Proposal for CHI '91

Work In Progress:

Empiricism Versus Judgement: Comparing User Interface Evaluation Methods on a New Telephone-Based Interface

Poster or Short Talk (Preferred Format:Poster)

Heather Desurvire, Debbie Lawrence, and Michael Atwood
NYNEX Science and Technology
500 Westchester Avenue
White Plains, N.Y. 10604
(914) 683-2757
hwd@nynexst.com
(please address correspondence to Heather Desurvire)

Keywords: Usability Analysis; Heuristic Evaluations; Telephone-Based Interface

Abstract

New Techniques for evaluating user interfaces are being developed by human factors researchers, in an attempt to reduce the time and cost of empirical usability testing. In heuristic evaluations (Nielsen and Molich, 1990) experts evaluate interfaces, using pre-established usability guidelines. The current study attempts to test the heuristic method of evaluation under several conditions and compare results to laboratory performance testing results. So far, the study has compared user laboratory performance on an interactive touch-tone telephone interface with heuristic evaluation by the same subjects. Two forthcoming conditions will be the heuristic evaluations of a paper specification of the interface by non-experts and experts. The study addresses: (1) whether heuristic evaluations and laboratory performance testing detect the same set of problems; and (2) whether users and two levels of experts produce similar heuristic evaluations with a live versus paper version of the system.

Work In Progress:

Empiricism Versus Judgement: Comparing User Interface Evaluation Methods on a New Telephone Based Interface

Introduction

Establishing new techniques for evaluating user interfaces is a growing concern in industry, due to the high cost of traditional laboratory testing of subjects. A few researchers have proposed alternative evaluation methods which are less costly than traditional performance testing, but their efficacy has not yet been thoroughly tested. In "heuristic evaluation" (Nielsen & Molich, 1990) experts evaluate interfaces, using established guidelines of usability. As an alternative to laboratory performance testing, the procedure would avoid the time and expense of running subjects. If the experts perform the procedure on paper specifications, early prototyping could also be avoided. We tested several different implementations of the heuristic technique during development of an interactive touch-tone telephone interface. Our main questions were: (1) would heuristic evaluations detect the same usability problems which emerged in laboratory performance testing, and what differences would occur; (2) how would heuristic evaluations by users compare to those of experts; and (3) how do heuristic evaluations of paper specifications versus live interfaces compare?

Experiment

The design includes (1) laboratory performance (task completion and error rates) by target users of the system (2) heuristic evaluation of the system by the same group, (3) heuristic evaluation of a paper specification of the interface by non-experts and (4) heuristic evaluation of the same paper specifications by human factors experts. The first two conditions have been run, and their data analysis is in process. The remaining two conditions will be conducted in February 1991. In the performance condition, fifteen subjects with no prior experience with the system were given a brief introduction and asked to use the system to perform a list of tasks. The experimenters collected the performance data as non-participant observers via videotape in a laboratory setting. There were two sessions for each subject. The same subjects then performed a heuristic evaluation of the interface. For each task, they rated the interface on 10 usability guidelines (Nielsen and Molich, 1990)[1] using a 10-point scale[2];

Discussion

The performance subjects' heuristic ratings identified as problems all tasks with low completion rates. However, the heuristic ratings also identified as problems two tasks which had high completion rates. These two tasks also had unreliable system performance. These subjects' heuristic ratings seem to reliably reflect their actual experience of where the system was hard to use. Their heuristic ratings reflected both design problems and system performance problems. The remaining conditions, using raters who have not used the system, will indicate how well heuristic ratings based on paper specifications can predict performance problems. Error rates as a measure of original subjects' performance will also be included in the analysis.

Further research could also look at the difference between heuristic ratings using a scale format, as in this study, vs. dichotomous ratings, (i.e., scoring a feature as "problem/not problem") and vs. a more open ended format used by Nielsen and Molich (1990).

References

Molich, R. and Nielson, J. Improving a human-computer dialog: What designers know about traditional interface design. *Communications of the ACM* 33, 3 (March, 1990).

[1] These guidelines originated from Smith and Mosier's (1986) 60 guidelines. We used the same 9 usability guidelines as Nielsen and Molich, and added one on documentation (i.e, "would or would not need documentation to complete task").

[2] This approach differed from Nielsen and Molich's approach in that we used a 10 point bipolar rating scale. Alternatively, Nielsen asked the user-evaluators to write a report on the problems with the usability guidelines kept in mind.

Nielson, J. and Molich, R. Heuristic evaluations of user interfaces. *CHI 1990 Proceedings*, 249-256.

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