Triangulation within and across HCI disciplines

Wendy E. Mackay
Université de Paris-Sud

How should we evaluate interactive software? Gray and Salzman (1998) examine five influential studies that compare methods of evaluating usability. In a carefully-argued critique, they identify numerous threats to validity and point out serious flaws in both the designs and claims made in each study. I think there will be various reactions to this article. Some people will react with alarm, concerned that if frequently-cited articles in respected publications are so full of errors, perhaps the field of human-computer interaction is in serious trouble. Others may attack the article, mis-interpreting it as the blind application of laboratory-style experiments to field studies. My own reaction is to commend Gray and Salzman for daring to critique some of the most influential work in the field, showing not only how to identify certain kinds of problems, but also how to address them. Their study helps to advance the field by raising the level of critical analysis in a graceful and constructive manner. This is how research disciplines make progress: through self-analysis and the creation of a solid research context.

Field studies vs. laboratory experiments

Like many others in HCI, I started out with academic degrees in experimental psychology. Moving into industry, I quickly realized that laboratory experiments were rarely useful for addressing the HCI problems we were interested in. We borrowed and reinvented design and evaluation methods, keeping what worked and exploring ways of improving them. It was not until I later returned to academia for a Ph.D. in a different field (management of technological innovation) that I discovered the wealth of relevant literature from other social sciences. My well-thumbed copy of Cook & Campbell proved far more useful than my aging Psychology texts and research articles. As a psychologist, I was conditioned to believe that other social sciences were not particularly scientific. I discovered that there exists a solid foundation for conducting field studies, grounded in scientific principles but tempered with humility: the constraints of the real world make this type of research difficult to do well.

For those readers who are discouraged after reading Gray & Salzman's critique, I strongly recommend Cook & Campbell (1979) for concrete advice on how to conduct and analyze "quasi-experiments" or field studies. They begin by explaining the philosophical history of causal inference and the scientific method, including various measures of validity. The rest of the book is devoted to descriptions of quasi-experimental designs and modes of analyzing the data that result from them.
For each design, they point out the most likely sources of invalid conclusions and offer suggestions for how best to avoid them. "Threats to validity" exist in any research setting. What Cook and Campbell offer are methods for conducting scientifically-grounded research, given the variety of constraints found in real-world settings.

Building a scientific research context for HCI

All natural sciences assume that individual studies build upon each other: no single study can ever address all possible threats to validity or potential biases by the researchers. Because individual studies are never viewed as conclusive, subsequent studies are required to present new results and to replicate or challenge earlier findings. Multiple studies, run by different people in different settings, help to reduce the number and severity of threats to validity and increase the strength and power of the results.

Human-Computer Interaction, as a field, lacks an agreed-upon research context. Individual studies rarely replicate results from previous studies, much less build upon them. This is partly due to its youth: it takes time to develop a shared research

foundation. This is also due to the multidisciplinary nature of the field: researchers trained in one discipline (such as computer science) often feel compelled to perform studies outside of their expertise (such as laboratory experiments) in order for their work to be accepted. Even people with scientific training (such as experimental psychologists) may not be trained in research techniques that are most relevant to the problem at hand (i.e., field studies). Directly applying laboratory methods in field settings gives poor results; yet providing only anecdotal evidence misses an opportunity. HCI researchers must develop a set of research techniques that apply to the problems specific to HCI, either borrowed from other fields or invented for the purpose.

Another critical aspect of a shared research context is the peer review process. This is particularly complex in HCI, with people trained in many disciplines, sometimes with little or no overlap. For example, we do not have a good notion of "expertise". Is a computer scientist with extensive experience in graphical user interfaces but no background in statistics considered an "expert" reviewer of an experiment comparing different graphical interfaces? Should a university statistics course taken a decade ago be considered adequate background for critiquing an experimental paper? Asking people with diverse backgrounds to review a paper helps to ensure that the resulting papers are accessible to a wider, more diverse audience. It also avoids a common phenomenon in many sciences, in which tiny groups of researchers concentrate on increasingly-detailed aspects of a problem and end up speaking only to themselves. On the other hand, we need a better classification of expertise; distinguishing between expertise in the discipline, in the specific content, and in the development or evaluation techniques used. We must agree upon standards for particular types of papers and try to ensure that papers are reviewed by people with the necessary technical expertise. We also need to learn to value papers with muted claims, rather than looking only for "exciting" results that may simply be the result of a less-than-careful analysis.

Triangulation within research studies

For HCI to mature as a field, we must develop a shared research context, despite our internal diversity. Replication of studies is not enough. We need triangulation: using different techniques to operationalize behavior (i.e., specifying specific, measurable actions) while attempting to measure the same phenomenon. Gray and Salzman describe how Bailey et al. (1992) used triangulation when they conducted two independent experiments to evaluate Molich's and Neilsen's (1990) Heuristic Evaluation technique:

"A strength of this report is that experiment 2 essentially replicates the experiment 1 findings using a different style of interface. This greatly increases the generality and construct validity of the findings." (p.42)

Another type of triangulation involves examining different types of data gathered within the same study. For example, I conducted a controlled study of an "intelligent tutor" for a text editor, in an industrial setting. In the early 1980's, applying artificial intelligence to on-line teaching was all the rage. We wondered just how much "intelligence" was necessary. Before actually implementing our tutor, we used a Wizard of Oz technique to decide. We used a within-subjects design, providing identical amounts of tutoring to pairs of subjects. For one third of the commands, the wizard would watch one subject performing a specified task and send a tutoring message when it was relevant to the task the subject was currently performing. The other subject of the pair would receive the same tutoring message, independently of what she was doing. For the second set of commands, the researcher would focus on the other subject in the pair, providing relevant tutoring messages to her and effectively random advice to the first subject. No tutoring was given for the remaining third of the commands. The quantitative results were interesting: subjects learned both sets of tutored commands, regardless of whether or not the tutoring was related to their current behavior. (In the control condition,

subjects learned very few of the untutored commands.) The purely quantitative analysis suggested that subjects did not "need" an intelligent tutor: random presentation of advice was equally effective. However the qualitative results were quite different. All but one of the subjects reported that they really enjoyed receiving tutoring when it was contingent on their behavior, and disliked it when it appeared randomly.

By triangulating, we gain a deeper understanding of what is going on. But, of course, the study still leaves major design questions. Do we decide to ignore the qualitative data and simply implement a random tutor, since it's much cheaper and equally effective from a learning standpoint? Or do we listen to the qualitative data and try to figure out a different method for presenting messages that users don't find as intrusive, perhaps by letting them decide when they want their "random" tutor activated.

Triangulation across scientific disciplines

McGrath, Martin and Kulka (1982) discuss the dilemmas faced by researchers when choosing among research methods. They identify eight distinguishable research strategies: Laboratory experiments, Experimental simulations, Field experiments, Field Studies, Computer Simulations, Formal Theory, Sample Surveys and Judgment Tasks.

They illustrate the relationships among these research methods by laying them out as pie slices in a circle. The eight methods are classified as occurring in natural settings, in contrived or created settings, independent of natural settings or methods in which no observation of behavior is required. So, for example, field experiments and field studies are classed together as "settings in natural systems", whereas formal theory and computer simulations are "methods that do not require direct observation of behavior". Methods are also classified as obtrusive or unobtrusive and as universal or particular. They point out that, other things being equal, the researcher hopes to maximize three mutually-conflicting goals:

> " A. generalizability with respect to populations
> B. Precision in control and measurement of variables related to the behaviors of interest, and
> C. existential realism, for the participants, of the context within which those behaviors are observed.

But alas, ceteris is never paribus, in the world of research...The very choices and operations by which one can seek to maximize any one of these will reduce the other two; and the choices that would "optimize" on any two will minimize on the third. Thus, the research strategy domain is a three-horned dilemma, and every research strategy either avoids two horns by an uneasy compromise but gets impaled, to the hilt, on the third horn; or it grabs the dilemma boldly by one horn, maximizing on it, but at the same time "sitting down" (with some pain) on the other two horns." (p.74)

This is the crux of the problem. Field studies maximize "existential realism", but suffer with respect to generalizability and precision in control. Laboratory experiments maximize precision of measurement but suffer with respect to generalizability and existential realism. Experimental simulations and field experiments fit between these too, but they too involve compromises. In Mackay and Fayard (1997), we describe a set of research studies at the C.E.N.A., the French center for studies in air traffic control. The studies include biological analyses of sleep patterns of controllers, laboratory experiments of different user interface strategies, computer simulations of new tools used by controllers, cognitive models of air traffic control's activities and ethnographic studies of controllers at work. We argue that triangulation across these different research methods provides a deeper understanding of the problems faced by controllers and leads to better design solutions. Triangulation across scientific research methods is not always obvious. We need to develop methods of reporting research results to make it easier to make cross-disciplinary comparisons.

Triangulation across science and design disciplines

HCI needs a sophisticated model of triangulation; comparing quantitative and qualitative results within individual studies, performing different kinds of experiments to test the same phenomenon, and comparing results from studies derived from different scientific disciplines. But even this is not enough. All of these forms of triangulation assume a shared scientific context. Yet human computer interaction has another very important component: design.

By designers, I do not mean people trained in computer science or psychology. I mean people with academic training in a design discipline, whether it is graphic design, architecture or typography. The concept of building upon previous work exists within design disciplines, but in a manner very different from the sciences. Designers are trained how to see (or hear) as well as how to produce. Design students are taught design rules and are then evaluated on how well they break them. Design mixes craft and artistic sense; designers learn by exposing themselves to a variety of ideas, good and bad. Creating quality interactive software has an important design component: a good user interface is more than the lack of usability problems. For designers, triangulation concerns the use of multiple techniques for creating new artifacts, rather than evaluating them.

Human computer interaction is a strange field. The artifacts we study are not natural phenomena in the ordinary "scientific" sense; they are artifacts created by people. We may study the artifacts themselves, but more commonly, we study the interaction between people and these artifacts. We can benefit greatly from techniques borrowed from other scientific fields, but we must reexamine them and in some cases, recreate them. Triangulation is probably the only realistic solution for dealing with different underlying assumptions and clashing paradigms. Yet we do not really know how to triangulate across scientific and design disciplines.

Conclusion

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 scientifc and design disciplines promises to facilitate communication among researchers and designers and improve the scientific basis for HCI.

References:

Bailey, R.W., Allan, R.W. and Raiello, P. (1992) Usability testing vs. heuristic evaluation: a head-to-head comparison. In Proceedings of the Human Factors Society 36th Annual Meeting (pp. 409-413). Santa Monica, CA: Human Factors Society.

Cook, T. and Campbell, D. (1979) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston, MA: Houghton-Mifflin Company.

Gray and Salzman (1998) Damaged Merchandise? A review of experiments that compare usability evaluation methods. To appear in HCI.

Molich, R. and Nielsen, J. (1990) Improving a human-computer dialogue. Communications of the ACM. 33 (3), pp. 338-348.

Mackay, W.E. and Fayard, A.L. (1997) HCI, natural science and design: A framework

for triangulation across disciplines. In Proceedings of Designing Interactive Systems, DIS'97, ACM: Amsterdam, the Netherlands.

McGrath, J., Martin, J. & Kulka, J. (1982) Judgment Calls in Research. CA: Sage Publications.

Runkel, P. and McGrath, J. (1972) Research on Human Behavior: A Systematic Guide to Method. NY: Holt, Rinehart & Winston.