REPAIRING DAMAGED MERCHANDISE: A REJOINDER

Wayne D. Gray & Marilyn C. Salzman
Human Factors & Applied Cognition
George Mason University
Fairfax, VA 22030

VERSION ACCEPTED BY JOURNAL
PUBLISHED VERSION IS SLIGHTLY DIFFERENT

Gray, W. D., & Salzman, M. C. (1998). Repairing damaged merchandise: A rejoinder. *Human-Computer Interaction, 13*(3), 325-335.

Send correspondence to:
Wayne D. Gray
George Mason University
m/s 3f5
Fairfax, VA 22030
(703) 993-1357
gray@gmu.edu

REPAIRING DAMAGED MERCHANDISE: A REJOINDER

Wayne D. Gray & Marilyn C. Salzman

Wayne D. Gray is a cognitive scientist with an interest in methodology as well as in how artifact design affects the cognition required to perform tasks. He has worked in government and industry; he currently heads the Human Factors and Applied Cognitive Program at George Mason University. Marilyn C. Salzman has worked as a usability engineer for industry, currently she is a doctoral student in the Human Factors and Applied Cognitive Program at George Mason University. Her interests include human-computer interaction design and evaluation.

1. Introduction

Our goal in writing DM[1] was not to have the last word on the subject but to raise an awareness within the human-computer interaction (HCI) community of issues that we felt had been too long ignored or neglected. Upon reading the ten commentaries from distinguished members of the HCI community, we were pleased to see that they had joined the debate and broadened the discussion. Subsequently, we were somewhat torn by how to proceed. Our first thought was to respond point-by-point, paper-by-paper. However, we refrain from addressing many specific issues here as a full discussion would involve a paper at least as long as DM. Instead we focus on a few important themes that emerged throughout our paper and the ensuing discussion:

- What is usability, how do we measure it, and what do we need to know about our usability evaluation methods (UEMs)?
- Why do we find ourselves where we are?
- What is the role of experiments vs other empirical studies in HCI? Are there common issues in the design of empirical studies?
- How to judge the value of a study?
- Where do we go from here?

2. What Is Usability, How Do We Measure It, And What Do We Need To Know About UEMs?

After completing our review of validity problems with the UEM studies (section 5 of DM), we began our Observations and Recommendations (section 6) by stating our belief that "the most important issue facing usability researchers and practitioners alike [is] the construct of usability itself" (pp. 238). We argued that we need to broaden our definition of usability so that it goes beyond the problem-counting approach epitomized by the papers reviewed in DM. Among our commentators, Mackay, McClelland, Monk, Oviatt, and Newman took up and elaborated this issue.

Mackay urges the use of multiple measures to triangulate upon what is meant by usability. McClelland points out that "usability is not one thing" and that it is an evolving construct. In addition to traditional usability issues (objective performance, subjective impressions, safety, and learning), we may also need to consider factors such as pleasure of use when evaluating the strengths and weaknesses of our UEMs. Oviatt presents a similar argument, encouraging the HCI community to consider more than our traditional measures of usability when examining UEMs and to rely on triangulation to pinpoint each UEM's strengths and weaknesses.

Newman's commentary touches upon a number of key points that deserve careful consideration. First, he faults us for a sin of omission; namely finding "no fault with the results of user testing" and ties this into an excellent discussion of the type of information a designer needs to have to improve a design. Although we chose not to focus on the strengths and weaknesses of user testing in DM, it was clear to us that how user testing was implemented and how its outcomes were interpreted varied greatly among the five studies reviewed. In fact, in section 6.1 we warned that deciding how or what to change in an interface based upon problems identified during user testing is not simple and that the "problems-to-features" mapping "cannot be assumed and the links must be carefully forged." As Monk points out, this is an issue for which experimentation can be particularly useful. Another useful mechanism for accomplishing this goal may be simulation, an argument made by both Carroll and Newman.

Second, we believe Newman is correct in asserting that designers need more than measures of "overall performance" as they are seldom truly interested in the tasks performed in the usability lab. Designers are not interested in "the performance of tasks" but in how the interface affects that performance. "If a particular task is performed conspicuously slowly, the designer needs to know that the slow appearance of a dialogue box is a major contributor" (Newman). This sentiment echoes John and Mashyna's (1997) concerns with attempts to classify usability problems into a small number of categories. They maintain that developers need to know the specific problem (e.g., a problem with an item in a particular menu) and not the general one (e.g., "there are menu problems" or "speak the users' language").

Third, we endorse Newman's argument that we need to go beyond "piecewise approaches to evaluation" to identify "the critical parameters of interactive applications." We believe that one way to do so is to combine user testing with careful, fine grained task analyses such as those that can be accomplished via GOMS (John and Kieras, 1996a; b). An excellent example of this is Franzke's study of the use characteristics of different software packages (Franzke, 1994; Franzke, 1995). Although such studies are currently painstaking to perform, a large part of the pain is caused by the absence of off-the-shelf tools to automate the process of putting user data into correspondence with the various components of a fined grained procedural task analysis.

Perhaps this is an area in which the tool-building component of the HCI community can come to the aid of those interested in measurement and methodology.

3. Why Do We Find Ourselves Where We Are?

In exploring the question of why some of the best industry laboratories in our field could have produced research with the flaws delineated in DM, our commentators seem to be of two minds. The first camp argues, as do Jeffries and Miller, that valid research is not interesting research. They pose a dichotomy between doing valid experiments versus experiments that have *ecological validity*. The second camp (which is well represented by Lund and Mackay) argues that research is valued within corporations based upon its *face validity* rather than its methodological purity. Thus, as Lund states "it is almost axiomatic, therefore, that much corporate research undertaken in this environment will have limited validity." When practitioners review conference and journal submissions, they are heavily influenced by the values of the corporate culture within which they work and look for "content that *appears* to provide value" (italics added). Mackay sees this problem as exacerbated in HCI since it is a field that draws upon people with diverse backgrounds, where experts in one area may lack training in other areas such as experimental design. Whereas diversity is a strength of the field, aligning reviewer expertise with the conceptual, methodological, and statistical issues raised by various research papers is a challenging endeavor.

We believe the distinction between *face validity* and *ecological validity* is important. The former *appears* good, the latter *is* good; the former *is* easy, the latter *is* hard; and, unfortunately, *the former is often mistaken for the latter*. When empirical techniques, experimental or otherwise, are applied to address difficult real-world problems, it is much easier to do something that *appears* good, rather than something that *is* good.

This tendency towards face validity may be exacerbated (as suggested by Lund) by the absence of a focus on methodological issues within the HCI-oriented, academic research community. He points out that "the absence of comparative academic research in this area is noticeable and distressing" and surmises that gatekeeping with respect to the conference review process may have failed in part due to the "relatively limited academic interest in the topics of methodologies."

4. What Is The Role Of Experiments Vs Other Empirical Studies In HCI? Are There Common Issues In The Design Of Empirical Studies?

DM focused on experiments and problems with experimental design for the simple reason that the five studies we reviewed were cast by their authors as experimental studies. It was

problems with the use of the experimental method in these studies and their acceptance by the HCI community that motivated DM. DM should not be interpreted either as advocating experiments to the exclusion of other empirical techniques, or as maintaining that other techniques do not need to be concerned with the types of validity discussed in DM.

Although experiments were our focus, throughout DM we included small statements of our regard for other empirical methods. For example, our conclusion section begins, "The multitude of empirical methodologies is a strength of the HCI research community" and in section 3 we acknowledge the "plethora of alternatives" to experiments. Apparently, these small statements were not enough. In her commentatory, John expresses her fear that our "unremitting focus on problems with experimental design" might make it easy for some readers to conclude that we believe that all HCI research must be cast as experiments. Unfortunately, this "unremitting focus" has apparently mislead Carroll who states that we advocate experiments as "the royal road to causal analysis." We do not.

As Cook and Campbell discuss in their chapter two, experiments enable the researcher to draw strong inferences concerning causality. (Indeed, we suspect that this is why so many researchers attempt to cast their work in experimental form despite the plethora of alternatives.) This power comes at a price -- experiments do not degrade gracefully. Although there may be a continuum of experimental designs ranging from well-designed to poorly designed, there is not a concomitant continuum of inferences, from strong to weak, that can be drawn from such designs. At some point along the design continuum, valid inferences no longer can be made. We need to value not only the strength of the inferences that an experiment permits, but to understand that small problems of experimental design can have large effects on what we can legitimately conclude. Flawed experiments simply do not enable us to draw valid inferences.

When circumstances do not permit a valid experiment to be designed, we urge the reader to consider alternatives to experiments. One alternative is to do as Mackay recommends -- to go beyond chapter 2 to the core of Cook and Campbell's excellent book. Chapters 3 and beyond are devoted to *quasi-experimental design*: the detailed consideration of how to design reliable and valid empirical studies when valid experimental designs are not an option. We urge the reader not to be bothered, as Carroll apparently is, by Cook and Campbell's "quirky language." These researchers were not writing for experimental psychologists or for computer scientists. They were addressing the public policy oriented program evaluation community (i.e., large social and education programs, not computer programs). It would be a mistake for the discipline of HCI to ignore the lessons embodied in that book, simply because its language comes from another discipline.

Other alternatives to experimental design mentioned by our commentators include ethnographic studies, simulation studies, case studies, diary studies, critical incident techniques, task analytic methods, and verbal protocol analyses. Such alternatives can be used to study issues and relationships that cannot easily be brought under experimental control. For example, if our goal is to understand how UEMs are applied in practice, some of the alternatives (e.g., ethnographic studies, case studies, etc.) cited above might be more effective than experimentation. If our goal is to better define the usability construct, simulation studies and task analytic methods might be particularly useful. Finally, if our goal is to understand how successful UEMs are at discovering rare but critical events, we might turn to the critical incident technique, which is a form of task analysis (Kirwan & Ainsworth, 1992).

Importantly, in turning to non-experimental forms of empirical inquiry it is necessary to remember that these are not simply sloppy experiments but have their own requirements of methodological rigor. For example, John rightly criticizes our statement that:

> If Jeffries, et al. (1991) had been cast as a case study (and appropriate changes made throughout), the paper would have provided a snapshot of the trade-offs facing Hewlett-Packard in deciding how to do usability analyses in the late 1980s.

Our assumption that re-casting this paper as a case study would alleviate the validity issues we raised was incorrect. As John's comment suggests, even techniques that do not lend themselves to rigorous statistical analysis must face concerns with internal validity, construct validity, and external validity.[2]

In choosing an empirical technique, experimental or otherwise, several factors need to be carefully considered. First, different empirical techniques support stronger or weaker inferences regarding causality. Second, there is no best empirical technique, only a best technique given the question and circumstances. Third, there may well be a trade-off between power and the breadth of questions that are asked so that techniques such as case study or verbal protocol analysis may allow weaker conclusions to be made about a wider range of issues than a comparable experimental study. Fourth, our statements about the strength of causal inferences that various techniques permit, apply only to single studies. A well-designed series of, for example, case studies, may well support causal inferences as strong as any permitted by experimentation. Fifth, along with all of our commentators who took a stance, we believe in empirical pluralism or triangulation. Difficult questions, especially those confronting HCI, are unlikely to be adequately addressed using any one technique. Sixth, triangulation involves more than simply using different empirical techniques to study the same issue. It requires that the results from these divergent empirical approaches be combined and integrated into powerful analytic frameworks. Such frameworks will provide a source for practitioners to draw upon, as well as, clearly defining the

gaps and issues that researchers should explore. Undoubtedly, the resulting frameworks will be as diverse as the needs of the practitioner and researcher communities. However, if done well, the frameworks will be partially overlapping and mutually referencing. For issues concerning individual cognition, emerging frameworks can be found in the form of cognitive architectures and their use by the HCI research community in computational cognitive modeling (see, e.g., Gray, Young & Kirschenbaum, 1997).

5. How To Judge The Value Of A Study?

Many of our commentators weigh in on the issue of how to judge the value of a study. Several take a narrow perspective, focusing on the problems involved in gaining access to large numbers of usability specialists or software engineers. A pair of commentators complain that DM misjudged the value of the study they had conducted. Finally, some of our commentators chose to comment upon the place of individual studies in the context of the larger research enterprise. In this section, we briefly comment upon the first two approaches and elaborate upon the third.

A narrow focus on statistical conclusion validity led several of our commentators to discuss the difficulty of obtaining large samples of the type of subjects needed for UEM research. Although gaining access to the right type of subjects is important, valid and useful results can be achieved with a limited subject pool. The most serious problems in inferring cause-and-effect in the studies we reviewed did not stem from lack of access to or low numbers of the right people; rather, the problems stemmed from other features of the experimental design. Three of the studies we reviewed had adequate sample sizes but had important problems with other validity issues. Likewise, for the two studies with lowest sample sizes, it is fair to say that their overall validity would not have been improved by a tenfold increase in sample size. While in experimental design as in much of life, more is generally better; having a large sample size does not guarantee success nor, as witnessed by several studies mentioned in DM, does a low sample size guarantee failure.

Jeffries and Miller maintain that we misjudged the value of Jeffries et al. (1991), by failing to appreciate that the value of a study depends upon understanding "*what* kind of study is being done, *why* it's being done, and *whether* it has a chance of answering the questions it poses" (our italics). Although this strikes us as a reasonable set of principles, apparently we disagree with Jeffries and Miller as to how these principles apply to their study. Specifically, for Jeffries et al., (1991) we evaluated *what* and *why* based upon the questions posed in their introduction and the conclusions drawn in their discussion section. The *whether* was judged based upon their methodology and results section. We stand by our statements that the questions posed cannot be addressed by the study as conducted; that the conclusions drawn are not warranted by the results

obtained. We refer readers to DM's Appendix 1, where we carefully delineate why the conclusions drawn by Jeffries, et al. are not supported by their study.

Other commentators, including Carroll and Karat, take a broad perspective and speak about the difficult challenge of balancing rigor and relevance in HCI research. We share this perspective. Finding this balance requires the careful identification of research questions and empirical methods to answer those questions. Such a process involves satisficing not optimizing. Our choices must be tempered by the resources we have available. First, we must limit the scope of our questions so that they are important but answerable. Second, we must design a study or series of studies that can adequately address those questions. When resources such as subjects are limited, we need to look for ways to utilize those resources effectively. This may mean using repeated measures in an experimental design framework or using non-experimental techniques such as case studies. Third, we need to recognize that each method has its own rules of inference and its own standards of methodological rigor. When we chose to cast our study in an experimental framework, we need to carefully construct our design so that it has statistical conclusion, internal, construct, and external validity. These issues cannot be ignored when we turn to non-experimental techniques. In fact, we should use our knowledge about validity to guide our selection of appropriate methods. When a method's validity is violated, inferences cannot be drawn.

6. Where Do We Go From Here?

Among the majority of our commentators, the consensus on what the field of HCI should do to improve our knowledge of usability and UEMs can be summarized by saying that "we just have to think bigger" (B. E. John, personal communication, Feb. 13, 1998). In DM, we argued that no single measure of usability could establish the reliability and validity of a UEM and that multiple converging *measures* (section 6.1.3) were needed. We join with several of our commentators (Mackay, Oviatt, John, Newman, Monk, Carroll, and Karat) in extending this argument to include multiple converging *techniques*. Triangulation is critical to the advancement of our understanding, whether our focus is on UEMs or on other issues of HCI. In this issue, Mackay provides a thought provoking discussion on how triangulation can help improve the conference and journal review process as well as facilitate more effective studies of UEMs. John describes one use of triangulation to discover the strengths and weaknesses of UEMs.

An important component of thinking bigger is having a research agenda that extends beyond one's current study. There are many ways to characterize UEMs and many criteria by which to judge their effectiveness. Finding usability problems is central, critical, and complex in it own right. However, other factors such as when a UEM can be used, who should administer it, what types of products it is suited for, and the time and resources it demands are also very important

(John, this issue; Olson & Moran, 1996). As Karat points out, UEMs have different purposes, making it difficult, and perhaps undesirable to compare them solely by the type of usability problems they find.

Many of the points made in DM, when taken together with many of the points made by Oviatt, Newman, John, McClelland, Mackay, Monk, and Karat, may be seen as laying out a research agenda for *repairing damaged merchandise.* To be useful to practitioners, research needs to:

- develop definitions and measures of usability [construct validity issues] that are applicable in a variety of context (e.g., real-time and/or safety-critical systems, consumer appliances, point-of-sales issues, traditional office automation, etc.)
- address the interaction between UEM and the skills required to apply the UEM [an external validity of persons issue]
- determine which UEMs work best at what stage of the design [a construct validity issue concerning both mono-operation and mono-method biases]
- determine which UEMs work best with what types of software [another construct validity issue, also concerning both mono-operation and mono-method biases]
- determine in which settings (e.g., lab versus end-user's setting) various UEMs can be best applied [an external validity of setting issue]
- develop UEMs that avoid piecewise evaluation by considering the usability of software in its context of use [a causal construct validity issue as well as an external validity of setting issue]

To achieve this agenda, it is clear that multiple measures of usability are essential, that measures other than usability are necessary to understand the potential of UEMs, that experimentation must be supplemented by alternative methods, and that neither industry nor Academe can answer these questions alone. Of our commentators, Lund is very articulate on this last point. Speaking from a perspective informed by years of corporate experience, he points out that "the environment in which practitioners and many researchers work within corporations is unlikely to change" and suggests that mechanisms need to be created to enable the academic community to work with industry to "produce more generalizable results." "Teaming between industrial and academic researchers" is strongly endorsed by Oviatt. It should not be surprising that we concur. It is still true that as academics we have much freedom to set our own research agenda. However, this freedom is necessarily constrained by resource availability. It would be interesting and important to develop mechanisms by which university researchers could work

with industry developers without the research being subverted by the corporate pressures that Lund so convincingly details.

7. Conclusions

Although this rejoinder addressed a few of the specific points raised by our commentators, for the most part we avoided considering each of the many points they made in order to focus on a few of the themes that emerged overall. In doing so we see this paper as a contribution to an ongoing discussion of the methodology, empirical techniques, and rules of inference that should guide human-computer interaction research, rather than as a defense of DM We look forward to being involved in these discussions with those who wish to move the field forward.

In this vein, we eschew having the last word in this first round of discussion, and give that privilege to one of our commentators:

> Gray and Salzman have touched the tip of an iceberg. Their detailed analysis of usability studies has highlighted the need for a better research foundation for HCI. The HCI community has the opportunity to address and discuss the underlying research context and peer review process. Adding the concept of triangulation within and across scientific and design disciplines promises to facilitate communication among researchers and designers and improve the scientific basis for HCI (Mackay).

8. Notes

8.1 Acknowledgment

We give our thanks to Deborah A. Boehm-Davis for her comments on earlier versions of this paper. We likewise thank our commentators for engaging us in debate and for striving to keep the tone of the discussion focused on the highest of professional purposes. We also thank our editors, Gary M. Olson and Thomas P. Moran for highlighting the importance of the issues we raised by turning an individual submission into a special issue, for soliciting and refereeing the commentaries, and for proactively soothing the many ruffled feathers that this process inevitably entailed.

8.2 Support

The work on this paper was supported by a grant from the National Science Foundation (IRI-9618833) to Wayne D. Gray.

8.3 Authors present address

Wayne D. Gray, Human Factors and Applied Cognitive Program, George Mason University, msn 3f5, Fairfax, VA 22030, USA. E-Mail: gray@gmu.edu. Marilyn C. Salzman, Human Factors and Applied Cognitive Program, George Mason University, msn 3f5, Fairfax, VA 22030, USA. E-Mail: msalzman@gmu.edu

9. References

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. (Revised ed.). Cambridge, MA: The MIT Press.

Franzke, M. (1994). *Exploration and experienced performance with display-based systems* (Ph.D. Dissertation ICS Tech. Rpt. 94-07): University of Colorado.

Franzke, M. (1995). Turning research into practice: Characteristics of display-based interaction. In *Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems*, (pp. 421-428). New York: ACM Press.

Gray, W. D., Young, R. M., & Kirschenbaum, S. S. (1997). Introduction to this Special Issue on Cognitive Architectures and Human-Computer Interaction. *Human-Computer Interaction, 12*(4), 301-309.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the ACM CHI'91 Conference on Human Factors in Computing Systems* (pp. 119-124). New York: ACM Press.

John, B. E., & Kieras, D. E. (1996a). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction, 3*(4), 320-351.

John, B. E., & Kieras, D. E. (1996b). Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction, 3*(4), 287-319.

John, B. E., & Mashyna, M. M. (1997). Evaluating a multimedia authoring tool with Cognitive Walkthrough and think-aloud user studies. *Journal of the American Society of Information Science, 48*(9).

Kirwin, B., & Ainsworth, L. K. (Eds.). (1992). *A guide to task analysis.* Washington, DC: Taylor & Francis

Olson, J. S., & Moran, T. P. (1996). Mapping the method muddle: Guidance in using methods for user interface design. In M. Rudisill, C. Lewis, P. G. Polson, & T. D. McKay (Eds.), *Human-Computer interface designs: Success stories, emerging methods, and real world context*, . San Francisco: Morgan Kaufmann Publishers, Inc.

vanSomeren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. New York: Academic Press.

Yin, R. K. (1994). *Case study research: Design and methods.* (Second ed.). Thousand Oaks, CA: Sage Publications.

10. End Notes

[1] Throughout this rejoinder, we refer to Damaged Merchandice? as DM All references to sections are to sections within that paper.

[2]Excellent discussions of methodological rigor in non-experimental empirical studies include Yin's (1994) discussion of case studies and Ericsson and Simon (1993) or van Someren, Barnard, and Sandberg (1994) discussion of methodological issues for verbal protocol analysis.