Usability is a property of devices (interfaces, work methods, software, and so on) but it’s also a field—and some would say it’s a profession. Here, we examine the concept of usability from the standpoint of human-centered computing. Manifestly, the two should be quite closely related. The principles of HCC that have been explored in this department dictate a prime goal for intelligent technology: It must be useful, understandable … and usable.

Is Usability Analysis At Risk?
At the 2008 South by Southwest–Interactive Conference in Austin, Texas, consultant Jared Spool announced to a roomful of more than a thousand attendees that the era of user-centered design (UCD) was dead, UCD having been proven ineffective. He supported this claim with an assertion that Apple, with their broadly acknowledged successful user experience, doesn’t use UCD processes, and with the additional claim that there had been no empirical report of successful implementation of UCD.

In stark contradiction to Spool’s claim of UCD being proven ineffective, hundreds of published (and likely thousands of unpublished) studies every year demonstrate an improvement to user interfaces thanks to various types of human factors studies (usability testing, usability inspections, or user requirements analysis). Usually, it’s easy for individuals with a background in human factors, cognitive psychology, or business ethnography to examine software tools or interfaces and design something better, because the products of technologists (however smart and well-intentioned) were based on designer-centered design, with the naive view that the designer-technologist can easily walk in the shoes of the workers who would use the tools.

We might be witnessing a perfect storm of converging forces that make usability more important than ever before. The Internet’s growth and the spread of intelligent technologies across work environments underscore the claim that the pool of usability analysts needs to be “scaled up by a factor of 100.” But there are also forces that put usability analysis at risk. One such force is the widespread belief in the myths of automation, which foster the belief that simply injecting more technology into the workplace will solve the problems of cognitive and adaptive complexity.

Also putting usability analysis at risk are the problems with procurement processes—experienced most recently in the problems with website development in support of the Affordable Care Act. When a software developer does a poor job of coding, this gets discovered (usually) at system test time. When someone does a poor job of usability engineering, it doesn’t get discovered until the product is out the door and the customer support phones start ringing off the hook, or until the site is live and visitors are leaving it in droves. And then, the manager who paid for that usability work doesn’t say, “We received poor usability support,” he or she says, “Usability isn’t worth it.”

The pace of change is yet another force that puts usability analysis at risk; the endless parade of upgrades and releases. How long does it take you to relearn how to “mail merge” when sending out holiday cards each year? Any of a plethora of quotations could be used to name the problem. The response is the abrogation of responsibility and the default reliance on mere satisficing. “Let’s get something out there and rely on user feedback to steer the design of release 2.”

And thus we explore the question: If usability is to be a valuable, empirical methodology … then where’s the science in usability analysis?

Where’s the Science?
We conducted a Google search on “usability testing” looking for links to, say, articles published in peer-reviewed scientific journals or links to scientific books.
We had to go many clicks deep to find a link to anything with a modicum of empirical respectability, and that link led to a book about UCD. The prior 51 links were primarily commercial sites, touting “many happy customers,” offering software that does quick-fix tests of webpage usability. Next, we did a similar search on “performance measurement,” hopeful that this search phrase would certainly yield copious links to actual scientific work. Our hopes were quickly dashed. Twenty links deep we found a link to the journal Performance Measurement and Metrics. The prior 19 were all commercial.

There seem to be many people out there whose stock-in-trade is convincing other people that they can help them make better decisions. Most of the sites seem to have a business pitch that essentially says that people without any particular experience at the application of an established empirical methodology can follow some simple steps and do fast usability testing themselves. And thus, the savvy business leader can avoid having to pay the “expensive professionals.” Ironically, the commercial sites tout their own cadre of experts.

In our search, we were particularly struck by the ISO definition of usability (and its derivatives, such as the Wikipedia definition): “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.”6 Effectiveness, efficiency, and satisfaction are certainly good goals, from a certain perspective.6 But how are these things to be evaluated or studied? ISO specification of the concepts of effectiveness, efficiency, and satisfaction entail measures of affect and measures of performance (including accuracy and completeness). Too often, the activity called usability analysis consists of questionnaires, interviews or “expert review,” which can tap the former but not the latter. Where the rubber meets the road, there isn’t much rubber. It’s hard to find science, even where you might expect to see some.

**The Training and Certification Gap**

Our Google search notwithstanding, there are authentic and highly competent usability researchers and practitioners—and these should be given due credit for their excellent contributions. We are, however, skeptical about the extent to which usability professionals put themselves under a sufficiently powerful empirical microscope. Too much usability analysis is, bluntly put, amateurish. Worse, the lack of requisite skill sets and an agreed-upon list of best practices is actively promoted in the field. The theme of the Usability Professionals’ Association annual meeting in 2006 was “Usability through Storytelling.” Storytelling has value, as does any method. Learning from stories is arguably an expression of the empiricist philosophy that knowledge comes from experience. But usability analysis must be more than storytelling. The theme of the 2008 Usability Professionals’ Association annual meeting was “The Many Faces of User Experience: Usability through Holistic Practice,” as the society splayed open its arms—“Marketing specialists, graphic designers, computer scientists, business analysts, psychologists, information architects, technical writers and others bring valuable perspectives to usability and user experience” (see www.usabilityprofessionals.org/conference/2008). Welcoming everyone into the tent isn’t likely to produce an identifiable discipline, nor robust results.

There’s no well-accepted or agreed-upon certification in usability engineering. There’s no standard curriculum for usability professionals. In the range of fields that span psychology to computer science, we provide insufficient education in the history of cognitive task analysis and insufficient training in the conduct of experiments—which is what usability analyses really are, or should be (at least more often). The lack of certification to ensure a level of quality has led to the undermining of usability’s recognition as a scientifically grounded discipline, and an understandable reluctance to give usability analysts a seat at the software development table.

The individuals who conduct usability analysis should be demonstrably proficient at cognitive task analysis. And in addition to the need for properly trained personnel and the need for some sort of certification system, we need to advance the methodology.

**The Methodology Gap**

Of course, usability as a discipline is itself influenced by emerging technology. Tools such as WebEx or GoToMeeting have enabled remote usability evaluations,7 allowing easier access to representative test participants around the globe. Crowdsourcing is also a possibility for usability testing.8 But to achieve rigor, there must be more than just increasing the sample size by leveraging the Internet.

There has recently been interest in using psychophysical measurements such as eye movements, galvanic skin response, heart rate, and fMRI. These approaches are a narrow window to the mind. What a person is looking at doesn’t tell you what they’re thinking, or what they know, or what they’ve learned. Hence, none of the physiological methods alone establishes the kind of rigor that’s called for; no matter how much they might pander to a quantitative quest for objective purity.

Usability analysis using the method of professional judgment is almost
always the analysis of some single design, the one presented by the customer. Rarely are independent variables manipulated or controlled within a formal experimental design. Anyone happy with "3–5 users"19 is unlikely to conduct an actual experiment and apply statistical tests. It should be admitted that the irrational limitations on formal experimentation and the number of participants necessary for rigorous usability analysis often are imposed by shortsighted managers and programmatic constraints rather than chosen voluntarily by usability professionals. Recognize-and-fix usability evaluations are also imposed.

Nevertheless, regrettably few resources are devoted to evaluating the choice of methods used,10 including their effectiveness and sufficiency. Very little research has been conducted to determine which usability methods are best applied in which situations for a given purpose. Thus, a creative series of comparative usability evaluations by Rolf Molich11 is noteworthy. A number of professional usability teams independently and simultaneously evaluated websites, Web applications, and Microsoft Windows programs. These studies shed light on usability practice, though they didn’t systematically compare usability analysis methods.

A further limitation is that researchers who have empirically investigated usability analysis methods12 have focused on single, simple measures (that is, the number of usability problems identified)—which is an inadequate measure, even of productivity. A review of five experimental studies comparing usability evaluation methods found that "small problems in the way these experiments were designed and conducted call into serious question what we thought we knew about the efficacy of various methods."13

Kasper Hornbaek14 conducted an excellent review of usability measures from 180 studies published in core human-computer interaction journals and proceedings. We note some of the issues of measurement validity and reliability that he identified:

- Approximately one quarter of the studies don’t utilize a measure of user performance effectiveness (for example, success and accuracy at achieving task goals, and error rate), and some studies confuse a measure of efficiency (such as resources used and time spent) with a measure of effectiveness.
- Measures of the quality of interaction are rarely utilized.
- Approximately one quarter of the studies don’t assess the outcome of the users’ interaction (that, the quality of the users’ work products).
- Measures are often interpreted directly when they actually have no simple or straightforward interpretation. (For example, how is the number of webpages visited a measure of usability? Is completion time an indicator of efficiency or an indicator or motivation?).
- Measures of learning and retention are rarely employed, despite being recommended in prominent textbooks.
- Measures of user satisfaction reinvent questions that can be found in available, validated questionnaires “... the diversity of words used in rating scales is simply astonishing” (p. 90).14
- Some studies conflate users’ perceptions (that is, judgments of the learnability of an interface) with objective measures (such as time needed to master an interface to a certain performance criterion).

We also note that Hornbaek selected the 180 studies from a larger set of 587 journal articles and conference proceedings papers. Some studies were deselected because they didn’t involve any systematic comparison (for example, of alternative interface designs); also, some didn’t report any quantitative measures and only recounted selected comments from informal user experiences.

While usability study can have a positive impact in product and service delivery, little controlled experimentation has been done to determine which usability methods provide the most significant impact.10,14

Toward Standards for Scientific Usability Analysis
Several methods developed by researchers in human factors, cognitive ergonomics, cognitive psychology, and cognitive systems engineering have a track record of success in eliciting requirements, analyzing tasks and goals, and evaluating performance. Indeed, rigorous methods for evaluating human performance and learning are more than 150 years old. Methods for what is today called contextual inquiry can be dated to the earliest days of industrial psychology.15 These venerable methods have been invigorated and adapted for task analysis for modern work systems and work contexts.16,17

All of these cognitive task analysis methods involve studying what happens when “users” interact with technology. The users can be engaged in a task walkthrough and can be asked questions. Measurements can be made of what they do and how they perform. Conclusions can be reached about what they learn and how they reason. This is called “end-user testing.”18 But there’s another path taken in the field of usability analysis. This is the path in which experienced usability professionals make their own evaluations and judgments, most commonly called heuristic
evaluation. Based on the experience of cognitive systems engineers and experimental psychologists—who have developed and applied methods for cognitive task analysis and performance measurement—what might be a good first stab at a description of best practice for scientific usability analysis?

We can take it as a foregone conclusion that no single method of usability analysis will be “best” for all circumstances. Methods must be adapted to the goals and focus of the analysis (for example, the software tool, interface, or webpage). It’s also a foregone conclusion that usability analysis projects should employ multiple methods and measures. Both of these aspects of methodology were recognized in the early days of expert systems, when attempts were made to compare methods for eliciting experts’ knowledge.

We believe an important step toward increased rigor in usability practice would be the systematic categorization and application of usability methods. Usability analysis methods can be described in terms of a handful of basic categories:

- **Retrospective** methods engage users (or an analyst) in the use of a tool, and then after some time at that, pause to reconstruct their reasoning while using (or learning to use) the tool or work method.
- **Concurrent** methods engage the user (or analyst) in the use of a tool—and at certain times they’re briefly interrupted and are presented with probe questions that elicit judgments concerning the tool or work method or statements about what they are thinking. While there may be some concern that interrupting the user might disrupt the flow of reasoning, research by experimental psychologists suggests that brief interruptions don’t make much difference.

- **Performance analysis** methods are focused on measuring actual performance, using such measures as time-on-task, goal achievement, and so forth. Ultimately, the goal for usability analysis is to reach conclusions about the quality of performance using some tool and the method it instills.

The retrospective, concurrent, and performance analysis methods are all focused on creating a cognitive decomposition. This is a description of user activity with reference to what they know, what they learn, and how they reason. Many different questions can be boiled out from the general question of “Is this tool any good?”

Many different questions can be boiled out from the general question of “Is this tool any good?”

Questions that are formative of retrospective methods include:

- What knowledge or previous experience did you rely on to use this tool? Did you rely on knowledge of similar tools (interfaces and so on)?
- At what points did you find yourself asking questions? Did you look for answers to those questions?
- Were you reminded of any previous experience?
- Does this tool (or interface) fit a standard or typical task you were trained to deal with?
- If your decisions and actions weren’t the best, what training, knowledge, or information could have helped?
- How would you describe the tool (or interface) to someone else?

Questions that are formative of the concurrent methods include:

- What are you looking for? Why are you looking for it?
- Why has this caught your attention and what will you do with it?
- What were your specific goals and objectives?
- What do you need to proceed further?
- Is there anything you are concerned, confused, or frustrated about?
- What did you decide to do next?
- What did you expect to happen next? To what degree were your expectations met?
- Did you imagine the events that would unfold?
- What mistakes might have occurred here?
- What were the alternative choices that you could have made?
- Were any alternatives rejected?
- At this point, what might be done by someone having less experience than you?

A primary methodological question regarding concurrent methods is when to interrupt and ask questions. On this, too, there are a number of alternatives. There’s no a priori methodological cookbook. A probe question might be asked whenever a user shows visible signs of frustration. A probe might be asked every other time the user enters information into a field, or clicks the mouse. Certainly, not all of the above questions are to be asked every time. Nor are these lists exhaustive.

The third class of methods is Performance Analysis. Many attempts at performance analysis in the field of usability employ single performance measures, such as the time taken to achieve specific primary task goals. There are many possibilities that can be adapted from the field of training...
and performance analysis.22 One of the more powerful performance analysis methods, particularly suited to the analysis of software tools, focuses on learnability.23

**Step 1.** Users perform until they reach some criterion of performance on a goal or a conjoint set of goals. A criterion can be based on time to achieve a goal or subgoal, or the number of attempts that must be made before a goal or subgoal is achieved, or even the number of attempts that must be made before a goal or subgoal is achieved twice in a row.

**Step 2.** There is a hiatus period of perhaps a few days.

**Step 3.** Users are once again placed into a task performance context, and again the researcher determines the number of attempts or the time it takes for the user to reach the criterion. Comparison of the number of attempts needed to reach the criterion performance can be used to estimate the original learning strength, to derive learning curves, and determine the extent to which the skill is retained.

We call out the study of learnability because it should be emphasized in usability studies. As Hornbaek noted, out of 180 studies “only five studies measured changes in efficiency over time” (p. 87).14 Table 1 summarizes our characterization of methods. This scheme helps us clarify the often-asked question of how many participants to involve in a usability study.24 There are really two questions here, one about the number of “usability experts” to be involved in an analysis and the other about the number of “representative users” to be involved.

If the participant-analysts are usability “experts,” the analysis should employ no fewer than five participants but need employ no more than seven. The rationale for 5–7 is that studies of the diminishing return on cognitive task analysis have shown that after about five to seven attempts decompositions by domain experts, there’s usually little more to be learned.25 Ideally, the “experts” who are doing the evaluation would be individuals who are proficient in the content domain (of the tool, website, interface, and so on) and are genuine experts in usability analysis. But for now, all we really have to go on is the finding about diminishing returns for experts doing task decomposition.

The low end accords with the “five users” benchmark set by Jakob Nielsen9 and Robert Virzi.26 They found that the percentage of usability problems that participants identified asymptoted at about 80 to 85 percent for about five participants. This was for a single task and measure—the identification of usability problems. We wonder if the same asymptote obtains for other tasks and measures. But more to the point, typical usability analyses use fewer than five participants. Even more to the point, Nielsen’s study involved “representative users,” and not usability experts. It’s hard to avoid seeing this as a rigor gap. What’s needed are more experiments, using certified experts, showing that you get the diminishing return at three (or some number fewer than five). Since no one has yet shown that, we choose to err on the side that would be taken by mainstream experimental psychologists, and advocate for no fewer than five. We’ve never seen five usability experts tackle the same problem (except in the Molich 2012 studies11). Given the evaluator effects that Rolf Molich and his colleagues found, how would usability professionals justify conducting studies with fewer than five usability professionals?

<table>
<thead>
<tr>
<th>Question</th>
<th>Analyst</th>
<th>Perspective</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>What do people desire?</td>
<td>“Expert” analyst</td>
<td>Self</td>
<td>Retrospective Concurrent Performance analysis Combinations of methods</td>
</tr>
<tr>
<td>Layperson</td>
<td>Self</td>
<td>Retrospective Concurrent Performance analysis Combinations of methods</td>
<td></td>
</tr>
<tr>
<td>Is it any good?</td>
<td>“Expert” analyst</td>
<td>Self</td>
<td>Retrospective Concurrent Performance analysis Combinations of methods</td>
</tr>
<tr>
<td>Layperson</td>
<td>Self</td>
<td>Retrospective Concurrent Performance analysis Combinations of methods</td>
<td></td>
</tr>
</tbody>
</table>

It's hard to avoid seeing this as a rigor gap. What’s needed are more experiments, using certified experts, showing that you get the diminishing return at three (or some number fewer than five). Since no one has yet shown that, we choose to err on the side that would be taken by mainstream experimental psychologists, and advocate for no fewer than five. We’ve never seen five usability experts tackle the same problem (except in the Molich 2012 studies11). Given the evaluator effects that Rolf Molich and his colleagues found, how would usability professionals justify conducting studies with fewer than five usability professionals? If the participant-analysts are representatives of some user population (“Layperson” in Table 1) then the usability analysis should be thought of as being a proper psychological experiment, in which case a sample size of about seven is certainly too small; it puts someone at risk of confounding any salient effects with individual differences. When conceived in reference to the traditional psychology laboratory, a sample size of at least 15 would be preferred. Interestingly, the Nielsen study used 15 participants as the benchmark for deciding what 100 percent meant in their analysis of diminishing returns. But we would take this a step further. As Kasper Hornbaek argued, there needs to be some sort of comparison, some manipulated independent variable. In the case
of participant characteristics, there should be some specific sampling plan that makes a sample type an independent variable, to ensure empirical rigor. For instance, one group of 15 participants might be individuals having little or no experience using the given type of software or interface, and a second group might be individuals having had some or considerable experience. The excellent work of Jeff Sauro and James Lewis offers direction on sample size decisions, and is a great resource for those interested in making informed decisions about sample size.

**Keeping Hope Alive**

The concerns about the methodology of usability study expressed by Hornbaek in 2006 remain pertinent. Rigor in usability studies could be improved if the following ideas were implemented:

1. Individuals who tout themselves as usability experts should provide up-front convincing empirical evidence that they are, in fact, experts.
2. Developers of software (tools, interfaces, and webpages) should conduct thorough analyses of user desirments prior to or during the initial prototyping, and thereby avoid needing to have a group of “representative users” rattle off design problems that the developers should have been able to anticipate and avoid in the first place.
3. Focus usability methodology on the analysis of actual performance and learnability rather than relying exclusively on satisfying through the use of questionnaires or surveys.

In 1988, Donald Norman characterized usability as “the next competitive frontier” (p. vi). With the commoditization of hardware and, increasingly, software, what might differentiate products would be how usable they are. We believe that this is true, but the lack of rigor in the application of usability practices has retarded acceptance of this truth; we advocate for usability analysis as an applied science motivated by the concepts and laws of human-centered computing.

**Acknowledgments**

Randolph Bia would like to gratefully acknowledge financial support from the University of Texas at Austin IC Institute.

**References**


Randolph G. Bias is a professor in the School of Information and director of the Information eXperience (IX) Lab at the University of Texas at Austin. Contact him at rbias@ischool.utexas.edu.

Robert Hoffman is a senior research scientist at the Institute for Human and Machine Cognition. Contact him at rhoffman@ihmc.us.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.

---

Expert Online Courses — Just $49.00

Topics:

www.computer.org/intelligent

IEEE INTELLIGENT SYSTEMS