The Evaluator Effect in Usability Tests

Niels Ebbe Jacobsen
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3891
jacobsen@cs.cmu.edu

Morten Hertzum
Department of Computer Science
University of Copenhagen
DK-2100 Copenhagen, Denmark
hertzum@diku.dk

Bonnie E. John
HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3891
bej@cs.cmu.edu

ABSTRACT
Usability tests are applied in industry to evaluate systems and in research as a yardstick for other usability evaluation methods. However, one potential threat to the reliability of usability tests has been left unaddressed: the evaluator effect. In this study, four evaluators analyzed four videotaped usability test sessions. Only 20% of the 93 unique problems were detected by all four evaluators and 46% were detected by only a single evaluator. Severe problems were detected more often by all four evaluators (41%) and less often by only one evaluator (22%) but a substantial evaluator effect remained.

Keywords
Usability, user testing, usability test, evaluator effect

INTRODUCTION
Usability testing—also known as think-aloud studies—is probably the single-most important method for practical evaluation of user interfaces [2]. Although there is no standardized procedure for running usability tests they are typically conducted in controlled environments and include a number of sessions involving a user working on set tasks while thinking out loud. Based on live observations, or analyses of video tapes, an evaluator constructs a problem list from the difficulties the users have accomplishing the tasks. Previously, it has been shown that four or five users detect roughly 80% of the problems in an interface [4] as long as the average likelihood of a user detecting a problem ranges between 0.32 and 0.42 [1].

While the sufficient number of users has been studied in different ways, the effect of different evaluators analyzing the same sessions has been left untouched. This study investigates how evaluators differ in analyzing identical video tapes of a usability test. So far the evaluator effect has only been studied for Heuristic Evaluation, where it has a substantial impact. In a test of a voice response system a single novice evaluator found 22% (on average) of the problems found by all evaluators in the study. Usability specialists and so-called double experts did somewhat better finding on average 41% and 60% of the problems, respectively [3].

METHOD
Evaluators
Four HCI research evaluators, all familiar with the theory and practice of usability testing, analyzed four video tapes. Table 1 shows the evaluators' experience with the system evaluated in this study, and their evaluation experience in terms of the total number of users previously analyzed.

Evaluator	Occupation	System experience	Evaluation experience
E1	Associate professor, HCI	10 hours	52 users
E2	Doctoral student, HCI	5 hours	4 users
E3	Assistant professor, HCI	2 hours	6 users
E4	Usability lab manager	12 hours	66 users

Table 1. Evaluator profiles

Video tapes
Four Macintosh users spent about an hour thinking aloud as they worked through four or five set tasks in a multimedia authoring system [6]. Although they were experienced computer users, none of the users had previous experience with the system and no instructions were given. The system resembles an advanced word processor insofar that the user can create documents consisting of plain text, still graphics, movies, and animations. The users were asked to create several pages consisting of some text, a figure, and an animation, to add some items to the glossary and the table of contents, and to switch two pages. The same experimenter ran all four experiments, and he did not interrupt the users unless they forgot to think aloud, explicitly gave up solving a task, or got stuck for more than three minutes.

Procedure
Evaluators E1 and E2 knew the authoring system well, while evaluator E3 and E4 were asked to familiarize themselves with it prior to their analysis. The evaluators had access to a written system specification (35 pages) and the running system throughout their participation in the study. The evaluators were asked to detect and describe all problems in the interface based on analyzing the four tapes in a preset order. No time constraints were enforced. The evaluators were requested to report three properties for each problem detected: (a) evidence consisting of the users' action sequence and/or verbal utterances, (b) a free-form problem description, and (c) the criteria for identifying the problem. The evaluators used nine predefined problem criteria: (1) The user articulates a goal and cannot succeed in attaining it within three minutes, (2) the user explicitly gives up, (3) the user articulates a goal and has to try three or more actions to find a solution, (4) the user produces a

result different from the task given, (5) the user expresses surprise, (6) the user expresses some negative affect or says something is a problem, (7) the user makes a design suggestion, (8) a system crash, and (9) the evaluator generalizes a group of previously detected problems into a new problem.

RESULTS AND DISCUSSION

Based on the four evaluators' individual problem reports a master list of 93 unique problem tokens (UPTs) was constructed. This was done independently by the first two authors. They agreed on 84% of the UPTs; disagreements were resolved through discussion and a consensus was reached.

The percentage of the UPTs reported by E1, E2, E3 and E4 were 63%, 39%, 52%, and 54% respectively. Thus, a single evaluator detected on average 52% of the problems, which is only slightly more than the 41% found by usability specialists in a Heuristic Evaluation [3].

Table 2 shows the evaluator effect for all UPTs and for just the severe problems. We define a severe problem as a UPT judged by one or more evaluators to violate problem criteria 1, 2 or 8. The evaluator effect for all UPTs is substantial, 46% of all UPTs were found by only a single evaluator. Though the severe problems were generally detected by more than one evaluator, only 41% of the severe problems were detected by all four evaluators. Furthermore, 73% of the severe UPTs were judged as violating problem criteria other than 1, 2 or 8 by some evaluator. Thus, any given evaluator should not be expected to use problem criteria as a reliable method to judge severity on UPTs.

	1 evaluator	2 evaluators	3 evaluators	4 evaluators
Severe UPTs	8 (22%)	7 (19%)	7 (19%)	15 (41%)
All UPTs	43 (46%)	19 (20%)	12 (13%)	19 (20%)

Table 2. Number of UPTs detected by groups of 1, 2, 3, and 4 evaluators.

The effect of adding more evaluators to a usability test resembles the effect of adding more users; both additions increase the overall number of UPTs found. As illustrated in Figure 1, the average increase was 46% going from one to two evaluators, 23% going from two to three evaluators, and 17% going from three to four evaluators when all four users were included in the calculation (the four points on the rightmost vertical). Calculating the effect of running more users we found an increase of 55% going from one to two users, 26% going from two to three users, and 23% going from three to four users when all evaluators were included in the calculation (the topmost curve). The declining number of new UPTs detected as more users are added confirms the results from similar studies [1, 4, 5].

We were able to describe the number of UPTs found based on the number of users and the number of evaluators:

$$\text{No. of UPT} = 19.35 * (\text{no. of evaluators})^{0.505} * (\text{no. of users})^{0.661} \quad \text{(eq.1)}$$

The fit between the mathematical model and our data is highly significant (R^2=0.997; SEE=2.6%). The four or five users that others have found to be sufficient in a usability test could, in our study, be traded for two evaluators running two or three users without decreasing the number of detected UPTs.

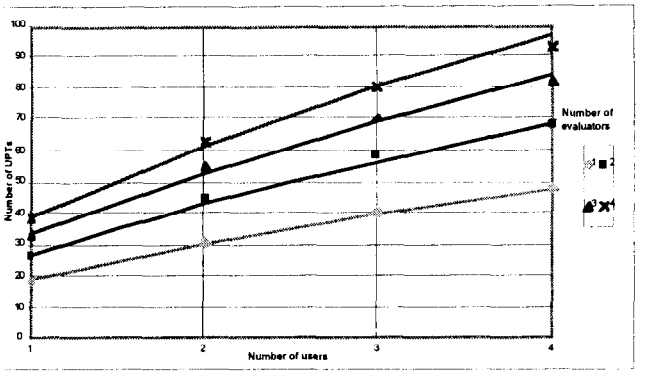

Figure 1 Data points are the observed average number of UPTs detected by all sets of users and of all sets of evaluators. The curves plot Eq. 1 for 1, 2, 3, and 4 evaluators.

CONCLUSION

This study demonstrates that four evaluators, each analyzing video tapes of four usability test sessions differed a great deal in their analysis. Only a fifth of the UPTs were detected by all evaluators, while almost half of the UPTs were detected by only one evaluator. Severe problems faired better, with two fifths detected by all evaluators and only one fifth by a single evaluator. Given previous studies of individual differences the results of this study is not very surprising but certainly overlooked or neglected in the field of usability testing. The evaluator effect puts usability testing in perspective, and questions the use of data from usability tests as a baseline for comparison to other usability evaluation methods.

Additional research is needed to understand how this effect varies by evaluator experience, problem severity, task-type, system-type, and other variables important to usability practitioners and researchers.

REFERENCES

1. Lewis, J. Sample Sizes for Usability Studies: Additional Considerations. *Human Factors 36, 2* (1994), 368-378.
2. Nielsen, J. *Usability Engineering.* Academic Press, Boston, 1993.
3. Nielsen, J. Finding usability problems through heuristic evaluation, in *Proceedings of CHI'92* (Monterey, CA, May 1992), ACM Press, 373-380.
4. Nielsen, J. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies, 41* (1994), 385-397.
5. Nielsen, J. & Landauer, T.K. A Mathematical Model of the Finding of Usability Problems. in *Proceedings of INTERCHI'93* (Amsterdam, Holland, April 1993), ACM Press, 206-213.
6. Pane, J.F. & Miller, P.L. The ACSE multimedia science learning environment, in *Proceedings of the 1993 International Conference on Computers in Education* (Taipei, Taiwan, 1993).