In *Extended Abstracts of the ACM CHI 2002 Conference* (Minneapolis, MN, April 20-25, 2002), pp 662-663. ACM Press, New York.

Preprint version

Usability Inspections by Groups of Specialists: Perceived Agreement in Spite of Disparate Observations

Morten Hertzum[1], Niels Ebbe Jacobsen[2], and Rolf Molich[3]

[1] Centre for Human-Machine Interaction, Risø National Laboratory, DK-4000 Roskilde, Denmark
[2] Nokia Mobile Phones, Frederikskaj, DK-1790 Copenhagen V, Denmark
[3] DialogDesign, Skovkrogen 3, DK-3660 Stenløse, Denmark
morten.hertzum@risoe.dk, niels.ebbe.jacobsen@nokia.com, molich@dialogdesign.dk

ABSTRACT

Evaluators who examine the same system using the same usability evaluation method tend to report substantially different sets of problems. This so-called evaluator effect means that different evaluations point to considerably different revisions of the evaluated system. The first step in coping with the evaluator effect is to acknowledge its existence. In this study 11 usability specialists individually inspected a website and then met in four groups to combine their findings into group outputs. Although the overlap in reported problems between any two evaluators averaged only 9%, the 11 evaluators felt that they were largely in agreement. The evaluators perceived their disparate observations as multiple sources of evidence in support of the same issues, not as disagreements. Thus, the group work increased the evaluators' confidence in their individual inspections, rather than alerted them to the evaluator effect.

Keywords

Usability evaluation methods, inspections, evaluator effect

INTRODUCTION

Current usability evaluation methods suffer from a substantial evaluator effect in that different evaluators who examine the same system tend to identify substantially different sets of problems [2]. To work around the evaluator effect it is generally recommended that several evaluators examine a system and combine their individual findings into a group output. This recommendation is commonplace for inspection methods which do not involve users [3, 4] but it appears to be equally valid for evaluators analyzing usability test sessions with users [2].

This study compares the calculated evaluator effect with the evaluators' own perceptions of the evaluator effect after they had met in groups to combine the findings from their individual inspections. Group meetings are an opportunity for evaluators to cope with the evaluator effect and improve their evaluation skills by learning from the differences between their own findings and those of their colleagues. However, research on decision processes has previously found that group discussion after individual work increases people's confidence in their individual work, but not its quality [1]. This is an important issue because recommendations to involve more than one evaluator in evaluations will remain academic and ineffective unless practitioners feel that it pays to spend the extra resources.

METHOD

Evaluators

Twelve professional usability specialists participated in the study as evaluators. One evaluator was, however, removed during the data analysis because he failed to comply with the procedures of the study. On average, the remaining 11 evaluators had 7.3 years of experience with usability work and had conducted 37 usability tests (with users) and 35 usability inspections (without users).

Website and Task Scenario

The inspected website, www.avis.com, is a comprehensive e-commerce site that enables people to rent cars all over the world for specified periods of time. The evaluators were asked to focus their inspections on five user tasks: finding the cost of renting a car, making a reservation, getting an overview of the kinds of cars available, finding out about pick-up locations, and getting information about special deals.

Procedure

The evaluators individually inspected the website and subsequently met for 2 hours in 3-person groups to combine their individual inspections into group outputs. Finally, all evaluators participated in a 2-hour plenary session to discuss their experiences from the evaluation.

The individual inspections were conducted in August, 2001, and documented in written reports listing the usability problems and positive aspects encountered by each evaluator. Each report also contained an executive summary with the 3 most important positive aspects of the website, the 3 most important problems, and up to 3 managerial recommendations. As long as the evaluators complied with

Table 1. Problems broken down by the number of evaluators reporting the problems. No problem was reported by more than 7 evaluators. The evaluators' observations are almost as disparate for the 33 severe problems (second row) as for all 220 problems (first row).

No. of evaluators reporting problem	1	2	3	4	5	6	7	Total
All problems	135	39	25	9	7	4	1	220
Severe problems	14	3	7	2	5	1	1	33

this format they were free to inspect the website in any way they wanted. The evaluators spent an average of 15 hours on their individual inspection and they all performed some type of heuristic evaluation [4].

RESULTS

Based on the 11 evaluators' individual reports a master list of 220 unique problems was compiled. This was done independently by the first two authors. The authors classified 68% of the problem instances identically; disagreements were resolved through discussion and a consensus was reached. Most of the authors' disagreements were resolved by combining the problem instances into fewer, more frequently reported problems.

Each evaluator reported an average of 35 (16%) of the 220 problems. The overlap in reported problems between any two evaluators averaged only 9%. Table 1 breaks the problems down by the number of evaluators reporting the problems. As many as 174 (79%) of the problems were reported by just one or two evaluators. Hence, the 11 inspections exhibit a substantial evaluator effect.

The evaluator effect would be less critical if severe problems were reported more consistently than cosmetic problems, which have little impact on the usability of the website. A problem was defined as severe if it appeared in one or more executive summaries. Each evaluator reported an average of 24% of the 33 severe problems. Table 1 shows that 17 (52%) of the severe problems were reported by just one or two evaluators. Hence, the evaluator effect persists for severe problems.

The substantial differences in the individual reports stand in stark contrast to the perception the evaluators acquired during the group work. The evaluators left the group work with a strong, and reassuring, feeling of agreement. This became evident during the plenary session as exemplified by the following quotes from five of the evaluators:

- "I was surprised to see how little we disagreed."
- "A very high level of agreement."
- "It is not that subjective after all. There is consensus about what the problems are."
- "General agreement, but a number of concrete details differ."
- "We are all in agreement. We haven't made the same observations, t hough."

Nobody countered these statements. The evaluators did not have access to data like Table 1 but during the group work each evaluator was confronted with the findings and opinions of two colleagues. Thus, the evaluators knew there were differences in their observations, but they did not perceive these differences as disagreements.

DISCUSSION

The evaluators generally perceived the differences in their observations as multiple sources of evidence in support of the same issues, not as evidence of an evaluator effect. For example, when one evaluator reported one unclear error message and another evaluator reported another unclear error message they tended to experience increased confidence in their individual inspections. This reinforced confidence may stem from a failure to appropriately distinguish specific problems from categories of problems. However, the specific problems reported by any one evaluator point toward substantially different revisions of the evaluated website when compared to the specific problems reported by any other evaluator. This is clearly illustrated by the absence of problems – severe as well as non-severe – that were reported by more than 7 evaluators.

Evaluators who feel that they are in agreement with their peers are unlikely to believe that it pays to involve multiple evaluators in their inspections. Thus, evaluators will tend to consider the evaluator effect negligible, and as a result the evaluator effect remains a major threat to the reliability of usability inspections.

ACKNOWLEDGEMENTS

Morten Hertzum was supported by a grant from the Danish National Research Foundation. We thank the usability specialists who participated in the study, known as CUE-3.

REFERENCES

1. Heath, C., and Gonzalez, R. Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes* 61, 3 (1995), 305-326.
2. Hertzum, M., and Jacobsen, N.E. The evaluator effect: A chilling fact about usability evaluation methods. To appear in *International Journal of Human-Computer Interaction*.
3. Lewis, C., and Wharton, C. Cognitive walkthroughs, in M. Helander, T.K. Landauer, and P. Prabhu (eds.), *Handbook of Human-Computer Interaction. Second, completely revised edition* (pp. 717-732). Elsevier, Amsterdam, 1997.
4. Nielsen, J., and Molich, R. Heuristic evaluation of user interfaces, in *Proceedings of CHI´90* (Seattle WA, April 1990), ACM Press, 249-256.