) The evaluated microwave interface should be improved regarding explicitness of functionality and display, consistency of task completion, and the certain display of icons.

References:

Brooke, J., 1996. SUS – A quick and dirty usability scale, in Jordan, P.W. *et al.* (Eds), Usability Evaluation in Industry, London: Taylor & Francis. pp. 189-194,

Desurvire, H., Lawrence, D., and Atwood, M.E., 1991. Empiricism versus judgment: Comparing user interface evaluation methods on a new telephone-based interface. SIGCHI Bulletin, 23, 4, pp. 58-59.

Desurvire, H., Kondziela, J., and Atwood, M., 1992a. (short paper version). What is gained and lost when using evaluation methods other than empirical testing. A short talk presented at CHI'92 (Monterey, California, May 3-7, 1992). ACM, collection of abstracts, pp. 125-126.

Desurvire, H., Kondziela, J., and Atwood, M., 1992b. (full paper version). What is gained and lost when using evaluation methods other than empirical testing. In: Proc. of HCI'92, Cambridge University Press, edited by Monk, A., Diaper, D., and Harrison, M.D., (University of York, U.K., September 15-18, 1992).

Doubleday, A., Ryan, A., Springett, M., Sutcliffe, A., 1997. A comparison of usability techniques for evaluating design. In: Proc. Of Designing Interactive Systems (DIS'97): Processes, Practices, Methods, and Techniques, Amsterdam, the Netherlands, ACM Press, pp. 101-110.

Dumas and Redish, 1993. A Practical Guide to Usability Testing, Norwood, NJ: Ablex Publishing. pp. 31.

Dutt, A., Johnson, H., and Johnson, P., 1994. Evaluating evaluation methods. In: Proceedings of the conference on people and computers IX (Glasgow, August 23-26). pp. 109-121.

Gray, W.D., and Salzman, M.C., 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer interaction*, 13, pp. 203-262.

Hartson, H.R., Andre, T.S., and Williges, R.C., 2003. Criteria for evaluating usability evaluation methods. *Human-Computer interaction*, 15(1), pp. 145-181.

Hertzum, M., Jacobsen, 2001. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4), pp. 421-443.

Jordan, P.W., 1998. An Introduction to Usability. Taylor & Francis Ltd. U.K.

Nielsen, J. and Mack, R.L., 1994. Usability Inspection Methods. John Wiley & Sons, Inc. U.S.A.