Enhancing the Performance of Interface Evaluators Using Non-Empirical Usability Methods

Heather Desurvire and John C. Thomas
NYNEX Science and Technology, Inc.
500 Westchester Avenue
White Plains, NY 10604

ABSTRACT

Heuristic Evaluation has been shown to be a quick cost-effective methodology that can lead to early identification of many of the same user interface errors as laboratory usability studies. In this paper, we describe a method designed to enhance the performance of expert, system developer, and non-expert evaluators. The evaluators most proficient at Heuristic Evaluation are Human-Factors Experts (Desurvire, Lawrence and Atwood, 1991; Desurvire, Kondziela and Atwood, 1992; Jeffries, Miller, Wharton and Uyeda, 1991) and most notably, "double experts" (Nielsen, 1992). Similar results were obtained for the Cognitive Walkthrough developed by Lewis, Polson, Wharton and Rieman, 1990 (Desurvire, et al., 1992; Jeffries, et al., 1991). We were interested in whether a non-empirical method could be developed in which evaluators other than Human Factors Experts can perform nearly as well as Experts.

Desurvire, et al. (1992) found that Heuristic Evaluation and Cognitive Walkthroughs not only predicted problems observed in laboratory studies but also encouraged evaluators to suggest improvements. In addition, non-empirical methods stimulated evaluators to point out problems that would be likely to occur in actual use, but would not be observed in laboratory studies. We were interested in expanding this finding by developing a method that encouraged a broader scope of thinking, and thus a broader evaluation.In this paper, we describe the method Programmed Amplification of Valuable Experts (PAVE) and how it enhanced the performance of System Developers and Non-Human-Factors-Expert evaluators. Future work is discussed in which real users in the field will be compared to these results.

INTRODUCTION

For at least two decades, empirical laboratory usability evaluation of user interfaces has been the major methodology of choice to ensure usable systems. In many cases, such studies can be conducted in advance of actually building the system, via prototyping, pencil and paper simulations, or "Wizard-of-Oz" studies (Thomas and Gould, 1975; Thomas, 1976). As businesses have had to face increasing competitive pressures, there has been a growing recognition of the necessity of more usable internal and commercial products. At the same time, there is increasing pressure to make the process of ensuing usable products faster and cheaper. Major techniques aimed at achieving these goals include the use of Guidelines, (e.g., Smith and Mosier, 1986), Cognitive Walkthroughs (Lewis, Polson, Wharton and Rieman, 1990), formal models (e.g., GOMS; (Card, Moran, and Newell, 1983), and especially Heuristic Evaluation (Nielsen and Molich, 1990).

While Heuristic Evaluation is especially promising, it is most effective when used by Human Factors Experts (Desurvire, Kondziela and Atwood, 1992; Karat, Campbell and Fiegel, 1992), especially when multiple Experts are used (Nielsen and Molich, 1990). Evaluation performance is enhanced further when the Experts are "double-experts", knowledgeable in both general human-computer interaction and in the specific type of interface being evaluated (Nielsen, 1992). Likewise, the CognitiveWalkthrough (Lewis, Polson, Wharton and Rieman, 1990) was also most useful when performed by Experts, as opposed to Non-Experts and System Developers (Desurvire, et al., 1992; Jeffries, Miller, Wharton and Uyeda, 1991). Clearly, for these methods, human factors expertise is essential to carry out effective evaluations. Yet, the reality is that many development efforts do not have the right expertise to utilize any of the aforementioned techniques. We were interested in developing a method whereby evaluators other than Experts could find nearly as many errors as Experts in Heuristic Evaluation and Cognitive Walkthroughs.

Earlier studies also found that many of the interface "errors" found by non-experts were "false positives"; that is, these user errors could not actually occur (Desurvire, Lawrence and Atwood, 1991; Desurvire, et al., 1992. We were also interested therefore, in whether this new approach, PAVE (described below) would reduce the proportion of these false positives.

A limitation of laboratory studies is that they may fail to identify problems that will occur in the real world due to the inevitable differences between the laboratory and the field in tasks, users, and context (Thomas and Kellogg, 1989). Another motivation then, in the development of PAVE was to encourage evaluators to consider alternative users, tasks, and contexts. Since we used PAVE on a real system that will actually be deployed, at some future date, we will be able to evaluate the effectiveness of PAVE in that regard. A field trial is planned to compare lab and non-empirical data from the Desurvire, et al, 1992 study, to learn what is lost and gained from the real world. For the current study, the laboratory data from Desurvire, et al, 1992, is used as the benchmark.

We began developing PAVE by building on previous work and utilizing methods from brainstorming. An earlier study found organizing information useful for evaluators (Scapin, 1990). Our approach is to augment the evaluator's existing knowledge through the use of a set of perspectives that stimulates the evaluator to think about usability more broadly. We developed these ideas by grouping them into a set of knowledgeable experts (Perspectives) that would focus the evaluator on various aspects of an interface. For example, by using the Perspective of sociology, we were interested in whether the evaluator might detect social issues that might arise from using a new interface. For example, if the interface was made so easy for all members of a work group to use, as opposed to a few, this might change the social hierarchy and motivations of work and success within the group. Ten such perspectives were utilized in the experiment to focus on ten areas we felt might be of importance when evaluating a user-interface.

METHOD

Three evaluators from each of three groups: Human Factors Experts, Non-Human-Factors-Experts (Non-Experts), and System Developers were asked to study flow charts of the same voice interface that was used in the earlier study (Desurvire, et al., 1992). Each evaluator then was asked to study the interface several times; once from each of several quite different perspectives. These Perspective included: whatever knowledge the evaluator typically brought to bear (Self), a Human Factors Expert, a Cognitive Psychologist, a Behaviorist, a Social/Community Psychologist, an Anthropologist, a Freudian, a Health Advocate, a Worried Mother, and a Spoiled Child.

All evaluators received the same order of Perspectives as listed above, starting with the most typical and professional, proceeding toward the broader and more unusual. After reading a short orientation toward each perspective, the evaluators looked at a flow chart for each of the same three tasks, noted probable user errors, and suggested improvements; they were instructed not to repeat the same comments from each successive perspective. When the evaluations were completed, they were asked to comment on PAVE and list which Perspectives were most valuable.

It took approximately 2 to 3 hours for each evaluator to complete the PAVE method evaluation. This was the same amount of time it took the evaluators to complete the cognitive walkthrough and heuristic evaluation reported in Desurvire, et al. (1992).

RESULTS

Scope of Problems

Experts, Developers, and Non-Experts using PAVE found additional problems unforeseen by any of the previous non-empirical usability methods.

Reliability

Non-Experts were still the most unreliable; 10% of their named problems could not occur; however, this is much better than the 55% false positives they named with heuristic evaluation (Desurvire, et al., 1992).

Which Perspectives Enhance Evaluation?

Each of the 10 Perspectives contributed some amount of help to the evaluation for predicting lab results, naming improvements, and identifying problems that could potentially occur in the interface. Of the more unusual Perspectives, the Worried Mother and Freud contributed substantially to predicting laboratory results. For suggesting improvements, the Sociologist was the most facilitative. For problems that could potentially occur in the field, but did not occur in the laboratory studies due to the constraints of the task sets, the Worried Mother and the Anthropologist contributed substantially.

PROCEEDINGS of the HUMAN FACTORS AND ERGONOMICS SOCIETY 37th ANNUAL MEETING—1993

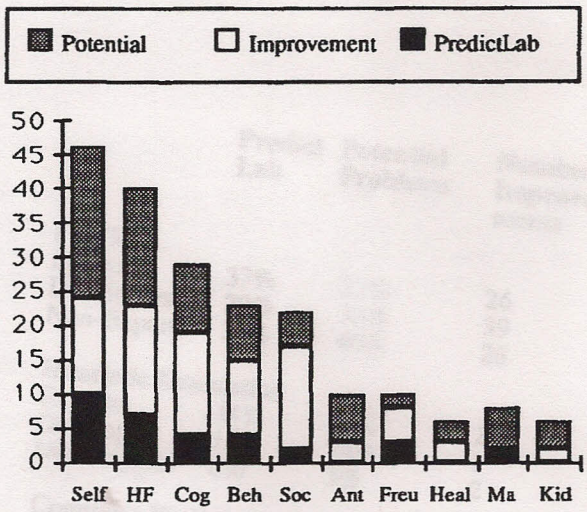

Figure 1. Number of problems each successive PAVE Perspective added to predicting lab problems, potential, and suggested improvements

Accumulated Severity Scores

Evaluators were asked to rate the severity of each problem they found with the user-interface. The scale is a 3-point scale, in which 1 was the least severe problem, and a 3 was the most severe. The severity ratings were added together and averaged for each evaluator group. The Experts found a severity rating of problems of 20 when performing their own evaluations, as compared to a severity of 5 and 7 for Non-Experts and System Developers. Experts' evaluations dropped dramatically for the Perspectives subsequent to the Self evaluations; the System Developers showed increases; and the Non-Experts showed fairly consistent severity ratings across all Perspectives. One caveat is that the first few perspectives, since they were the first to be performed, had an unfair advantage over the last few perspectives. Since we had so few subjects, we thought it to be a moot point to counterbalance the order of perspectives given to the evaluators. It is with this understanding that both figures should be read.

Contribution Of Perspectives By Evaluator

Experts and non-experts found more problems with the "self" Perspectives, than any other perspective. System Developers found nearly the same number of problems from the Perspectives: Self, Human Factors Engineer, Cognitive Psychologist, and Behaviorist. Non-Experts benefited from all the Perspectives, especially the Self, Social Psychologist and Cognitive Psychologist.

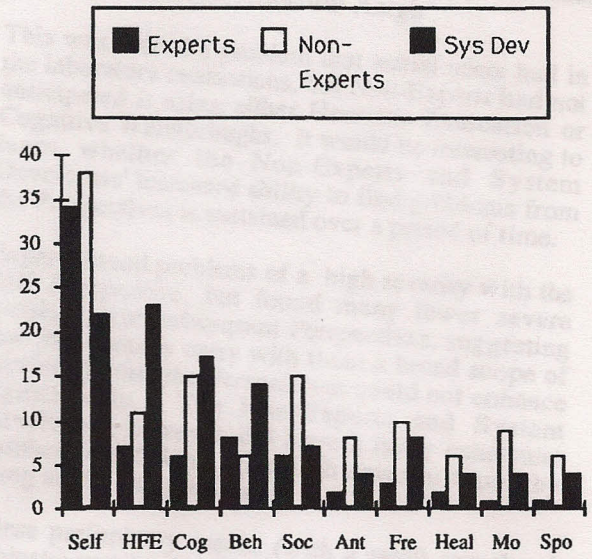

Figure 3. Number of problems per Perspective contributed by each evaluator

Predicting Laboratory Results

Expert evaluators using PAVE were fairly comparable to Experts using the Cognitive Walkthrough method and generally did somewhat worse than Experts using Heuristic Evaluation. However, Developers and Non-Experts did better with PAVE in all categories than with either Heuristic Evaluation or Cognitive Walkthrough; in some categories, the improvement was quite substantial (see Table 1).

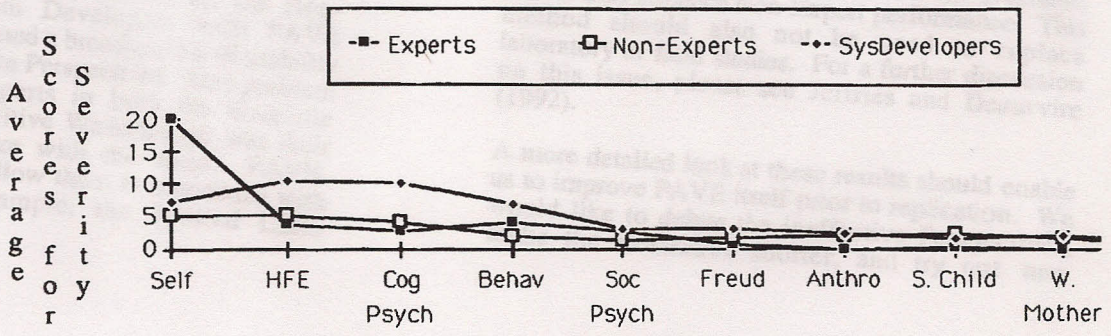

Figure 2. Accumulated severity scores averaged for each evaluator group for each perspective.

	Predict Lab	Potential Problems	Number of Improvements
PAVE			
Experts	37%	27%	26
Developers	29%	33%	39
Non-Experts	34%	40%	26
Heuristic Evaluation			
Experts	44%	31%	24
Developers	16%	24%	1
Non-Experts	8%	3%	2
Cognitive Walkthrough			
Experts	28%	31%	5
Developers	16%	21%	1
Non-Experts	8%	7%	2

Table 1. Percentage of problems evaluators found, predicting lab, potential problems, and number of improvements.

Evaluators Post-Test Comments

The Human Factors Experts and the System Developers both thought the Human Factors Perspective was the most useful for evaluation, and the majority of these evaluators pointed out that the Cognitive Psychologist, Sociologist, and Anthropologist were helpful. Interestingly, the Non-Experts especially thought their own evaluations, as well as the Spoiled Child were the most helpful. Evaluators suggested that we include the User, Legal, and Security Perspectives. Almost half of the evaluators thought the Worried Mother Perspective should be deleted, and that the evaluation should be shorter since their focus and concentration tended to diminish at the end.

DISCUSSION

The improvement of performance for the Non-Experts and System Developers with PAVE suggests that they gained a broader view of usability when exposed to these Perspectives. One problem found with Non-Experts in both the Heuristic Evaluation and Cognitive Walkthrough was their inability to empathize with the users. PAVE, however seemed to allow them to empathize with the users. For example, the Spoiled Child Perspective allowed one Non-Expert to find that it would be frustrating to try to find the correct password. The Non-Expert reports:

"Hey I hit number <xx> to add a name, and the stupid machine just keeps giving me a dial tone, I can't find the right password! Aargh"

This was indeed a problem that actual users had in the laboratory evaluations, but Non-Experts had not anticipated it using either Heuristic Evaluation or Cognitive Walkthroughs. It would be interesting to learn whether the Non-Experts and System Developers' increased ability to find problems from the Perspectives is sustained over a period of time.

Experts found problems of a high severity with the Self Perspective, but found many fewer severe problems with subsequent Perspectives, suggesting that they already carry with them a broad scope of knowledge that the Perspectives could not enhance significantly. The Non-Experts and System Developers, however, did have a fairly consistent number of problems with high severity found by using all the Perspectives.

These preliminary results (with a small sample of subjects) suggest that the PAVE approach may offer substantial promise as a technique to enhance interface evaluations by Non-Experts and Developers in several ways including: avoiding the flagging of false positives, finding real problems, and offering suggestions for improvements. Like Heuristic Evaluation and Cognitive Walkthroughs, PAVE can be used on paper specifications of a system, an important feature that shows these methods can be performed early on in the development cycle.

While this study suggests that Non-Expert evaluators and System Developers can use PAVE fairly effectively on some criteria, this does not mean that they replace Human Factors Experts. Rather, to the extent such expertise is not available, PAVE may enhance Non-Expert performance. This method should also not be used to replace laboratory or field studies. For a further discussion on this issue, please see Jeffries and Desurvire (1992).

A more detailed look at these results should enable us to improve PAVE itself prior to replication. We would like to delete the ineffective Perspectives, make the evaluation shorter, and try out new

Perspectives we think may be efficacious. The future deployment of this voice system in the field will also allow us to evaluate the effectiveness of PAVE in predicting difficulties observed in the real world which we did not find in the laboratory.

Acknowledgments

We would like to thank the co-authors of the original work from which this study is based: Mike Atwood, Jim Kondziela, and Debbie Lawrence, as well as Wendy Kellogg for useful comments.

References

Card, A., Moran, T.P.., Newell, A. Psychology of Human-Computer Interaction, Erlbaum, New Jersey, 1983.

Desurvire, H.W., Lawrence, D. and Atwood, M.E. Empiricism versus judgment: Comparing user interface evaluation methods on a new telephone-based interface. SIGCHI Bulletin, 23, 4, 58-59, 1991.

Desurvire, H.W., Kondziela, J.M.& Atwood, M.E. What is gained and lost when using evaluation methods other than empirical testing. In Proceedings of HCI International 1992, Cambridge University Press, A. Mark, D. Diaper, M.D. Harrison (Eds), 1992.

Jeffries, R.J., & Desurvire, H. Usability Testing vs. Heuristic Evaluation: Was there a contest? SIGCHI Bulletin, 24,4, 39-41, 1992.

Jeffries,R.J., Miller, J.R., Wharton, C. & Uyeda, K.M. User interface evaluation in the real world: A comparison of four techniques. Proceedings of CHI'91, (New Orleans), ACM, S. Robertson, G. Olson, & J. Olson , (Eds), Addison-Wesley, NY pp. 119-124, 1991.

Karat, C., Campbell, R.& Fiegel, T. Comparison of empirical testing and walkthrough methods in user interface evaluation. In Proceedings of CHI'92 (Monterey), ACM, 1992, 397-404.

Lewis, C., Polson, P., Wharton, C. and Rieman, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In Proceedings of CHI'90 (Seattle), ACM, Chew, J.C. & Whiteside, J. (Eds), Addison-Wesley, NY, 235-242, 1990.

Nielsen, J.& Molich, R. Heuristic evaluations of user interfaces. In Proceedings of CHI'90, (Seattle), ACM, Chew, J.C. and Whiteside, J. (Eds), Addison-Wesley, NY, 249-256, 1990.

Nielsen, J. Finding Usability Problems Through Heuristic Evaluation. In Proceedings of CHI, 1992 (Monterey) , ACM, 1992, 373-380.

Scapin, D.L. Organizing human factors knowledge for the evaluation and design of interfaces.International Journal of Human-Computer Interaction,203-229,2(3), 1990.

Smith, S.L. and Mosier, J.N. Guidelines for Designing User Interface Software. Report MTR-10090, The MITRE Corp., Bedford, MA, August 1986.

Thomas, J.C.& Kellogg, W.A. Minimizing ecological gaps in interface design. IEEE Software, 78-86, January, 1989.

Thomas, J.C. & Gould, J.D. A Psychological Study of Query by Example. Presented at May, 1975 National Computer Conference, Anaheim CA. AFIPS Conference Proceedings, 1975, 44, 439-445.

Thomas, J.C. A Method for Studying Natural Language Dialogue, IBM Research Report RC 5882, February, 1976.