

Feature

Experimental Design: Does External Validity Trump Internal Validity?

Jeremiah Still | Missouri Western State University | jstill2@missouriwestern.edu

Like many who submit manuscripts to the CHI conference each year, I look forward to reading the reviewers' reflections on my submissions. This year my coauthors and I were asked to justify the validity of our highly controlled research; similar requests have been made of our recent submissions to journals. In these experiments, we went to great lengths to ensure a high degree of internal validity. Our research goal was to establish a cause-and-effect relationship between what we were manipulating and what we were measuring. The only way to establish cause-and-effect relationships is by using designs with high internal validity.

It is not uncommon for researchers to be asked to reframe their highly controlled experimental designs in applied terms—people want to know the real-world application of the work. However, based on recent feedback from review committees, it has been suggested studies high in internal validity are of little value to CHI participants. To illustrate this point, let's explore one of my students' CHI reviews. The associate chair provided an excellent summary regarding the "issue" of design's high internal validity: "The reviewers agree that this idea has merit and could be a useful tool for HCI designers, but raised some concerns about validity of the results (this issue was also discussed by the authors)." One reviewer stated, "...The nature of the experiment was not particularly ecologically valid..." Another reviewer wrote, "The authors are aware of this shortcoming and point out that 'the situations presented in the experiment were relatively minimal and artificial.' I contend that the situations were minimal and artificial enough that it is not possible to draw conclusions... [this variable] should be tested in a more realistic task in order to evaluate viability."

What follows is a typical rebuttal I might offer:

This experiment does indeed lack a high amount of external ecological validity (which we hung a lantern on), but this design comes with the benefit of having a high amount of internal validity. Our goal was to understand the cognitive operations typically involved with the interaction types, not to investigate its use in a particular setting. For instance, we did not set out to only understand "X" interaction type within the context of "Y." However, we do hope that future research will test the external validity of the claims presented in this manuscript.

Describing Internal and External Validity

Let's take a moment to define internal and external validity in their purest forms. In an experiment with high internal validity, a variable of interest (i.e., independent variable) is manipulated to determine its effect on something being measured (i.e., dependent measure). The advantage of using a design with high internal validity is its ability to control for confounding variables. In other words, it allows researchers to be confident that the variable being manipulated is causing the change seen within the dependent measure. In contrast, a study with high external validity might be conducted in a more natural setting, for instance, studying user interactions with a piece of equipment in the specific situation in which that equipment is typically used. Studies with high external validity are beneficial because they allow one to observe many variables interacting with one another; these interactions are purposely minimized in studies with high internal validity. The primary drawback is the presence of co-varying variables precludes the researcher from establishing cause-and-effect relationships. Most research falls somewhere between these two extremes of validity; those are referred to as quasi-experiments.

We suspect that some have the incorrect mental model of experimental validity. They believe validity is dichotomous, consisting of two independent categories—a study can have internal validity or external validity—and therefore, for instance, a study with internal validity cannot have external validity. Conversely, it is the case that any particular study's validity fails on a continuum between the two extremes. With this more appropriate mental model, it becomes clear that identification of a study as having "pure" internal or external validity is difficult (See Figure 1). Most HCI studies would be considered quasi-experimental. That is, some independent variables are manipulated by the researcher, allowing for cause-and-effect conclusions (e.g., simulator training: yes or no), while other experimental variables of interest are not manipulated (e.g., age, working memory capacity of participants). When variables of

interest are not manipulated, cause-and-effect conclusions cannot be drawn because other variables may co-vary with those variables of interest. For example, participants' working memory capacity, although it may be of interest, cannot be manipulated. Working memory capacity may co-vary with a number of variables, including intelligence, experience with the task/material, and attention; therefore, one cannot say that differences in working memory alone result in differential task performance.

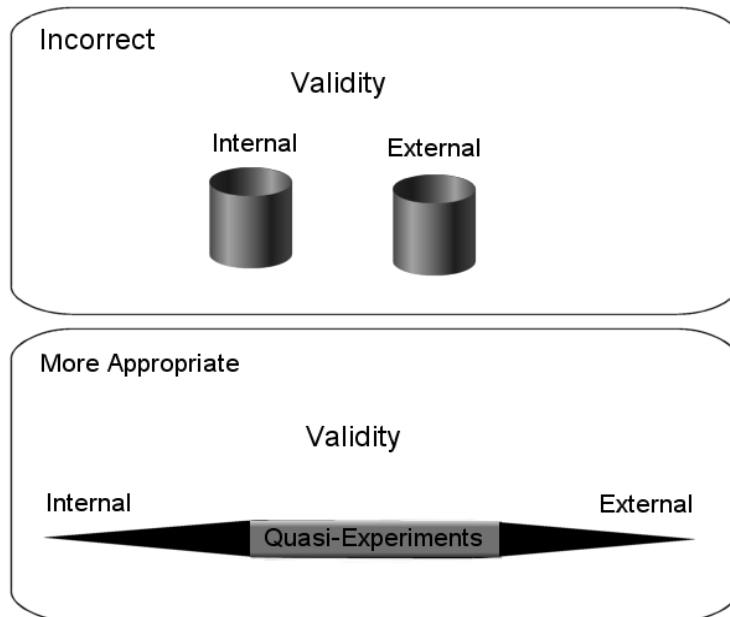


Figure 1 displays two mental models of internal and external validity. The top model is incorrect, as each type of validity is categorically separate. The bottom model is more appropriate, as the continuous bar reflects the continuum between the two extreme states of validity. The thickness of the bar represents hypothetical prevalence of design types used in interaction design studies—quasi-experiments are most common, while experiments with more internal or external validity are less common.

Associating Validity with Basic and Applied Research

I have suggested studies with high internal validity do have a place in the HCI literature, but there is a second conceptual issue that reviewers may take issue with. They may have incorrectly equated internal validity with basic research and external validity with applied research, as is often done. The difference between basic and applied research lies not in experimental validity; rather, the difference lies in their intended purpose [2]. Basic research is most often performed without an immediate practical purpose. Many who execute basic research hope it will someday have a direct and practical benefit but primarily hope to further scientific understanding of a particular topic by discovering and describing the principles that govern a particular phenomenon. On the other hand, applied research is performed with the intent of providing a solution for a specific, practical problem. In a mature research program, like that directed by Paul Fitts, basic and applied research are used to inform each other. In addition, it is important to note either basic or applied research can have a high degree of internal or external validity. An example of applied research with high internal validity would be searching for a cure to cancer within a laboratory environment. An example of basic research with high external validity would be performing a case study of a newly discovered animal's behavior within its natural environment. These two examples highlight research that clearly violates the incorrect conceptualization that internal validity is found only in basic research and external validity is found only in applied research. It could be the case that the reviewers had this incorrect conceptualization. Of course, it is also possible that there is a correct understanding of experimental validity but that some reviewers simply do not value basic research.

Why Our Future Depends on Basic Research

Let me share a fictional scenario to illustrate the importance of basic research. I believe it especially rings true for the field of HCI, as we never know what future technologies will come to fruition. Imagine that a researcher studies the game of croquet—a lawn game that involves whacking balls through hoops. This

researcher manipulates the velocity of the balls, lawn conditions (e.g., slope, type of grass), ball placement (e.g., angle, distance), and ball mass. He measures where the whacked ball stops, how long it took to stop, and one ball's displacement of the other balls in play. Through his experimental work, he hopes to be able to predict where a ball will stop under a variety of conditions. Other academics on campus mock the research, saying it has no real value. Further, administrators pressure him to quit his current work and explore more funding-attractive research. The research simply does not provide a solution to a currently recognized problem. However, after many years (and the rise and fall of many popular problem sets), the world suddenly fears that a giant asteroid, crossing through the asteroid belt between Mars and Jupiter, may hit the Earth, resulting in Armageddon. Now the professor's croquet-game research is relevant! Based on his previous research, he is able to predict where the asteroid would strike and what force would be required to alter its course. This scenario illustrates the potential of basic research for the rapid resolution of unforeseen, practical problems.

Conclusion

My goal here is simply to remind researchers that with any design decision—experimental design or interaction design—there is a trade-off between benefits and costs. We strongly believe the field of HCI benefits from the acceptance and support of research falling near the internally valid end of the validity spectrum (bottom panel of Figure 1). Again, this type of experimental validity produces knowledge about the cause-and-effect relationship between variables and measurements of interest. However, in using this design, we accept that it does come at the cost of not being able to apply our findings to specific instances of interaction with absolute confidence.

It is vital for the future development of the field of interaction design that authors recognize the trade-offs associated with their experimental design decisions. Remember, “any measure device is valid if it does what it is intended to do” [1]. We ought to be transparent about our research validity goals and seek quality research, whether it is high in internal or external validity or somewhere between.

Endnotes

[1] Carmines, E. G. and Zeller, R. A. *Reliability and Validity Assessment*. Series: Quantitative Applications in the Social Sciences. Sage Publications, London, 1979.

[2.] Nickerson, R. S. Basic versus applied research. In R. J. Sternberg (Ed.), *The nature of cognition*. MIT Press, Cambridge, 1999, 409-444.



About the Author

Jeremiah D. Still is an assistant professor in the Department of Psychology at Missouri Western State University. He directs a graduate program that integrates HCI and MBA coursework. Jeremiah focuses on understanding human-centered design from a cognitive viewpoint; how a user will perceive, process, and respond to a product's design. He received his Ph.D. from Iowa State University in Human-Computer Interaction.

Copyright ACM, (2011). This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in *Interactions*, VOL18, ISS13, (May + June) <http://doi.acm.org/10.1145/1962438.192453>