INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *13*(4), 343–349

Introduction: Current Issues in Usability Evaluation

James R. Lewis
IBM Corporation

In this introduction to the special issue of the *International Journal of Human–Computer Interaction*, I discuss some current topics in usability evaluation and indicate how the contributions to the issue relate to these topics. The contributions cover a wide range of topics in usability evaluation, including a discussion of usability science, how to evaluate usability evaluation methods, the effect and control of certain biases in the selection of evaluative tasks, a lack of reliability in problem detection across evaluators, how to adjust estimates of problem-discovery rates computed from small samples, and the effects of perception of hedonic and ergonomic quality on user ratings of a product's appeal.

1. INTRODUCTION

1.1. Acknowledgements

I start by thanking Gavriel Salvendy for the opportunity to edit this special issue on usability. I have never done anything quite like this before, and it was both more difficult and more rewarding than I expected. The most rewarding aspect was working with the contributing authors and reviewers who volunteered their time and effort to the contents of this issue. The reviewers were Robert Mack (IBM T. J. Watson Research Center), Richard Cordes (IBM Human Factors Raleigh), and Gerard Hollemans (Philips Research Laboratories Eindhoven). Several of the contributing authors have indicated to me their appreciation for the quality of the reviews that they received. To this I add my sincere appreciation. Providing a comprehensive, critical review of an article is demanding work that advances the literature but provides almost no personal benefit to the reviewer (who typically remains anonymous). Doing this work is truly the mark of the dedicated, selfless professional.

1.2. State of the Art in Usability Evaluation

I joined IBM as a usability practitioner in 1981 after getting a master's degree in engineering psychology. At that time, the standard practice in our product develop-

Send requests for reprints to James R. Lewis, IBM Corporation, 8051 Congress Avenue, Suite 2227, Boca Raton, FL 33487. E-mail: jimlewis@us.ibm.com

ment laboratory was to conduct scenario-based usability studies for diagnosing and correcting usability problems. Part of the process of setting up such a study was for one or more of the usability professionals in the laboratory to examine the product (its hardware, software, and documentation) to identify potential problem areas and to develop scenarios for the usability study. Any obvious usability problems uncovered during this initial step went to Development for correction before running the usability study.

Alphonse Chapanis had consulted with IBM on the use of this rather straightforward approach to usability evaluation. Regarding an appropriate sample size, he recommended 6 participants because experience had shown that after watching about 6 people perform a set of tasks, it was relatively rare to observe the occurrence of additional important usability problems (Chapanis, personal communication, 1979).

A clear problem with this approach was that to run the usability study you needed a working version of the product. Since I first started as a usability practitioner, there have been two important lines of usability research that have affected the field. One is the development of additional empirical and rational usability evaluation methods (e.g., think aloud, heuristic evaluation, cognitive walkthrough, GOMS, QUIS, SUMI), many of which were motivated by the need to start initial usability evaluation earlier in the development cycle. The other is the comparative evaluation of these usability methods (their reliability, validity, and relative efficiency). The last 20 years have certainly seen the introduction of more usability evaluation tools for the practitioner's toolbox and some consensus (and still some debate) on the conditions under which to use the various tools. We have mathematical models for the estimation of appropriate sample sizes for problem-discovery usability evaluations (rather than a simple appeal to experience; Lewis, 1982, 1994; Nielsen & Landauer, 1993; Virzi, 1990, 1992). Within the last 12 years usability researchers have published a variety of usability questionnaires with documented psychometric properties (e.g., Chin, Diehl, & Norman, 1988; Kirakowski & Corbett, 1993; Lewis, 1995, 1999).

Yet many questions remain. For example

- Do we really understand the differences among the various usability evaluation methods in common use by practitioners? Do we have a good idea about how to compare usability evaluation methods?
- How do the tasks selected for a scenario-based usability evaluation affect the outcome of the evaluation? How do we decide what tasks to include in an evaluation and what tasks to exclude?
- How does the implicit assumption that we only ask participants to do tasks that are possible with a system affect their performance in a usability evaluation?
- Usability evaluation based on the observation of participants completing tasks with a system is the golden standard for usability evaluation, with other approaches generally considered discounted by one criterion or another. Usability practitioners assume that the same usability problems uncovered by one laboratory would, for the most part, be discovered if evaluated in another laboratory. To what extent do they know that this assumption is true? What are the implications for the state of the art if the assumption is not true?

- Sample size estimation for usability evaluations depends on having an estimate of the rate of problem discovery (p) across participants (or, in the case of heuristic evaluation, evaluators). It turns out, though, that estimates of this rate based on small-sample usability studies are necessarily inflated (Hertzum & Jacobsen, this issue). Is there any way to adjust for this bias so practitioners can use small-sample estimates of p to develop realistic estimates of the true problem-discovery rate and thereby estimate accurate sample size requirements for their studies?
- Is usability enough to assure the success of commercial products in a competitive marketplace? To what extent is "likeability" or "appealingness" affected by or independent of usability? How can we begin to measure this attribute?
- We feel like we know one when we see one, but what is the real definition of a *usability problem?* What is the appropriate level at which to record usability problems?

2. CONTRIBUTIONS TO THIS ISSUE

The contributions to this issue do not answer all of these questions, but they do address a substantial number of them.

2.1. "Usability Science. I: Foundations"

In this first article of the issue, Gillan and Bias describe the emerging discipline of usability science. The development and qualification of methods for usability design and evaluation require a scientific approach, yet they may well be distinct from other similar disciplines such as human factors engineering or human–computer interaction. In their article, Gillan and Bias reach across various modern disciplines and into the history of psychology to develop their arguments.

2.2. "Criteria for Evaluating Usability Evaluation Methods"

Hartson, Andre, and Williges (this issue) tackle the problem of how to compare usability evaluation methods. The presence of their article is additional testimony to the importance of the recent critique by Gray and Salzman (1998) on potentially misleading research ("damaged goods") in the current literature of studies that compare usability evaluation methods. Developing reliable and valid comparisons of usability evaluation methods is a far from trivial problem, and Hartson et al. lay out a number of fundamental issues on how to measure and compare the outputs of different types of usability evaluation methods.

2.3. "Task-Selection Bias: A Case for User-Defined Tasks"

Cordes (this issue) provides evidence that participants in laboratory-based usability evaluations assume that the tasks that evaluators ask them to perform must be possible and that manipulations to bring this assumption into doubt have dramatic effects on a study’s quantitative usability measures (a strong bias of which many usability practitioners are very likely unaware). Cordes discusses how introducing user-defined tasks into an evaluation can help control for this bias and presents techniques for dealing with the methodological and practical consequences of including user-defined tasks in a usability evaluation.

2.4. *“Evaluator Effect: A Chilling Fact About Usability Evaluation Methods”*

This title of the article by Hertzum and Jacobsen (this issue) might seem a bit extreme, but the evidence they present is indeed chilling. Their research indicates that the most widely used usability evaluation methods suffer from a substantial evaluator effect—that the set of usability problems uncovered by one observer often bears little resemblance to the sets described by other observers evaluating the same interface. They discuss the conditions that affect the magnitude of the evaluator effect and provide recommendations for reducing it.

For me, this is the most disturbing article in the issue, in part because other investigators have recently reported similar findings (Molich et al., 1998). The possibility that usability practitioners might be engaging in self-deception regarding the reliability of their problem-discovery methods is reminiscent of clinical psychologists who apply untested evaluative techniques (such as projective tests), continuing to have faith in their methods despite experimental evidence to the contrary (Lilienfeld, Wood, & Garb, 2000). Although we might not like the results of Hertzum and Jacobsen (this issue), we need to understand them and their implications for how we do what we do as usability practitioners.

Often usability practitioners only have a single opportunity to evaluate an interface, so there is no way to determine if their usability interventions have really improved an interface. In my own experience though, when I have conducted a standard scenario-based, problem-discovery usability evaluation with one observer watching multiple participants complete tasks with an interface and have done so in an iterative fashion, the measurements across iterations consistently indicate a substantial and statistically reliable improvement in usability. This leads me to believe that, despite the potential existence of a substantial evaluator effect, the application of usability evaluation methods (at least, methods that involve the observation of participants performing tasks with a product under development) can result in improved usability (e.g., see Lewis, 1996). An important task for future research in the evaluator effect will be to reconcile this effect with the apparent reality of usability improvement achieved through iterative application of usability evaluation methods.

2.5. *“Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples”*

For many years I have promoted the measurement of *p* rates from usability studies for the dual purpose of (a) projecting required sample sizes and (b) estimating the proportion of discovered problems for a given sample size (Lewis, 1982, 1994, 2001).

I have always made the implicit assumption that values of *p* estimated from small samples would have properties similar to those of a mean—that the variance would be greater than for studies with larger sample sizes, but in the long run estimates of *p* would be unbiased. I was really surprised (make that appalled) when Hertzum and Jacobsen (this issue) demonstrated in their article that estimates of *p* based on small samples are almost always inflated. The consequence of this is that practitioners who use small-sample estimates of *p* to assess their progress when running a usability study will think they are doing much better than they really are. Practitioners who use small-sample estimates of *p* to project required sample sizes for a usability study will seriously underestimate the true sample size requirement.

This spurred me to investigate whether there were any procedures that could reliably compensate for the small-sample inflation of *p*. If not, then it would be important for practitioners to become aware of this limitation and to stop using small-sample estimates of *p*. If so, then it would be important for practitioners to begin using the appropriate adjustment procedure(s) to ensure accurate assessment of sample size requirements and proportions of discovered problems. Fortunately, techniques based on observations by Hertzum and Jacobsen (this issue) and a discounting method borrowed from statistical language modeling can produce very accurate adjustments of *p*.

2.6. *"Effect of Perceived Hedonic Quality on Product Appealingness"*

Within the IBM User-Centered Design community and outside of IBM (e.g., see Tractinsky, Katz, & Ikar, 2000), there has been a growing emphasis over the last few years to extend user-centered design beyond traditional usability issues and to address the total user experience. One factor that has inhibited this activity is the paucity of instruments for assessing nontraditional aspects of users' emotional responses to products. Hassenzahl (this issue) has started a line of research in which he uses semantic differentials to measure both ergonomic and hedonic quality and relates these measurements to the appealingness of a product. Although there is still a lot of work to do to validate these measurements, it is a promising start that should be of interest to practitioners who have an interest in the total user experience.

3. *CONCLUSIONS*

I hope this special issue will be of interest to both usability scientists and practitioners. The contributors are from both research (Gillan, Hartson, Andre, Williges, and Hertzum) and applied (Bias, Cordes, Jacobsen, Lewis, and Hassenzahl) settings, with two of the articles collaborations between the settings (Gillan & Bias, this issue; Hertzum & Jacobsen, this issue).

The articles in this special issue address many of the topics listed in Section 1.2 (although there is still much work to do). One important outstanding issue, though, is the development of a definition of what constitutes a real usability problem with which a broad base of usability scientists and practitioners can agree. This is a topic that comes up in half of the articles in this issue (Hartson et al.; Hertzum & Jacobsen; Lewis) but is one for which I have not yet seen a truly satisfactory treatment (but see the following for some current work in this area: Cockton & Lavery, 1999; Connell & Hammond, 1999; Hassenzahl, 2000; Lavery, Cockton, & Atkinson, 1997; Lee, 1998; Virzi, Sokolov, & Karis, 1996).

Despite the unanswered questions, I believe that the field of usability engineering is in much better shape than it was 20 years ago (both methodologically and with regard to the respect of product developers for usability practitioners), and I look forward to seeing and participating in the developments that will occur over the next 20 years. I also look forward to seeing the effect (if any) that the articles published in this special issue will have on the course of future usability research and practice.

REFERENCES

Chin, J. P., Diehl, V. A., & Norman, L. K. (1988). Development of an instrument measuring user satisfaction of the human–computer interface. In *Conference Proceedings of Human Factors in Computing Systems CHI '88* (pp. 213–218). Washington, DC: Association for Computing Machinery.

Cockton, G., & Lavery, D. (1999). A framework for usability problem extraction. In *Human–Computer Interaction—INTERACT '99* (pp. 344–352). Amsterdam: IOS Press.

Connell, I. W., & Hammond, N. V. (1999). Comparing usability evaluation principles with heuristics: Problem instances vs. problem types. In *Human–Computer Interaction—INTERACT '99* (pp. 621–629). Amsterdam: IOS Press.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction, 13,* 203–261.

Hassenzahl, M. (2000). Prioritizing usability problems: Data-driven and judgement-driven severity estimates. *Behaviour and Information Technology, 19,* 29–42.

Kirakowski, J., & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology, 24,* 210–212.

Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour and Information Technology, 16,* 246–266.

Lee, W. O. (1998). Analysis of problems found in user testing using an approximate model of user action. In *People and Computers XIII: Proceedings of HCI '98* (pp. 23–35). Sheffield, England: Springer-Verlag.

Lewis, J. R. (1982). Testing small-system customer set-up. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 718–720). Santa Monica, CA: Human Factors Society.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors, 36,* 368–378.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction, 7,* 57–78.

Lewis, J. R. (1996). Reaping the benefits of modern usability evaluation: The Simon story. In A. F. Ozok & G. Salvendy (Eds.), *Advances in applied ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics* (pp. 752–757). Istanbul, Turkey: USA Publishing.

Lewis, J. R. (1999). Trade-offs in the design of the IBM computer usability satisfaction questionnaires. In H. Bullinger & J. Ziegler (Eds.), *Human–computer interaction: Ergonomics and user interfaces—Vol. 1* (pp. 1023–1027). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lewis, J. R. (2001). *Sample size estimation and use of substitute audiences* (Tech. Rep. No. 29.3385). Raleigh, NC: IBM. Available from the author.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., & Kirakowski, J. (1998). Comparative evaluation of usability tests. In *Usability Professionals Association Annual Conference Proceedings* (pp. 189–200). Washington, DC: Usability Professionals Association.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Conference Proceedings on Human Factors in Computing Systems—CHI '93* (pp. 206–213). New York: Association for Computing Machinery.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting With Computers, 13,* 127–145.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 291–294). Santa Monica, CA: Human Factors Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34,* 443–451.

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). Usability problem identification using both low- and high-fidelity prototypes. In *Proceedings on Human Factors in Computing Systems CHI '96* (pp. 236–243). New York: Association for Computing Machinery.