

The Validity of UEM Comparison Studies

Today, practitioners have a number of usability evaluation methods (UEMs) at their fingertips. Various factors can affect which UEM a given usability practitioner decides to use, ranging from a project's constraints to their own familiarity with the UEM. However, studies that "prove" the superiority of one UEM over another—especially those that relate the number of problems uncovered by a UEM to their cost—have also had "an important influence" on usability practitioners making this choice (Gray and Salzman, pp. 203).

Gray and Salzman imply that many readers have accepted the conclusions of UEM comparison studies without looking carefully at the methodology used to conduct them. Fortunately (or unfortunately, depending on one's point of view), Gray and Salzman took it upon themselves to do just that: closely examine the methodology of "5 experiments that compared UEMs" (Gray and Salzman, pp. 203). Gray and Salzman's ultimate judgement, that "the studies fell far short on the criteria by which good experimental studies are designed and interpreted," had clearly shaken up the usability community at the time their work was published, and still sends ripples of doubt through this community today (Olsen and Moran, pp. 199-200). For new usability practitioners entering the field, this paper provides an introduction to Gray and Salzman's work, and advice about how to interpret their analysis of UEM comparison studies.

Validity Defined

Gray and Salzman examined what is called the "validity" of 5 studies that compared UEMs. By validity, they meant "whether the design of [each] study support[ed] the claims that were made" (Gray and Salzman, pp. 207), or more simply, whether the experiments "test[ed] what [they were] supposed to test" (Gray-Salzman slides, 2). For example, if a researcher claimed that one UEM was better than another UEM, the study that led to this claim should have actually tested whether or not this was true.

Because validity can be broken down into several types, threats to validity can creep into a study in a number of areas and are sometimes difficult to recognize. Extreme vigilance is required at every stage of a study to uphold its validity. The remainder of

this section briefly describes the different types of validity and highlights the main problems Gray and Salzman identified in the 5 UEM comparison studies they evaluated.

Statistical Conclusion Validity

Statistical conclusion validity asks “whether the independent variable is *related* to the dependent variable.” The independent variable used in the comparison studies is the UEM, while the dependent variable is typically the number of serious usability problems that are identified by it (Gray and Salzman, pp. 209-210).

For the most part, the 5 UEM comparison studies Gray and Salzman reviewed lacked statistical conclusion validity because they had too few participants to have ample statistical power, which when too low, “may cause true differences not to be noticed” (Gray and Salzman, pp. 209). The small number of participants was also not enough to counteract the Wildcard effect, an “extreme variation in individual performance [that] could have...a large influence on the stability of measures” (Gray and Salzman, pp. 221). Additionally, researchers seemed to select comparisons that “capitalize[d] on chance factors” to achieve statistically significant results (Gray and Salzman, pp. 210).

Internal Validity

If there is statistical conclusion validity, internal validity asks “whether...the independent variable *caused* the observed change in the [dependent variable]” (Gray and Salzman, pp. 209). In other words, was it the UEM that caused the number of serious usability problems found to increase/decrease?

The internal validity of many UEM comparison studies was compromised because of the selection of participants, who may have differed in terms of experience with the UEM or with usability techniques in general, as well as in their background and prior experience (sometimes with the tasks or the product being evaluated). Further, the differences in locations or time frames in which the UEMs were used could have resulted in “different experiences that may have affected the findings in indeterminable ways” (Gray and Salzman, pp. 236). Lastly, internal validity was negatively impacted by inconsistencies in the instrumentation of the test itself, including ambiguous procedures for determining problem severity (Gray and Salzman, pp. 230) and estimating time on task (Gray and Salzman, pp. 236).

Construct Validity

Construct validity asks whether the “constructs” or variables one wants to study are those being manipulated, and whether the measures taken actually reflect changes in those variables (Gray and Salzman, pp. 213). For example, are researchers really measuring the number of serious usability problems, given that severity ratings are so subjective?

Construct validity became a problem in many UEM comparison studies because of differences in the meanings of each UEM and therefore, the exact procedures for using them (Gray and Salzman, pp. 222-223). There was also much ambiguity in terms of what types of problems were uncovered and counted (unique, excluding/including false positives, and so on) as well as how problems were classified (severity, significance, and so on). In other words, the “usability measure” itself was often unclear (Gray and Salzman, pp. 227). Moreover, there were differences in how UEMs were “carried out...[that may have] affect[ed their] ability to identify problems,” or differences in the software packages under evaluation (Gray and Salzman, pp. 215). Researchers’ failure to counterbalance UEMs among available participants or to use separate groups of participants also chipped away at the studies’ construct validity (Gray and Salzman, pp. 245-246).

External Validity

External validity asks whether conclusions can be “generaliz[ed to and across] particular target persons, settings, and times” (Gray and Salzman, pp. 217). For example, if a UEM comparison study shows that heuristic evaluations performed by experts uncover some proportion of usability problems, will newly trained usability practitioners achieve the same result?

Some of the UEM comparison studies examined by Gray and Salzman lacked external validity primarily because of diversity in the participants’ background and prior knowledge of the product being evaluated (Gray and Salzman, pp. 227), as well as in the environments where the studies were conducted (Gray and Salzman, pp. 223).

Conclusion Validity

Conclusion validity asks whether “the claims made by the authors [are] consistent with the results” (Gray and Salzman, pp. 217). Several of the UEM comparison studies made conclusions that “[went] beyond the data,” made broad claims (Gray and

Salzman, pp. 224), or were in direct opposition to the data reported (Gray and Salzman, pp. 227). Further, several researchers failed to distinguish between data-supported conclusions and “experience-based advice” (Gray and Salzman, pp. 218).

Good Samaritans or Cynical Critics?

It is unlikely that budding usability practitioners like ourselves have the knowledge required to verify the validity of each published study, and it is equally unlikely (based on many of the responses to Gray and Salzman) that even experienced usability practitioners have the resources necessary to do so. As “academics who were formerly practitioners,” Gray and Salzman have taken the initiative to do this work for the usability community and as such, invite us to closely examine our field’s “truths” (Gray and Salzman, pp. 206). While many in this community (especially those whose work was examined, like Jeffries and Miller) have become defensive toward Gray and Salzman, this student is of the opinion that the authors of “Damaged Merchandise?” should instead be applauded for their efforts.

Prior to Gray and Salzman’s work, consumers of UEM comparison studies had two choices:

1. Flush out the specific portions of a study that are valid, or
2. Design and conduct another study, and see how it fared.

As Jeffries and Miller admit, executing the first of these choices “places greater demands on the reader of a study, who must be sure to tie the results and discussion back to exactly what was—and what was not—done by the experimenters” (pp. 271-272). As a student learning that the job of a usability practitioner is to alleviate the burdens placed on users, the fact that our own research does not follow this advice seems hypocritical. Additionally, this choice assumes that the reader is aware of the fact that there may be invalid portions of a UEM comparison study, something this student (as a new reader) did not seriously consider before reading Gray and Salzman. Given the number of studies that have been published on this topic, however, it appears as though most usability practitioners have taken the second route. The UEM comparison studies Gray and Salzman evaluate are limited to 5 but were reduced from an original 11, alluding to their abundance (Gray and Salzman, pp. 219). However, quantity is not necessarily as important as quality, and many consumers of these studies (including experienced usability practitioners) are left to ponder contradictions like this one:

“Jeffries et al. found that the experts uncovered more problems than the usability test did, while other studies found that the experts uncovered less than half of the problems uncovered by a usability test” (Dumas and Redish, pp. 81).

Gray and Salzman’s careful examination of UEM comparison studies provides an explanation of why such contradictions run rampant, and reminds us that popularity may not always be indicative of usefulness.

What Do We Know?

According to Gray and Salzman, “up to now, no study fulfills the rigorous standards of experimental design,” and therefore none of the 5 UEM comparison studies can be considered valid (Gediga et al., pp. 24). Even more troubling, Gray and Salzman clearly state that “when [the] validity [of a study’s method] is violated, inferences cannot be drawn” (pp. 332). All but two commentaries written in response to Gray and Salzman explicitly state or imply their agreement that the 5 UEM comparison studies succumbed to threats to their validity. Given this wide-spread acknowledgement of the studies’ flaws, can we still salvage useful information from them?

Many experienced usability practitioners seem to think so. In their discussion of comparing usability testing with other UEMs, Dumas and Redish list a number of conclusions that are “suggested” by some of the same UEM comparison studies, based on the “consistent findings” and “available evidence” (pp. 82). Several would agree with the more definitive Karat, who asserts that these studies “all represent good work worthy of publication” (pp. 267), and asks us “consider what constitutes a ‘useful publication’” (pp. 266).

Empirical Validity Versus Ecological Validity

The heart of most seasoned usability practitioners’ counter argument lies in the type of validity against which the UEM comparison studies are evaluated. Usability practitioners overwhelmingly agree that reports of studies adhering to Gray and Salzman’s purely empirical definition of validity have “failed to provide the generality needed to make them relevant” (Karat, pp. 268). Instead, usability practitioners champion ecological validity, which focuses more on how UEMs are applied in real-world situations that involve less scientific factors such as resource constraints, development life cycles, corporate politics, and so on. From this point of view, one can see why the 5 UEM comparison studies were so popular: they helped usability practitioners do their jobs. And in the real-world, that equates to “extremely useful.”

Unfortunately, usability practitioners believe that generating empirically valid data (of the kind that would please Gray and Salzman) is impossible to do while maintaining ecological validity (Baber and Stanton, pp. 88; Jeffries and Miller, pp. 271). Gediga, et al. offer more support for this notion, stating that a “pragmatic view puts heavy restrictions on empirical studies, because a *valid* benchmarking is only feasible if the methods are applied exactly as they will be used in practice” (pp. 25, emphasis mine). John asserts that “experiments do not always tell enough about what caused the outcome to be of pragmatic use” (pp. 290). But perhaps the best summation of this conundrum comes from Carroll, who says: “we cannot adjust the requirements of the world to our models; we must in fact do the reverse” (pp. 309).

Challenging the Challengers

Given these statements, it would be interesting to know how usability practitioners would rate the ecological validity of the few “well-designed” studies that Gray and Salzman mention (Gray and Salzman, pp. 247-248). Usability practitioners may be vindicated in their assertion that one cannot have both empirical and ecological validity, or learn that their rationalizations are ill founded.

If it is truly impossible to balance both empirical and ecological validity, then “[the UEM comparison] studies are all examples of the reasonable compromises that were necessary” (Karat, pp. 267), and the studies’ “flaws” (such as confounding variables) may instead be perceived “as specific properties of the methods” (Gediga, et al., pp. 25). Flaws—especially those that Gray and Salzman would view as threats to internal validity—are often extremely important to the usability practitioner. In other words, usability practitioners feel that “[these were] not...bug[s] in the experimental design, but...feature[s]” (Jeffries and Miller, pp. 273). From this angle, the inconsistencies in the results of UEM comparison studies are admittedly attributed to the different contexts under which the studies were carried out, and the generalizability of the studies depend on how closely a usability practitioner’s work environment, procedures, and processes match that of the studies’ authors. Based on the responses to “Damaged Merchandise?,” it appears as though a large portion of the usability community did in fact find the 5 UEM comparison studies Gray and Salzman examined to be useful. A reason for this conclusion is summarized by Lund’s statement that usability practitioners “[are] prepared to interpret the results [of the UEM comparison studies] based on the filter of their own experience” (pp. 280). However, this does little to help those of us starting out in the usability field, as we have yet to build up our knowledge reserves.

Another problem with this rationale is captured by Jeffries and Miller's declarations that the "better course to understanding a large question...is to begin with a broad, general overview of the question, emphasizing the real-world aspects of the problem even at the cost of experimental purity" and that "if successful, such a study will raise smaller, more focused questions that can be addressed through methodologically more rigorous experiments" (pp. 272). Jeffries and Miller believe this has happened, overlooking Gray and Salzman's point that "all [5 UEM comparison studies] have had enough time for fuller reports to appear in the literature (none have)" (Gray and Salzman, pp. 208).

Lastly, Oviatt urges the usability community to "abandon any self-defeating myths implying that HCI research is not the domain of 'real science'" (pp. 306). This student finds it particularly difficult to believe that no other behavioral sciences have found ways for academics and practitioners to work together to design empirically yet ecologically valid research studies. Doing one thing right always requires more time and energy than doing many things inadequately, and although usability practitioners are probably overwhelmed with day-to-day activities, it sounds like they just lack the motivation necessary to undertake this task.

Gray and Salzman's Legacy

Rather than debating over whether the UEM comparison studies are useful (because in many usability practitioners' minds, there is no debate), many responses to Gray and Salzman's work instead focus on the bigger picture of how to improve our thinking about different UEMs. Some, like Monk, implore the usability community to "think clearly about...fundamental issues," such as how to measure usability, and how to define UEMs in terms of how they are "really used in practice" (pp. 301) before attempting better comparison studies. Others, like John, Mackay, and Newman indicate that "controlled experiments" are not the only way to gain an "understanding of "UEM use and effectiveness" (John, pp. 289). Instead, they remind us about case studies, field studies, simulations, and triangulation as ways to obtain useful data. In some ways, Gray and Salzman's argument is a semantic one: unless a study is empirically valid, authors should not call it "research." It is quite possible that the usability community has inaccurately "applied a research metaphor" to more than just usability tests (Dumas, pp. 3).

Conclusion

Gray and Salzman's work had a number of important consequences. First, their examination of UEM comparison studies reminded the usability community about validity and the various threats to a study's validity. These discussions of scientific rigor may be based in academia, but they should not be entirely forgotten when one moves to practical application. Second, it caused practitioners to take a closer look at just what was being compared and the ultimate purpose of these comparisons. All of us are susceptible to tunnel vision, and sometimes it can be helpful to backtrack and figure out the "what's" and "why's" before proceeding down a path. Third, it heightened practitioners' awareness of the need for carefully constructed and reported test methods, as well as a discussion about the limitations of any given study. Fourth, it encouraged researchers to more clearly separate conclusions based on the data from their own, experience-based recommendations. Few people have time to perform careful, independent verifications or separate these claims. Last, Gray and Salzman's work urged researchers to consider whether every publication should be deemed an empirical study, rather than some other type of paper that carries with it less connotative overhead.

It is this student's opinion that Gray and Salzman proved the 5 UEM comparison studies to be empirically invalid, but that opponents have not proven Gray and Salzman's suggestions to be impossible nor worthless. The usability community (academics and practitioners alike) should come together to seriously design and conduct a few rigorous empirical studies so we would know once and for all whether or not they can work. If these studies can be valid *and* relevant, individuals starting out in the field would be more confident in their choices of UEMs. If not, the profession could move on to brainstorming other possibilities instead of continuing to expend resources weeding out inconsistencies, pondering contradictions, and debating this issue.

References

- Baber, C. and Stanton, N. (1996). "Observation as a Technique for Usability Evaluation." Usability Evaluation in Industry. Jordan, P.W., Thomas, B., Weerdmeester, B.A., and McClelland, I.L., Eds., London: Taylor & Francis, Ltd.
- Carroll, J.M. (1998). "Review Validity, Causal Analysis, and Rare Evaluation Events." In Human Computer Interaction, (13) 3, pp. 308-310.

Dumas, J.S. (2002). "Gray-Salzman Slides." Class slides for Testing and Assessment Programs course at Bentley College, Waltham, MA.

Dumas, J.S. (1999). "Usability Testing Methods: When Does a Usability Test Become a Research Experiment?" In *Common Ground*, 9 (May), pp. 1-5.

Dumas, J.S. and Redish, J.C. (1999). *A Practical Guide to Usability Testing*. Exeter, UK: Intellect Books, Ltd.

Gediga, G., Hambord, K.-C., and Duntsch, I. (2002). "Evaluation of Software Systems." Accessed 27 October 2002.

<http://www.cosc.brocku.ca/~duentsch/archive/softeval.pdf>.

Gray, W. and Salzman, M. (1998). "Damaged Merchandise? A Review of Experiments That Compare Usability Methods." (Moran, H.P., Ed.) In *Human Computer Interaction*, (13) 3, pp. 203-261.

Jeffries, R. and Miller, J.R. (1998). "Ivory Towers in the Trenches: Different Perspectives on Usability Evaluations." In *Human Computer Interaction*, (13) 3, pp. 270-276.

John, B.E. (1998). "A Case for Cases." In *Human Computer Interaction*, (13) 3, pp. 289-295.

Karat, J. (1998). "The Fine Art of Comparing Apples and Oranges." In *Human Computer Interaction*, (13) 3, pp. 265-269.

Lund, A.M. (1998). "Damaged Merchandise? Comments on Shopping at Outlet Malls." In *Human Computer Interaction*, (13) 3, pp. 276-281.

Monk, A.F. (1998). "Experiments Are For Small Questions, Not Large Ones Like 'What Usability Evaluation Method Should I Use?'" In *Human Computer Interaction*, (13) 3, pp. 296-303.

Olsen H.P. and Moran, T.P. (1998). "Introduction to This Special Issue on Experimental Comparisons of Usability Evaluation Methods." In *Human Computer Interaction*, (13) 3, pp. 199-201.

Oviatt, S.L. (1998). "What's Science Got to Do With It? Designing HCI Studies That Ask Big Questions and Get Results That Matter." In *Human Computer Interaction*, (13) 3, pp. 303-307.

