H.-J. Bullinger & J. Ziegler (eds.), *Human-Computer Interaction: Ergonomics and User Interfaces. Proceedings of HCI International '99* (Munich, August 22-26, 1999), vol. I, pp. 1063-1067. Lawrence Erlbaum Associates, London.

# The Evaluator Effect during First-Time Use of the Cognitive Walkthrough Technique[†]

**Morten Hertzum**
Centre for Human-Machine Interaction
Risø National Laboratory, Denmark

**Niels Ebbe Jacobsen**
Department of Psychology
University of Copenhagen, Denmark

## 1 Introduction

Practising system developers without a human factors background need robust, easy-to-use usability evaluation methods. The cognitive walkthrough (CW) technique (Lewis et al. 1990, Wharton et al. 1994) has been devised to provide such a method and is particularly suited to evaluate designs before testing with users becomes feasible and as a supplement to user testing.

While several studies have evaluated how well CW predicts the problems encountered in thinking-aloud studies (e.g. John and Mashyna 1997, Lewis et al. 1990), only Lewis et al. have assessed to what extent different evaluators obtain the same results when evaluating the same interface. Data from Lewis et al. suggests that the variability in performance among evaluators using CW is much lower than that of evaluators using heuristic evaluation or thinking-aloud studies (Jacobsen et al. 1998, Nielsen 1994). One reason for this seemingly higher robustness of CW might be that it is a quite structured process. CW has however evolved considerably since the study of Lewis et al. Moreover, their data was limited in sample size and applicability to actual CW evaluators.

To inform practitioners and methods developers about the robustness of CW this paper investigates to what extent novice evaluators who perform a CW of the same tasks detect the same problems in the evaluated interface. While acknowledging the importance of choosing the right tasks in a CW, we have decided to focus on the actual walkthrough process.

---

[†] This work was supported by grants from the Danish National Research Foundation and the Danish Research Councils. We wish to thank the evaluators for their time and effort.

## 2 The CW Technique

Our study is based on the version of CW described in Wharton et al. (1994). CW consists of a preparation phase and an execution phase. In the preparation phase the evaluator describes a typical user, chooses the tasks to be evaluated, and constructs a correct action sequence for each task. In the execution phase the evaluator asks four questions for each action in the action sequences: (1) Will the user try to achieve the right effect? (2) Will the user notice that the correct action is available? (3) Will the user associate the correct action with the effect trying to be achieved? (4) If the correct action is performed, will the user see that progress is being made toward solution of the task? With the description of the user in mind the evaluator decides whether each question leads to a success or failure story. In case of a failure story a usability problem has been detected.

## 3 Method

Eleven graduate students in computer science evaluated a prototype of a Web-based system against set tasks. Half of the evaluators had design experience from industry, but they had no prior knowledge of the system to be evaluated. The evaluated system, called HCILIB, was a prototype of a Web-based library giving access to a collection of scientific articles on human-computer interaction. HCILIB (Perstrup et al. 1997) integrates Boolean search with a scatter-gather inspired technique to display a browsable structure of the collection. Boolean searches can be expressed as conventional Boolean queries (using ANDs and ORs) or by means of a Venn diagram metaphor. The Venn diagram metaphor relieves the user from direct interaction with logical expressions. Instead, query terms are entered into two search boxes, A and B, and the search results are automatically sorted into three disjunctive sets corresponding to A–B, A∩B, and B–A.

The experiment was embedded in a grade-giving assignment where the students were asked to construct action sequences and do a cognitive walkthrough of three set tasks. Just before the assignment was handed out the evaluators received two hours of instructions in CW based on a lecture on the practitioner's guide to CW (Wharton et al. 1994). The instructions also offered the evaluators some hands-on experience followed by instant feedback. The evaluators documented their cognitive walkthroughs in a problem list describing each detected problem. As a rough estimate each evaluator spent 2-3 hours completing his/her CW. Based on the problem lists from the 11 evaluators the two authors independently constructed a master list of unique problem tokens. Combining these master lists we had an inter rater reliability of 80%; disagreements were resolved through discussion and a consensus was reached.

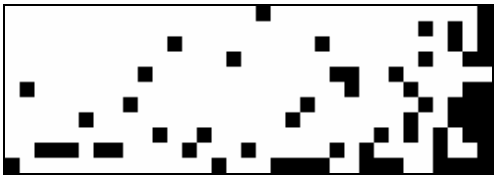

Figure 1. Matrix showing who found which problems. Each row represents an evaluator, each column a problem, and each black square that the evaluator detected the problem.

## 4 Results

The eleven evaluators reported a total of 74 problem instances from their CWs. These problem instances made up 33 unique problem tokens (in the following just termed problems). As much as 58% of the problems were detected by only a single evaluator, and no single problem was detected by all evaluators, see Figure 1. A single evaluator found on average 18% of the 33 known problems.

We were curious to know how groups of evaluators performed compared to single evaluators. Figure 2 shows the average number of problems that would be found by aggregating the sets of problems found by different groups of evaluators. For each group a given problem was considered found if it was found by at least one of the evaluators in the group. The results suggest a great deal of misses – or false alarms – in the performance of single evaluators. An analysis of problem severity could not explain this evaluator effect, as the detection rate for

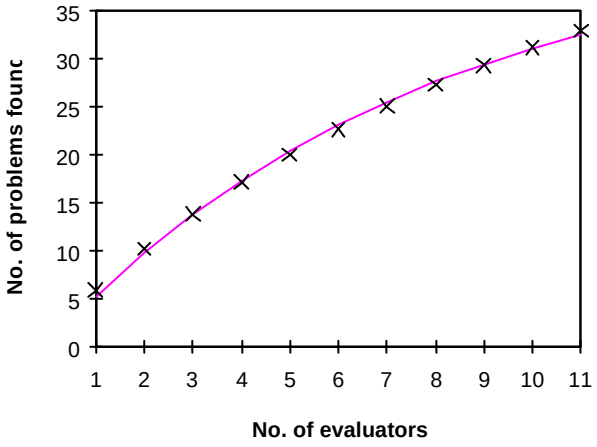


Figure 2. The number of problems detected shown as a function of the number of evaluators. The data points are the aggregated values from the experiment. The curve plots an estimate given by the formula $f(k) = n(1 - (1 - p)^k)$, for $n = 43$ and $p = 0.121$.

severe problems was only marginally higher than for the entire set of problems.

## 5 Discussion

As for other usability evaluation methods it is crucial to CW that the walkthrough leads to a reliable problem list. Studies on heuristic evaluation and usability studies have found substantial individual differences in the evaluators' performance (Jacobsen et al. 1998, Nielsen 1994). This suggests that our results are partly attributable to usability evaluation in general, rather than solely to CW. We believe, however, that the CW procedure falls short of providing the evaluators with a feel for the users and thus becomes inaccurate for two reasons: (1) Anchoring, i.e. despite the evaluator's efforts the walkthrough will end up evaluating the system against a user who is much too similar to the evaluator to be representative of the actual users. (2) Stereotyping, i.e. the walkthrough will end up reflecting a user that is much too homogeneous to accommodate the diversity of the actual users of the system evaluated.

We investigated the anchoring and stereotyping hypotheses by looking closer on how the evaluators answered the four questions on identical actions. Though the evaluators constructed their action sequences from the same three tasks only 4 out of an average of 15 actions were identical across all evaluators. One of these actions is to execute a query by activating a Query button. In evaluating this action three evaluators reported success stories on all four questions, while eight evaluators reported a total of five different problems: Three evaluators reported that the user will click a Venn pictogram situated above the Query button, rather than the button itself. Three evaluators reported that there is weak feedback from the system after clicking the Query button. Two evaluators reported that the Enter key does not execute the query, i.e. the user has to use a pointing device. One evaluator reported that the caption on the button should be changed. And finally, one evaluator reported that the user will forget to activate the Query button. It seems quite reasonable that all problems would actually happen for some users in a real situation, just as some users might experience no troubles using the Query button, as suggested by three evaluators. Though all evaluators' use of the four questions on the analysed action seems reasonable, the outcome is very different across evaluators. The same pattern was found for the three other actions that were identical across the evaluators.

The evaluators' descriptions of the target user in the preparation phase are similar in content, and they generally provide a broad description of a large, homogeneous group of users. The descriptions are in many respects similar to the descriptions of users given as examples by Wharton et al. (1994). Despite the formal description of the user, or perhaps because of the generality of these descriptions, the evaluators might not fully realise the heterogeneity of the user

group or their walkthrough might be anchored to their own experience with the system. Each of the four questions drives the evaluator to think of the user's behaviour in a certain situation. When the fictive user description becomes too fuzzy or lacks details to judge the user's behaviour, the evaluator unintentionally substitutes the description with a particular user much like herself/himself. Thus, evaluators tend to produce success stories if they imagine themselves having no troubles using the feature in question, and they report problems when they imagine themselves having troubles in the particular situation. In this sense a single evaluator using CW resembles an evaluator performing a thinking-aloud study with one user, namely himself/herself.

Wharton et al. (1994) state that CWs can be performed by individual evaluators as well as by groups of co-operating evaluators. For inexperienced CW evaluators our study strongly indicates that several evaluators are necessary to achieve a performance that is acceptable for practical use of the CW technique. Additional studies are required to learn how more experienced evaluators perform and to study more closely *why* we see these individual differences.

## 6 References


Jacobsen, N. E., Hertzum, M. & John, B. E. (1998). The evaluator effect in usability studies: problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (Chicago, October 5-9, 1998), pp. 1336-1340. Santa Monica: HFES.

John, B. E. & Mashyna, M. M. (1997). Evaluating a multimedia authoring tool. *Journal of the American Society for Information Science*, 48(11), 1004-1022.

Lewis, C., Polson, P., Wharton, C. & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the ACM CHI'90 Conference* (Seattle, April 1990), pp. 235-242. New York: ACM Press.

Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J. & Mack, R. L. (Eds.): *Usability Inspection Methods*, pp. 25-62. New York: John Wiley.

Perstrup, K., Frøkjær, E., Konstantinovitz, M., Konstantinovitz, T., Sørensen, F. S. & Varming, J. (1997). A World Wide Web-based HCI-library designed for interaction studies. In *Third ERCIM User Interfaces for All Workshop* (Obernai, France, November 1997).

Wharton, C., Rieman, J., Lewis, C. & Polson, P. (1994). The cognitive walkthrough method: a practitioner's guide. In Nielsen, J. & Mack, R. L. (Eds.): *Usability Inspection Methods*, pp. 105-140. New York: John Wiley.