Proceedings of Fourth Danish HCI Research Symposium, November 16 2004, Aalborg University.

Studying usability evaluation to improve its practical utility

Erik Frøkjær
Department of Computing
Universitetsparken 1
DK-2100 Copenhagen
+45 35321456
erikf@diku.dk

Kasper Hornbæk
Department of Computing
Universitetsparken 1
DK-2100 Copenhagen
+45 35321425
kash@diku.dk

ABSTRACT
Current research on usability evaluation has several limitations, including focusing on the evaluation outcome and on counting usability problems; a realistic understanding of how usability evaluation is used in practice has been largely ignored. We describe some of our recent work on addressing these limitations, including a diary study of evaluation processes, studying developers' assessments of usability problems, and generating redesign suggestions instead of problems. In addition, we speculate on future research that aims to address the limitations.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Evaluation/Methodology; D.2.2 [Software Engineering]: Design Tools and Techniques—User Interfaces.

General Terms
Measurement, Design, Experimentation, Human Factors.

Keywords
Usability evaluation, redesign, think aloud, metaphors of human thinking, empirical study, usability inspection, diaries.

1. INTRODUCTION
Research in usability evaluation is in a peculiar situation. On the one hand, a substantial number of studies have compared evaluation techniques and developed new ones, for reviews see [2-4]. On the other hand, several authors have pointed to severe limitations of that work. Gray & Salzman [3], for example showed how an often-cited selection of comparisons of evaluation methods suffers from low validity. Problems were identified not only with the statistical tests and the conclusions passed on to practitioners and researchers, but also with the measures used in the comparison of methods. A study of usability evaluation in industry also found that different teams of evaluators identified different usability problems [12]. Recently, Dennis Wixon [14] made the case that comparisons of UEMs do not appreciate that the real goal of such methods is to impact design, not to generate problems. Thus, comparisons may miss to assess properly the practical utility of UEMs.

Recently, we have begun to address these limitations by moving beyond some of the common ways of doing usability research. These include (a) disregarding the evaluation process and focusing only on the outcome of the evaluation, typically on sets of problems; (b) ignore the practical taking up and use of the evaluation's results in realistic system development contexts, and (b) counting usability problems, rather than investigating other outcomes of usability evaluation.

Below we describe studies that each moves beyond one of these common practices. In the final section we speculate on some possible research in which to further addressing the limitations of current usability research.

2. A DIARY STUDY OF EVALUATION PROCESSES
One way to study the evaluation process in more details is to have participants keep a diary. Diaries have previously been used to study the use of evaluation techniques [10,11]. In a recent paper [9] we described a study comparing two psychology-based inspection techniques, cognitive walkthrough and metaphors of human thinking (MOT). In comparison to the existing use of diaries, that study used diaries to compare evaluation techniques and combined diaries with quantitative measures of evaluation performance. In comparison to most of the existing literature in usability evaluation, the study not only focuses on the sets of problems that result from the evaluation, but also on the evaluation process.

Twenty participants evaluated web sites for e-commerce while keeping diaries of insights and problems experienced with the techniques. We will here only discuss the insight into the evaluation process that the analysis of the diaries gave. Especially two findings are relevant to mention.

First, the diaries show that usability problems are found in a variety of ways, not just by using the techniques as prescribed. At least ten participants identify problems already before reading the description of the inspection technique, or while initially orienting themselves on and gaining an overview of the web site. One participant writes, during her first visit on the web site before starting the evaluation procedure:

> Identification of immediate problems and some ideas for tasks. Especially the questionnaire [on the web site] is a disaster. The menu in the left side sometimes disappears. No systematic information on whether a word or a label is clickable…

That participant ends up reporting on her problem list three problems regarding the questionnaire. Even after finishing the evaluation procedure, some participants continue to identify problems.

Thus, participants seem to identify problems in many ways, not only through the techniques, reflecting large differences in individual working styles.

Second, during the course of the evaluation participants change their opinion on what they consider a usability problem, e.g. some participants change their opinion about problems when redesigning. One participant writes that

> [I] have come to the conclusion that the buying procedure is really not so complicated that it will give errors for the user.

The same participant had on his problem list noted as a serious problem the cumbersome buying procedure. Conversely, at least five participants identify problems when redesigning, problems they had not previously been aware of, for example:

> Looking at a screen dump makes me aware of new usability problems. What am I to do with problems I have just discovered?

These observations, and other from the paper [9], suggest that the process of usability evaluation are complex, somewhat disordered, and shaped to a high degree by participants' personal working habits. These findings appear to challenge common assumptions of the evaluation process as an orderly progression of steps that reflect the technique being used.

3. DEVELOPERS' ASSESSMENTS OF USABILITY PROBLEMS

In a couple of experiments, we have studied how developers assess usability problems. The main argument underlying these experiments is that developers' assessments heavily influence if a problem is addressed. Developers have a vested interest in minimizing redesign in order to meet time and cost-constraints and thus may be inherently biased in their assessment of usability problems. In practice, however, these are the circumstances that determine which and how problems are addressed.

In one study [8], MOT was compared to heuristic evaluation (HE). An experiment was conducted in which 87 novices evaluated a large web application. Of particular interest here is that the key developer of the web application assess the problems uncovered by MOT as more severe on users and also appeared more complex to repair than the problems uncovered by HE. The key point here is that the developer's assessment of usability problems helped identify differences between techniques; such differences could be relevant when selecting which technique to use.

In another study [7] we investigated how developers of a large web application assess output from usability evaluation. Problems and redesign proposals were generated by 43 evaluators using an inspection technique and think aloud testing. Of particular interest here is that developers' assessments of problems and our subsequent interviews with them provided insights in some of the reasons for taking up or ignoring a problem. For example, developers expressed that those problems which could be fixed easily and quickly were of particular utility. One developer explained:

> Typically if something can be easily and quickly fixed … that is a suggestion which requires four months of development is not as useful as some small suggestion, which corrects a small problem in 10 minutes, then I can correct it immediately

During all interviews, we asked developers if they could recall usability problems and redesign proposals. Usability problems were mostly remembered by developers as classes of problems, the particular instances was forgotten. One developer said that 'yes, there are several of them [usability problems] that I can still remember' and then—surprisingly—went on to expand on how specific redesign proposals on exploring similarities to standard search engines could be incorporated in the design. In contrast, all developers were able to describe in some detail redesign proposals which they had found interesting.

Another interesting finding was that developers find the problems identified to be mainly confirmations of issues they already know. In a comparative usability evaluation, Molich et al. [12] similarly found that only 4% of the problems identified were new to the usability team responsible for the system evaluated. One immediate reaction could be that this is not much. Yet, maybe we should be careful in concluding that developers get few new insights from usability evaluations. The developers in our study actually used the usability problems, and their thinking about the application seemed to have been influenced. Further, developers who for years have worked intensively with the application and its use context will not easily take up results of usability evaluations. On the contrary, changing their understanding is a process requiring time, during which new insights does not appear as something distinct and immediately clear. Rather, developers will experience nagging doubts, small changes in thinking, and challenges to their understanding. Studying how this develops over time would probably give a more valid picture of the impact of usability evaluations.

4. REDESIGN SUGGESTIONS AS SUPPLEMENTS TO PROBLEMS

Usability problems predicted by evaluation techniques are useful input to systems development; it is uncertain whether redesign proposals aimed at alleviating those problems are likewise useful. We have recently investigated this by having developers of a large web application assess usability problems and redesign proposals as input to their systems development [7]—the study also mentioned in Section 3.

Developers assessed redesign proposals to have higher utility in their work than usability problems. In interviews they explained how (a) redesign proposals help understand usability problems, i.e. redesigns contribute to characterizing and making more concrete the problems found, and illustrate why problems are important; and (b) redesign proposals are useful for inspiration and for seeking alternative solutions for problems that the development team has been struggling with. Point (b) is exemplified in the following quote from one of the developers:

> in some situations you may do things one way or the other, and then you can just choose, i.e. whether some list should be alphabetical or just split up…in other situations, like the three level hierarchical selection of job titles, no matter what we do we get into some complicated mess…so if one can find some way of making it more intuitive and usable than other ways, then we accept it eagerly, [because] we haven't quite figured out how to do it ourselves

The usability problems supported prioritizing ongoing development of the application and taking design decisions. One developer said that

> usability problems … what one cares about is the extent of them, how many is saying that some thing is a problem and how many is saying that some other thing is a problem, that help me prioritize what I should focus on

These comments do not mean, however, that developers did not appreciate usability problems, especially when they are well argued, clearly described, documented, and easy to fix. On the contrary, all developers wanted both problems and redesign proposals to form part of the input from usability evaluation to systems development.

5. FUTURE WORK

We are continuing to experiment with the problems and ideas introduced above. In addition, we wish to share a few further ideas for going beyond some of the limitations noted above.

The matching of usability problems underlies most usability research. Most usability problems are brief, often quite difficult to understand, and certainly incomplete in expressing the evaluators' thoughts, a problem of understandability more generally discussed in e.g. [13]. Thus, matching of just such problems descriptions form an insecure foundation for usability research. One example of this is the study by Molich et al. [12], where problems founds by different professional teams are matched. Another example is studies of the so-called evaluator effect [5], i.e. the observation that evaluators typically find different usability problems. In both cases, it could be worthwhile to explore if this matching really is sound. This could be done, for example, by (a) involving evaluators more in the matching process. Hertzum et al. [6] finds an interesting difference between 'objective matching' of usability problems and the opinions of the usability specialist who had produced the problems; (b) include different representations in the matching, for example both usability problems and suggested redesigns, and (c) study how this matching goes on in practice, to see if what we think are similar or different problems function in the same way for development teams.

When one looks to the literature on usability evaluation in industry, the results are surprisingly meager. More studies of industrial systems development could help us understand (a) what output is useful from evaluation techniques, and (b) at which stages different evaluation techniques give the best results.

Finally, in-depth studies of evaluation processes seem to give interesting data on evaluation performance. Above we gave one example on diary studies. Another example of an in-depth approach that gave interesting data is Boren and Ramey's [1] study of think aloud. They showed how practical think aloud studies are often far from the original content of the think aloud methodology.

REFERENCES

1. Boren, M. T. & Ramey, J. Thinking Aloud: Reconciling Theory and Practice, *IEEE Transactions on Professional Communication*, *43*, 3 (2000), 261-277.
2. Cockton G., Lavery, D., & Woolrych, A., Inspection-Based Evaluations, in Jacko, J. A. & Sears, A. *The Human-Computer Interaction Handbook,* Lawrence Erlbaum Associates, 2002, 1118-1138.
3. Gray, W. D. & Salzman, M. C. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods, *Human-Computer Interaction*, *13*, 3 (1998), 203-261.
4. Hartson, H. R., Andre, T. S., & Williges, R. C. Criteria for Evaluating Usability Evaluation Methods, *International Journal of Human-Computer Interaction*, *13*, 4 (2001), 373-410.
5. Hertzum, M. & Jacobsen, N. E. The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods, *International Journal of Human-Computer Interaction*, *13* (2001), 421-443.
6. Hertzum, M., Jacobsen, N. E., & Molich, R. Usability Inspections by Groups of Specialist: Perceived Agreement in Spite of Disparate Observations, *Proc. CHI2002*, (2002), 662-663.
7. Hornbæk, K. & Frøkjær, E. Comparing Usability Problems and Redesign Proposals as Input to Practical Systems Development, under review (2004).
8. Hornbæk, K. & Frøkjær, E. Usability Inspection by Metaphors of Human Thinking Compared to Heuristic Evaluation, *International Journal of Human-Computer Interaction*, *17*, 3 (2004), 357-374.
9. Hornbæk, K. & Frøkjær, E. Two Psychology-Based Usability Inspection Techniques Studied in a Diary Experiment , *Proc. Nordichi 2004*, ACM Press (2004)
10. Jacobsen, N. E. & John, B. E. Two Case Studies in Using Cognitive Walkthroughs for Interface Evaluation, *CMU-CS-00-132* (2000).
11. John, B. E. & Packer, H. Learning and Using the Cognitive Walkthrough Method: a Case Study Approach, *Proc. CHI'95*, ACM Press (1995), 429-436.
12. Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. Comparative Usability Evaluation, *Behaviour and Information Technology*, *23*, 1 (2004), 65-74.
13. Naur, P. *Knowing and the Mystique of Logic and Rules*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1995.
14. Wixon, D. Evaluating Usability Methods: Why the Current Literature Fails the Practitioner, *interactions*, *10*, 4 (2003), 29-34.