Usability Testing vs. Heuristic Evaluation: Was there a contest?

ROBIN JEFFRIES AND HEATHER DESURVIRE†

ABSTRACT

Recent research comparing usability assessment methods has been interpreted by some to imply that usability testing is no longer necessary, because other techniques, such as heuristic evaluation, can find some usability problems more cost-effectively. Such an interpretation grossly overstates the actual results of the studies. In this article, we, as authors of studies that compared inspection methods to usability testing, point out the rather severe limitations to using inspection methods as a substitute for usability testing and argue for a more balanced repertoire of usability assessment techniques.

Communicating experimental results and interpreting their implications is often a more difficult task than one might expect. In HCI research, findings can be particularly vulnerable to misinterpretation. Since we have found misunderstandings of our results to be fairly pervasive, we have written this article to clarify our findings and their correct interpretations.

We, (Desurvire, Lawrence, and Atwood, 1991; Desurvire, Kondziela, and Atwood, 1992a, 1992b; Jeffries, Miller, Wharton, and Uyeda, 1991) and others (Karat, Campbell, and Fiegel, 1992) have been engaged in research that compares different usability inspection methods, (i.e., usability problem identification techniques which do not involve testing with potential users) to laboratory usability testing. Alternative methods such as heuristic evaluation (Nielsen and Molich, 1990), and cognitive walkthroughs (Polson, Lewis, Rieman, and Wharton, 1991) are intended to augment usability testing, either by being applicable early in the design cycle when usability testing is not possible, or as "discount methods" (Nielsen, 1989), used when resources (e.g., money, time, and trained evaluators) are scarce. Currently, practitioners have only their experience and intuition on which to base decisions about the suite of formal and informal evaluations to apply in a given situation; our goal was to provide data about the strengths and weaknesses of the various techniques which would help the practitioner make such decisions more effectively.

Our studies showed that heuristic evaluation can work very well. In Jeffries et al (1991), four expert heuristic evaluators found more problems than any other evaluation technique, including a usability test. In Desurvire et al. (1992a, 1992b), three expert heuristic evaluators also found more problems than any method and predicted about half of the problems found in a usability test. However, these results do not constitute a blanket endorsement of heuristic evaluation over usability testing. First, they all involved evaluators who were trained in usability issues; less knowledgeable evaluators performed poorly. Second, the studies used the aggregate results of multiple evaluations; a single heuristic evaluation was consistently the *least powerful* evaluation technique. Finally, the kinds of problems found by the different usability techniques were quite different. Heuristic evaluation missed half of the problems found in the laboratory, and usability testing missed about the same number of problems found in heuristic evaluation (Desurvire et al., 1992a, 1992b). In the Jeffries et al. study, usability testing exposed more severe problems, more

† This paper was a joint effort. Order of authorship was determined by the flip of a coin.

recurring problems and more global problems than did the heuristic evaluation.

Several people appear to have interpreted our results as "proof" that usability testing is a waste of time. In a trip report on the CHI '91 conference (HICOM, 1991), Nigel Bevan reports: "Two studies reported that this [heuristic evaluation] is a far more efficient and effective technique than other formal procedures used by designers, or usability testing by human factors experts." In his book "Tog on Interface" (Tognazzini, 1992), Bruce Tognazzini states (p. 59-60) "Human interface designers deliver the highest return on investment of any member of a software team... The Jeffries [Jeffries et al., 1991] study is exactly in line with what I have witnessed during my years in this business; having trained, qualified human interface people involved in a software effort produces a major benefit at a most reasonable cost." In public forums, Tognazzini has made even stronger statements, to the effect that the "myth of usability testing" has been put to rest (paraphrased from a talk to BayCHI, June 9, 1992). Other colleagues have told us informally of statements made by their co-workers along the lines of "We no longer need to do usability testing, because we can get better results from a heuristic evaluation."

We are very concerned about these interpretations of our data. The goal of our papers was to shed light on the relative contributions of these techniques, not to denigrate the rich useful data that comes from laboratory usability testing. Thus, we would like to reiterate and clarify our conclusions about the different roles and costs of usability testing and heuristic evaluation.

It is not particularly surprising that heuristic evaluation can be a valuable inspection method, since it is, in essence, applying the educated intuitions of multiple experts. However, one can reap the benefits of heuristic evaluation only within limited constraints. First, the evaluators must be experts. The Jeffries et al. study that produced an advantage for heuristic evaluation used only experts. Desurvire et al. (1991, 1992a, 1992b) had three levels of evaluator expertise, and only the usability experts found anywhere near as many problems as usability testing. In fact, results from non-experts (people who were neither usability specialists nor software engineers) were found to be unreliable. Nielsen (1992) showed that experts do much better than software engineers on heuristic evaluations, with the greatest impact coming from "double experts" – people who were expert both on usability issues generally and on the domain of the application specifically. Useful results can be obtained with engineers doing heuristic evaluations with minimal training, but only if large numbers (i.e., ten to twenty) of independent evaluations are done. This would make sense when usability experts are the most scarce resource; while the large number of evaluations needed for effective coverage can be expensive, the data suggest that a useful fraction of usability problems could be identified in this manner.

The second caveat to the use of heuristic evaluation is that multiple expert evaluations are needed to produce the results seen in our studies – 3-5 evaluators at a minimum. A single heuristic evaluation is consistently the weakest way to evaluate an interface, in all the published studies. Such one-person evaluations may have their place – e.g., as an early way to weed out especially egregious problems – but to imagine that this could substitute for even a simple usability test is completely inconsistent with a growing mass of data. Few organizations will have access to the number of experts needed to do this sort of heuristic evaluation effectively; in fact, this limitation suggests that heuristic evaluation, as a substitute for usability testing, is more of a "Cadillac method" than a "discount method". Organizations who are looking for ways to focus on usability within tight resource constraints should definitely not concentrate all their efforts on expert heuristic evaluation.

The third concern about heuristic evaluation is that we must not neglect the associated costs of the method. These costs seem primarily to be in the stage after usability issues are identified. In the Jeffries et al. study a large number of the problems identified by heuristic evaluation are minor problems; many of these are also matters of "taste" – different HCI experts might disagree about the validity or value of the recommendations. The experts in Desurvire et al. (1992a, 1992b) only found about a third the most severe problems identified in the laboratory; in contrast, they found two-thirds of the laboratory problems of moderate severity, and 4/5 of the least severe problems. If the developers are given this large volume of minor problems to sort through, do the costs incurred at their end – the need to allocate resources among problems of different priority and the errors made by not focusing resources on the problems that will have the greatest user impact – outweigh the advantage of lower up-front costs in doing the evaluation? Even more problematic is the possibility that some of the issues identified by heuristic evaluation might be false alarms – non-problems whose correction could make the application less usable. Further research is needed on defining and determining the frequency of false alarms, but we cannot rule out their existence.

We also want to point out some of the hidden benefits of usability testing. Most obviously, a usability test identifies problems that will plague the actual users of the application. There is no need to sort or filter the problems according to their predicted impact on users; the impact can be assessed from the test. This was born out by all of our studies, where almost all the problems identified by the usability test were above the median in severity. In addition, as anyone who has run a usability test can confirm, data from users has an impact on the engineers developing the product that no "expert evaluation" can equal. Developers may doubt that a problem in the user-interface exists, but when they see the user actually experience that problem in the laboratory, they change their minds quickly. Finally, some problems found in usability tests are highly unlikely to be discovered by other methods; the ingenuity of people using a novel application greatly exceeds the imaginations of experts. For example, Jeffries et al. mention that the usability test in their study exposed a problem that, by a perfectly reasonable sequence of actions – i.e., discarding an apparently "unneeded" file – a user made it impossible to log back into the system. No other technique identified that problem, and, unless the evaluator had previous experience with very similar problems, it seems unlikely that any non-user-involved method would.

Probably the most pernicious aspect of the various misinter-

pretations we have encountered is the assumption that one should select a single method of usability evaluation. Our studies and other research have consistently shown that different methods have various strengths; the best evaluation of a user-interface comes from applying multiple evaluation techniques. The various techniques have differing constraints on their applicability and on the resources required to apply them effectively. Usability testing and heuristic evaluation require access to expert evaluators; in the case of heuristic evaluation, a group of them is required. If you have access to multiple such experts, then certainly doing both heuristic evaluations and usability testing will be better than doing one alone.

Other inspection methods have a role to play too. Both cognitive walkthroughs and usability guidelines can be used earlier in the development process than heuristic evaluation[1] and well before usability testing. The advantages of finding problems very early in the design process outweigh the rather modest number of problems guidelines or the cognitive walkthrough techniques identified in the comparison studies by Jeffries et al. (1992) and Desurvire et al. (1991, 1992a, 1992b). One must add to that the modest cost of the techniques, since they can be applied by the developers themselves, and they typically take no more than a couple of days to conduct. In many cases, it makes a lot of sense to include one or more of these techniques in the evaluation repertoire. In their enthusiasm to embrace heuristic evaluation as an "officially sanctioned" technique, people seem to have focused only on the advantages of the technique (i.e., it's fast, it's cheap, it finds a lot of problems) and not on either the disadvantages (i.e., it requires multiple evaluations; its works best with experts; it finds a distressing number of minor problems) or the complementary advantages of usability testing (i.e., it overwhelmingly finds severe problems; it finds problems that impact real users). Our goal here was to realign that pendulum where it belongs: all else being equal, a usability test will provide the highest quality assessment of an application. Usability testing has its disadvantages also, the primary ones being cost and that it can only be applied late in the development cycle. Usability inspection methods were developed to be used in circumstances where usability testing is impractical; they fill that niche very well. But if we begin to use inspection techniques to the exclusion of usability testing, we will have lost one of our most valuable tools for evaluation.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions of our coauthors to the original Desurvire et al. and Jeffries et al. papers: Mike Atwood, Jim Kondziela, Jim Miller, Kathy Uyeda, and Cathleen Wharton. Conversations with them greatly clarified and expanded our thinking.

REFERENCES

Desurvire, H., Kondziela, J., and Atwood, M. (1992a; short paper version). What is Gained and Lost When Using Evaluation Methods Other Than Empirical Testing. A short talk presented at CHI '92 (Monterey, California, May 3-7, 1992), ACM, collection of abstracts, pp. 125-126.

Desurvire, H., Kondziela, J., and Atwood, M. (1992b; full paper version). What is Gained and Lost When Using Evaluation Methods Other Than Empirical Testing. In the proceedings of HCI 1992, Cambridge University Press, edited by Monk, A., Diaper, D., and Harrison, M.D., (University of York, U.K., September 15-18, 1992).

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

Jeffries, R., Miller, J. R., Wharton, C., and Uyeda, K. M. User interface evaluation in the real world: A comparison of four techniques. Proceedings of CHI, 1991 (New Orleans, Louisiana, April 28 - May 2, 1991) ACM, New York, 1991. pp. 119-124.

Karat C., Campbell R., and T. Fiegel. Comparison of empirical testing and walkthrough methods in user interface evaluation. In Proceedings of CHI '92 (Monterey, California, May 3-7, 1992) ACM, New York, 1992. pp. 397-404.

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

Molich R., and J. Nielsen. Improving a human-computer dialogue. Communications of the ACM, 33,3, pp. 338-342, March 1990.

Nielsen J. Finding Usability Problems Through Heuristic Evaluation. In Proceedings of CHI '92 (Monterey), ACM, 1992, pp. 373-380.

Nielsen J. Usability Engineering At a Discount. In Salvendy, G., and Smith, M.J., (Eds.), Designing and Using Human-Computer Interfaces and Knowledge-Based Systems. Elsevier Science Publishers, Amsterdam, pp. 394-401, 1989.

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

1. Heuristic evaluation can in principle be done with partially specified prototypes also, but that is a very different sort of evaluation. We believe it would be inappropriate to extend our results to such evaluations. All the studies of heuristic evaluation to date have used either working systems or fully defined paper prototypes. In our experience, early prototypes underspecify things that are critical to usability, and the evaluators tend to give the developers the benefit of the doubt in those cases. Additional research needs to be done comparing heuristic evaluation of early and late prototypes.

SIGCHI bulletin

A PUBLICATION OF ACM/SIGCHI
Special Interest Group on Computer & Human Interaction

October 1992

Voting Members:

Please review the TWO proposed changes in Bylaws and the instructions for responding that are included in the plastic envelope containing this issue of the *SIGCHI Bulletin*.

Your ballots must be received by January 2, 1993.

VOLUME 24, NUMBER 4
ISSN 0736-6906

```
--MORE--(102)
Do you really want to mark everything as read? [yn]


End of newsgroup comp.dcom.telecon.

********  3 unread articles in comp.human-factors--read now? [ynq]

Article 3815 (2 more) in comp.human-factors:
From: thinman@netcom.com (Technically Sweet)
Subject: Re: Gnu I/F (Self-Destructive?)
Message-ID: <1992Dec17.182817.27173@netcom.com>
Organization: International Foundation for Internal Freedom
References: <1992Dec15.231336.4441@colorado.edu> <1992Dec16.153052.1003@xgml.com >
Date: Thu, 17 Dec 1992 18:28:17 GMT
Lines: 18

--MORE--(44%)

Article 3816 (1 more) in comp.human-factors:
From: asper@sbctri.sbc.com (Alan E. Asper)
Subject: Re: Usability Engineering book - comments needed on manuscript
Message-ID: <1992Dec17.184243.3173@sbctri.sbc.com>
Organization: Southwestern Bell Technology Resources, St.Louis, MO
References: <1992Dec10.170634.28877@walter.bellcore.com>
Date: Thu, 17 Dec 92 18:42:43 GMT
Lines: 13

In article <1992Dec10.170634.28877@walter.bellcore.com> Jakob Nielsen <nielsen@b ellcore.com> writes:
>I need several software developers and software project managers to comment
>on the manuscript for my new book, "Usability Engineering".
>

I urge anyone who reviews the manuscript to also read Jeffries & Desurvire's
article "Usability Testing vs. Heuristic Evaluation: Was there a contest?" in
the October 1992 issue of the ACM-SIGCHI Bulletin. It's a good article, and
provides some much-needed perspective on some of the methodologies described
in Nielsen's manuscript outline.

Alan

End of article 3816 (of 3817)--what next? [npq]

Article 3817 in comp.human-factors:
Newsgroups: ba.seminars,comp.human-factors,comp.cog-eng,comp.groupware
From: riander@well.sf.ca.us (Richard Ivan Anderson)
Subject: BayCHI (Jan 12) - ACTORS, AGENTS, AND SHOW BIZ - Spoonman
Message-ID: <BzFK38.5I1@well.sf.ca.us>
Sender: news@well.sf.ca.us
Organization: Whole Earth 'Lectronic Link
Distribution: ba
Date: Fri, 18 Dec 1992 00:59:35 GMT
Lines: 138

--MORE--(7%)
Terminal 1 (TELNET) -- TEXAS on INTERNET [Ambassador]
```

it's another FYI
- mike