What is Gained and Lost When Using Methods Other Than Empirical Testing

Heather Desurvire
Jim Kondziela
Michael E. Atwood

NYNEX Science and Technology
Artificial Intelligence Laboratory
Intelligent Interfaces Group
500 Westchester Avenue
White Plains, NY 10604
Phone: (914) 644-2757
Email: hwd@nynexst.com

Introduction

Traditional laboratory usability testing is frequently not performed due to a company's lack of funds, planning, or human factors expertise. Consequently, there is increasing interest in finding alternative usability testing methods that are easier and cheaper to implement than traditional laboratory usability testing. Recent studies are beginning to study and compare such techniques. These methods include Heuristic Evaluation (Nielsen and Molich, 1990), and Cognitive Walkthrough (Polson, Lewis, Rieman, & Wharton, 1990). For Heuristic Evaluation, Nielsen (1991) found that human-factors Experts were the best at finding an interface's usability problems, especially Experts who were also expert in the interface domain. Desurvire, Lawrence, & Atwood (1991) found Experts' evaluations were the most reliable, and their best guess predictions were predictive of laboratory performance. Karat, Campbell, & Fiegel (1992) similarly found that heuristic results were reliable and significantly predictive of laboratory data, yet empirical laboratory testing identified four and five times as many problems. Jeffries, Miller, Wharton, and Uyeda (1991) found that via Heuristic Evaluation, more severe problems were found than with laboratory testing or the Cognitive Walkthrough. This comparison study did, however, only utilize Experts in the heuristic condition, and System Designers in the Cognitive Walkthrough.

We were interested to set up a study where the methods and type of evaluator expertise would be compared to each other, and were comparable to the empirical laboratory testing data. We were interested in learning not only how predictive each method by group was of laboratory performance, but also what each type of evaluator and method brought to the evaluation. In other words, what might we lose if we did not perform laboratory tests, and what might we gain?

Methods

The study compares Heuristic Evaluation and the Cognitive Walkthrough evaluation methods to lab testing results on a telephone -based interface. Three evaluator groups were utilized in each method: human-factors Experts, Non-Experts, and the original system s System Designers. Each evaluator group had 3 people.

Laboratory

A traditional usability test of a telephone-based interface. was performed (n=18) where users performed 6 tasks. The experimenter observing the trial noted problems and the corresponding problem severity and the perceived attitude of the user towards these problems. The experimenter also collected task completion data, error data, time to complete task, and the number of tries to complete the task.

Alternative Usability Methods and Evaluators

The telephone-based interface being evaluated was conveyed to the evaluator groups by paper flow-charts, organized by task. Each evaluator group for both methods was given approximately 3 hours to perform a complete evaluation on a set of the same tasks performed in the lab.

Cognitive Walkthrough

The Cognitive Walkthrough is an evaluation method that attempts to simulate the interaction between the user and an interface while in the process of performing a task (Polson, Lewis, Rieman, & Wharton,1990; Lewis, Polson, Wharton & Rieman, 1991). . A series of questions are asked which attempt to reveal to the evaluator how the user's goals match the actions which are a result of the interface design. With this procedure, potential problems are named, and predictions of the percentages of users who will have this problem are also generated. Due to the amount of time it took to conduct this method, only 3 of the 6 tasks were evaluated.

Heuristic Evaluation

A method developed by Nielsen and Molich (1990), Heuristic Evaluation, utilizes a set of heuristics by which the evaluators must judge the interface. The group members were given a short lecture on the usability heuristics they were to use for evaluating the interface. After studying the user tasks and system via flow charts, the evaluators named all problems, using the same form as was used in the laboratory. "Best Guess" predictions were made on the same variables as were collected in the laboratory in a manner similar to predictions made in the first comparative study performed by Desurvire, et al.(1991).

Results

Evaluation Methods vs. Laboratory

The Experts heuristically evaluating the interface's tasks found the highest percentage of problems that actually occurred in the laboratory (44%), followed by the Experts using the Cognitive Walkthrough (28%). The next best at predicting the laboratory problems were the System Designers in both methods. System Designers and Non-Experts did not differ in their predictions between evaluation methods, (see Table 1 below for details).

Tab!e 1. The Percentage of Problems Evaluators Found that Did Occur in the Lab and Could Potentially Occur.

	Did Occur		Potential	
Lab	25		29	
	Percentage (Number of Problems)			
Heuristic Evaluation				
Experts	44%	(11)	31%	(9)
System Designers	16%	(4)	24%	(7)
Non-Experts	8%	(2)	3%	(1)
Cognitive Walkthrough				
Experts	28%	(7)	31%	(9)
System Designers	16%	(4)	21%	(6)
Non-Experts	8%	(2)	7%	(2)

Experts were best at predicting the most serious problems, especially when utilizing Heuristic Evaluation.

The evaluators also listed problems that we categorized as improvements to the interface. Improvement are based on an evaluators' assumption that it may avoid a user's potential problem or annoyance. For example, "prompt <x> is too condescending to the user". Of all the improvements named by all the groups, the heuristic Experts named the highest percentage at 77%, where they named only 16% from the Cognitive Walkthrough method. The other groups named even less (3% to 6%).

Finally, in their listing of problems, the evaluators also named those that were determined to be potential problems, but did not actually occur in the laboratory. For example, these were problems that may have arisen in cases that were not tested in the laboratory. The percentage of problems each evaluator group contributed to the total number of problems found from all the groups were taken. The Experts in both conditions found the greatest number of potential problems (31%). See table 1 for the other groups.

Discussion

Overall, Heuristic Evaluation is better than the Cognitive Walkthrough for predicting laboratory results, but only for Experts. This difference between methods for Experts may be due to the heuristic method's "reminding" Experts to analyze more dimensions of the interface, than the Cognitive Walkthrough is setup to. In addition, Heuristic Evaluation tends to facilitate naming improvements that go beyond the scope of the tasks for Experts. In absolute value, all groups' except for the Experts were quite low in both methods. These methods' strength, more than their predictive value, is information we gain beyond this.

It will be interesting in our next phase of research to compare our results to field data, for analyzing whether type of expertise and method are more predictive of field than laboratory data. Future research should also include studying other usability methods, such as Co-Operative Evaluation (Wright, Monk, & Carey, 1991) and Bastien and Scapin's (1991) Ergonomic Criteria for evaluation.

Acknowledgements

We would like to thank J. Rieman for his assistance and for his Cognitive Walkthrough software, J. Nielsen, C. Karat, and J. McKendree for useful discussions and comments on earlier drafts, and to those who participated in our study.

References

1. Bastien, C. & Scapin, D. L. A validation of ergonomic criteria for the evaluation of user interfaces. SIGCHI Bulletin,23,4,1991, pp.54-55.
2. Desurvire, H., Lawrence, D., and Atwood, M. Empiricism versus judgment: Comparing user interface evaluation methods on a new telephone - based interface.SIGCHI Bulletin,23,4,1991,pp.58-59.
3. Jeffries, R.J.,Miller, J.R., Wharton, C., &Uyeda. K.M. User interface evaluation in the real world: A comparison of four techniques. In Proceedings of CHI'91, (New Orleans), ACM, NY, pp. 119-124.
4 Karat, C.,Campbell, R., & Fiegel, T. Comparison of empirical testing and walkthrough methods in user interface valuation (in press).
5. Lewis, C. Polson, P., Wharton, C., & Rieman, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In Proceedings of CHI'90(Seattle),ACM, pp. 235-242.
6. Nielsen, J. Finding Usability Problems Through Heuristic Evaluation. In Proceedings of CHI'92, (Monterey), ACM, in press..
7. Nielsen, J. and Molich, R. Heuristic evaluations of user interfaces. In J.C. Chew & J. Whiteside (Eds.), Proceedings CHI' 90,(Seattle), ACM,pp. 249-256.
8. Polson, P., Lewis, C., Rieman, J., & Wharton, C. Cognitive Walkthroughs: A method for theory-based evaluation of user interfaces. in press,1991.
9. Wright, P.C.,Monk, A.F. and Carey. Co-Operative Evaluation. The York Manual, V1, 1991, University of York, England.