Evaluation and usability as a practice area has diversified its approaches, broadened the spectrum of UX issues it addresses, and extended its contribution into deeper levels of product-development decision making. This forum addresses conceptual, methodological, and professional issues that arise in the field's continuing effort to contribute robust information about users to product planning and design.

David Siegel and Susan Dray, Editors

Clothing the Naked Emperor: The Unfulfilled Promise of the Science of Usability

Randolph G. Bias
Florida Institute for Human and Machine Cognition and University of Texas at Austin | rbias@ischool.utexas.edu

Philip Kortum
Rice University | pkortum@rice.edu

Jeff Sauro
Measuring Usability | jeff@measuringusability.com

Doug Gillan
North Carolina State University | djgillan@ncsu.edu

Here are two scenarios with which all usability professionals, be they in-house staff or external consultants, will be familiar:

• *Scenario A*

Development VP: "Usability support? We don't have the time or the money for that on this release."

Usability professional: "Thank you for your time and consideration."

• *Scenario B*

Development VP: "Usability support? Of course. We wouldn't think of shipping or cutting live without it."

Usability professional: "All right. We'll get started right away."

But here's a third scenario that may not be quite as familiar:

• *Scenario C*

Development VP: "Usability support? Of course. But tell me, why did you select those particular user-centered design methods?"

We suspect that most usability or user experience (UX) professionals would have an answer for this, but the answer would be based on personal experience, or perhaps on case studies that were only somewhat related to the current situation. Randolph Bias and Deborah Mayhew anticipated that although a cost-benefit analysis approach might be useful for deciding *if* to incorporate usability support, in the near future it would be used to discern *which* methods to employ [1]. Well, not yet. This article is intended as a call to arms, hoping to justify and motivate a discussion about, and action on, an empirical basis for the selection of appropriate and optimal usability engineering methods.

Design is hard. There exist excellent individual designers—people who are creative and seemingly inherently in touch with what people will like, will pay for, and will be able to use. But most of us are not like that. And given the availability of tools that allow almost anyone to build a website, the Internet is replete with evidence that the vast majority of us are not excellent designers. Insert your own favorite example of a bad Web design here. A proven solution for this problem is a user-centered design (UCD) approach, wherein user data is gathered at various points in the design/development process to maximize, empirically, the usability of designs.

As defined by the International Standards Organization (ISO), usability is 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" [2]. Usability engineering entails the gathering of user data to inform (early on) and validate (later in the design/development cycle) the goodness of website designs and other software user interfaces (UIs).

Often situated as part of a UCD approach to software development, the pursuit of good usability involves different methods designed to collect that user data across the entire lifecycle of the products. Early in the design of a UI, methods such as task analysis and contextual inquiry are employed to gather and prioritize product requirements and help drive early designs. In the middle of product development, formative usability evaluations are conducted via methods such as paper-and-pencil tests, prototype tests, heuristic evaluations, pluralistic usability walkthroughs, and cognitive walkthroughs. As product development winds down, summative usability evaluations likely entail end-user lab testing.

Through these various usability evaluations, products and websites approach maturity, and improved usability, *before* a product ships or a website cuts live. Thanks to these evaluations, design decisions are not left up to the loudest or most well-liked product designer/developer, but rather are made empirically, based on user data. And thus products developed via a UCD approach employing multiple usability evaluations have been shown to be easier to use, to lead to higher customer satisfaction, and to lead to superior user experience than products designed without the benefit of user data.

In his excellent chapter on usability evaluation, Gilbert Cockton argues for an expansion of the concerns of usability to embrace "context of use" [3]. For our purposes, we wish to define our discipline broadly, incorporating Cockton's call for attention to *quality in use*. Concerned with the possible confounds inherent in comparing usability evaluation methods, Cockton suggests, "Consider a simple comparison of heuristic evaluation against user testing. Significant effort would be required to allow a fair comparison." This is true. Here we argue for expending that effort to good potential return.

In too many instances, Web and other software designers and developers have yet to recognize the importance of systematic, professional application of usability engineering methods [4]. Poorly designed HCIs cost time, money, and in the worst cases, lives. Given the state of software development, some usability engineering is almost always cost effective. Why then is usability given so little emphasis in the software development cycle? As Fiora Au et al. note, "There are many challenges and issues associated with traditional usability testing methodologies, … [contributing] to the industry's general reluctance to integrate usability testing as an essential activity on par with functional testing, despite its importance. Instead, it is often considered as a 'nice-to-have' reserved for larger projects with generous budgets" [5].

Conspiring with this lack of emphasis on usability is the positively accelerating need for it, as stimulated by the advent and popularity of the Internet and the associated growth of the number and types of tasks people carry out online. Because of this, usability guru Jakob Nielsen says "we will need to scale up by a factor of 100 in terms of available personnel" [6]. Thus, the tide is turning.

The popularization of usability is evident in the proliferation of how-to books in the field of usability and usability testing over the past decade (e.g., [7,8]). Although there is this increase in usability evaluation being accepted routinely as an important component in contemporary UI design and development, usability is more craft than science. While there is much activity in the world of Web and software development practice in the arena of usability, there is much less attention paid to the potential science of usability. Historically there has been attention paid to the application of perceptual and cognitive psychology to the design of UIs. There are also recent efforts to add psychophysical measurement to the usability analyst's tool belt, in the form of eye movements, galvanic skin response, heart rate, and fMRI. But what of empirically determining which usability engineering methods are best applied in a certain situation?

Photograph by David Rout

Hasn't This Been Done Already?

No. Although if you thought so, you wouldn't be alone.

When presented with the prospect of some research on usability engineering methods, Penn State professor of information sciences and technology and CHI Lifetime Achievement Award winner Jack Carroll opined, "That old saw?" but was quick to acknowledge that whereas in the latter 1980s and 1990s there were various studies on UCD methods, "the considerable amount of effort, debate, and passion that was poured into this topic did not resolve the issue of methods" [9].

Andrew Dillon, dean of the University of Texas at Austin School of Information and current president of the Association for Information Science and Technology, reports that in 1995 he proposed to a corporate funding agency an "evaluation comparing multiple methods against real world problems identified from corporate users, aiming to show if various methods could predict the actually occurring problems" [10]. His proposal was not funded because the reviewer thought "it had already been done."

University of Colorado computer science professor Clayton Lewis offers his recollection of a CHI workshop in Amsterdam in 1993. The workshop concluded with the participants selecting a particular design task (design of a UI for an ATM) and a metric (time to complete the transaction), and then "taking an oath" [11] to incorporate that task into any similar research in the future, with the notion being that eventually there would accrue a collection of data about certain methods applied to this one problem. The approach Lewis describes could provide real-world data relevant to answering our question of which method is best applied when. As far as Lewis knows, the work wasn't done (and he adds that certainly he didn't do it).

So, although usability methods have widespread application, the selection and measurement of these methods is still an inexact science—so inexact that it is based more on opinion and tradition than science. The methods that usability professionals believe to be most important are not always the ones they choose to use in actual practice. Perhaps most noteworthy in the realm of the pursuit of empiricism regarding usability engineering methods is a creative and valuable series of *comparative usability evaluations* (CUE), organized and run by Rolf Molich (see [12]), in which multiple professional usability teams evaluated the same set of interfaces. They found there was significant variance in the kinds of usability problems found across these teams (evaluator effects), to a degree that surprised many in the usability profession at the time. Molich's nine CUE studies shed empirical light on usability practice, but they do not comprise a systematic comparison of the relative value of various usability engineering methods, nor do they provide guidance as to when to apply the methods. Importantly, Molich and others who have tackled an empirical examination of usability engineering methods have tended to focus on simple output counts (i.e., the number of usability problems identified) as a measure of productivity.

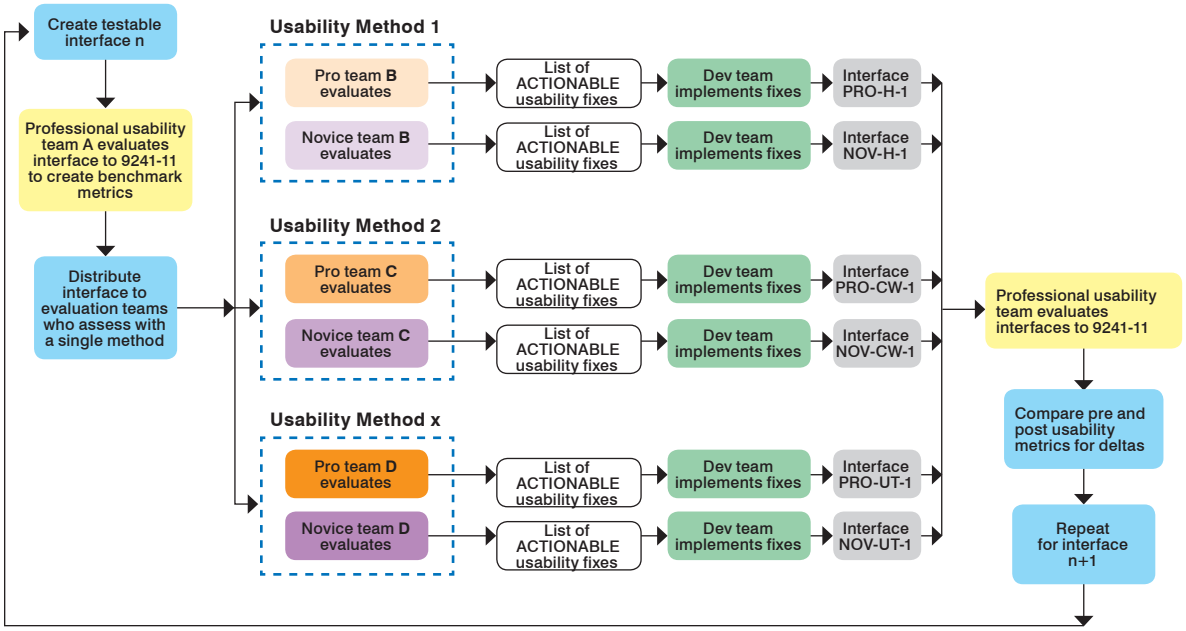
Yes, usability evaluation tends to yield high ROI, but which method or methods will yield the highest ROI for a particular development project? Yes, usability analysts are being welcomed onto more and more design/development teams, but for any particular project do we need a professional with much experience or might a tyro do? Yes, a heuristic evaluation is likely a satisfactory method to employ in evaluating an e-commerce site, but what about for a medical informatics site, the usability of which might mean the difference of hundreds or thousands of people seeking the medical attention they need? These are the sorts of questions we could start to answer only with a systematic, empirical comparison of the efficacy of various usability evaluation methods.

One Approach

We believe what is needed is a systematic approach to determine the relative efficacy of common usability assessment methods using end-user performance metrics, rather than assessment-method output counts, as the defining measure. One possible approach, detailed in Figure 1, consists of:

- selecting and developing representative interfaces for use such a study;
- performing benchmark usability testing of those interfaces;
- evaluating the interfaces using a range of usability methods employed in isolation with both expert and novice teams;
- creating specific, actionable usability outputs as a result of those evaluations;
- modifying the interfaces to reflect those specific outputs; and
- determining the impact of those modifications through a second set of benchmark tests.

One key to the proposed method is that it uses the ISO measures of effectiveness, efficiency, and satisfaction to establish an initial benchmark for an interface and then measures the changes to these metrics after a specific usability method



▸ Figure 1. One empirical approach to comparing usability engineering methods.

had been employed and its outputs implemented. Clearly, this would not be easy, but the benefits could be enormous. For the usability team to continue to compete for resources, especially as software engineering methods improve and marketing teams get more sophisticated, we will need to be able to point to measurements that show the ROI for the dollar and hour spent on our particular, chosen UCD methods is higher than that for other uses of that time and money.

Important considerations in the proposed approach. While the framework of the approach we propose is simple in concept, the devil, as they say, is in the details. There are several important implementation considerations.

• *Selecting appropriate interfaces that represent the entire design space.* Whether they are addressing complex computer-based military systems or simple consumer-grade products that have only physical controls, most usability methods and metrics are sufficiently robust that they can be employed without regard to the kind of interface being tested. This portability and robustness of usability methodology means we are relatively unconstrained in our selection of interfaces used to test the efficacy of these different methods. That said, the interfaces must be of sufficient complexity to actually have usability issues, be easy to build and replicate, and be easy to modify based on the testing recommendations. For the purposes of this research, then, we would suggest Web-based interfaces as the appropriate interface platform for three important reasons. First, the ubiquity of Web interfaces simplifies the recruitment and testing of participants. Second, because of the ubiquity of the platforms over which the Web is delivered, we would be able to employ state-of-the-art data-collection methodologies that allow for excellent participant representation across geography, experience levels, and economic circumstances. Third, Web-based interfaces are easily modifiable by readily available professionals.

• *Performing rigorous benchmarks with sufficient power.* Reliable, objective usability metrics must be established for each interface so they can serve as benchmarks that can be compared to metrics measured on the final, modified interfaces at the end of the projects. In order to ensure that there is a good chance of detecting differences between each iteration and condition in the study, a sufficient number of users would need to be tested to afford enough statistical power (see [13] for review of statistical power) to make these comparisons with any degree of confidence. Unlike formative field testing, these benchmark tests may require hundreds of users for each interface.

• *Identifying which methods to evaluate.* Once the interfaces have been benchmarked for usability using ISO-9241-11 metrics, the actual

assessment of the different usability methods could begin. There are many usability assessment techniques described in the literature. Christopher Nemeth [14] details 36, whereas Neville Stanton and his colleagues [15] identify 87. Further, there are many methods that are simply combinations of some of the fundamental methods described in the literature. How do we choose among this relatively large number of potential usability assessment methods? A good starting point would be to select those methods that are in most common use. Should these results prove promising, the research could eventually be expanded to include secondary and less commonly used methods in the assessment.

• *Establishing a clear and unambiguous output for the evaluation method.* The form of the output from these usability evaluations is of the utmost importance. Usability reports are often nebulous and nonspecific, encouraging the designers with ambiguous statements such as "the buttons on screen 12 are hard to use and should be modified to increase success rates." While accurate, these kinds of statements do not provide sufficient direction for unskilled development teams to take action, and they provide tremendous latitude on the actual fix that is implemented. We believe that the outputs of these usability evaluations need to be single, actionable modification instructions. Such an output would provide a specific instruction that could be implemented in an interface on a single interface element, with no interpretation required on the part of the person performing the implementation. These modification instructions could also serve as the basis for further assessments of the efficacy of the different methods. For example, raw counts along with confidence intervals could be analyzed to see if different methods yield significantly different numbers of correctable issues. They could also be categorized to more fully understand if different methods are better at identifying certain types of errors. For example, it might be the case that evaluation method A is better at identifying global navigation errors than evaluation method B.

• *Controlling the implementation of the actionable items from the evaluation.* Controlling the implementation of the modification instructions is also important in retaining a degree of experimental control. The development teams would need to be instructed to carry out the modification instruction exactly as delivered, seeking clarification on intent when necessary. Once the changes were implemented, regression testing would be conducted by the developer to ensure the fix had been implemented in a technically sound and stable fashion. The design would be reviewed to ensure the exact fix recommended by the modification instruction had indeed been implemented.

• *Ensuring consistency of the post-evaluation process.* The final benchmark testing would need to be performed in a manner identical to that of the initial benchmark test, since the improvement (we hope) in performance that each method has provided is the key variable of interest.

And So . . .

The results of this kind of research would have an immediate transformative impact on the practice of professional usability. Usability professionals would be able to select methods that have known impact, and businesses would be able to determine if the application of usability is effective. The research would also set the stage for understanding a broader suite of UCD methods, particularly if we find there is a large difference between the methods explored in this research and a robust level-of-experience interaction. The general method we have described in this research would be able to be used as a template to explore the full range of UCD methods; other researchers would be able to add to a global UCD methods efficacy table incrementally as they generate objective metrics for the entire suite of employed methods.

The kinds of results we describe would be invaluable for anyone involved in product design. By collecting objective performance data in addition to the usual (from historical, nonsystematic investigations of this question) quantity-of-output data, this line of research would be able both to better parse the relationship between these two measures and examine how these relate to the severity of the identified errors, as well. The findings of this project would also contribute explicitly and intentionally to the connection between usability theory and practice. The publications spawned by this study would inform both usability practice and usability education. Finally, computer users around the world would ultimately benefit from gaining usable access to functionality that would have been otherwise too hard for them to discover and use, thanks to the subsequently improved optimization of applied usability practice, as driven by usability science.

This research would also help to explore the novice/expert division in the application of the different UCD methods. Some usability professionals and other software developers assert that

"some usability is better than no usability" and that the so-called discount methods can have value even when applied by those with no training or demonstrated skill in the art. This research would help us better understand the relationship between the skill level of the practitioner and the requisite skill level for the method being applied.

We are not alone in thinking that there is work yet to be done in this arena. Commenting on Cockton's chapter on usability, David Siegel writes, "I do agree wholeheartedly with [Cockton] when he points out the many factors that can complicate the process of interpreting usability findings due to this lack of a cookbook of infallible methods and the presence of many confounds. These issues argue for the need for greater professionalism among usability practitioners, not for the downgrading of the profession or marginalizing it on the periphery of the product development team. Professionalism requires that practitioners have expert understanding of the limitations of methods, expertise in modifying them to address different challenges, the dedication to continually advance their own processes, and the skill to help drive the evolution of practice over time" [16].

Stephanie Rosenbaum offers another commentary on the Cockton chapter: "Thus a key element of usability evaluation is deciding when to employ guidelines and inspection (user-free methods) and when it's critical to perform empirical research such as usability testing or contextual inquiry with the target audience. Planning the activities in a usability evaluation program—and the schedule and budget appropriate to each—is central to the responsibilities of an experienced and skilled usability practitioner. An encyclopedia chapter on usability evaluation should help readers understand this decision-making process" [17].

But such a chapter is impossible today because the necessary research has not been done. We anticipate pursuing this course of study, as we believe that such empiricism is needed to make Emperor Usability's raiment actual and glorious. Will you join us? What other approaches would you propose? We look forward to participating in this dialogue with the usability/UX professional community, in the interest of improved usability/UX practice and the advancement of the science of usability.

Acknowledgments

Randolph Bias would like to gratefully acknowledge financial support from the University of Texas at Austin IC² Institute.

ENDNOTES:

1. Bias, R.G. and Mayhew, D.J., eds. *Cost-Justifying Usability: An Update for the Internet Age, Second Edition.* Morgan Kaufman, San Francisco, 2005.

2. ISO. *Ergonomic requirements for office work with visual display terminal (VDT's)–Part 11: Guidance on usability* (ISO 9241-11(E)). Geneva, Switzerland, 1998, 2.

3. Cockton, G. Usability evaluation. In *The Encyclopedia of Human-Computer Interaction, 2nd Edition.* M. Soegaard and R.F. Dam, eds. The Interaction Design Foundation, Aarhus, Denmark, 2013; http://www.interaction-design.org/encyclopedia/usability_evaluation.html

4. Nunes, N. What drives software development: Bridging the gap between software and usability engineering. *Human-Centered Software Engineering, Human–Computer Interaction Series I,* (2009), 9–25.

5. Au, F.T.W., Baker, S., Warren, I., and Dobbie, G. Automated usability testing framework. *Proc. Ninth Australasian User Interface Conference vol. 76.* B. Plimmer and G. Weber, eds. ACS, 2008, 55–64.

6. Nielsen, J. Usability for the masses. *Journal of Usability Studies 1,* 1 (2005), 2–3.

7. Reiss, E.L. *Usable Usability: Simple Steps for Making Stuff Better.* John Wiley and Sons, Hoboken, NJ, 2012.

8. Tullis, T. and Albert, W. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics.* Morgan Kaufman, Burlington, MA, 2008.

9. Carroll, J. Personal communication. June 29, 2013.

10. Dillon, A. Personal communication. June 19, 2013.

11. Lewis, C. Personal communication. June 19, 2013.

12. Molich, R. CUE - Comparative usability evaluation. 2012; http://www.dialogdesign.dk/CUE.html

13. Sauro, J. and Lewis, J.R. *Quantifying the User Experience: Practical Statistics for User Research.* Morgan Kaufmann, Burlington, MA, 2012.

14. Nemeth, C.P. *Human Factors Methods for Design: Making Systems Human-Centered.* CRC Press, New York, 2004.

15. Stanton, N.A., Slamon, P.M., Walker, G.H., Baber, C., and Jenkins, D.P. *Human Factors Methods: A Practical Guide for Engineering and Design.* Ashgate, Burlington, VT, 2005.

16. Siegel, D.A. Commentary on [3]. In *The Encyclopedia of Human-Computer Interaction, 2nd Edition.* M. Soegaard and R.F. Dam, eds. The Interaction Design Foundation, Aarhus, Denmark, 2013; http://www.interaction-design.org/encyclopedia/usability_evaluation.html

17. Rosenbaum, S. Commentary on [3]. In *The Encyclopedia of Human-Computer Interaction, 2nd Edition.* M. Soegaard and R.F. Dam, eds. The Interaction Design Foundation, Aarhus, Denmark, 2013; http://www.interaction-design.org/encyclopedia/usability_evaluation.html

ABOUT THE AUTHORS

Randolph Bias is a professor in the School of Information at the University of Texas at Austin, currently a visiting scientist with the Institute for Human and Machine Cognition, and principal with The Usability Team. He thinks usability/UX is the third leg (with functionality and schedule) of the design/development stool.

Philip Kortum is an assistant professor at Rice University in Houston, TX. His primary interests are the research and development of highly usable systems in the voting and mobile computing domains and the characterization of measures of usability and usable systems.

Jeff Sauro is a Six-Sigma-trained statistical analyst and pioneer in quantifying the user experience. He manages Measuring Usability LLC in Denver. He has published four books, including the recent *Quantifying the User Experience: Practical Statistics for User Research.* He has worked for GE, Intuit, PeopleSoft, and Oracle and has consulted with dozens of Fortune 500 companies.

Douglas J. Gillan is professor and head of the psychology department at North Carolina State University. His training in psychology focused on biopsychology and cognition. He has worked in both industry and academia on information visualization and human-technology interaction. He is a fellow of the Human Factors and Ergonomics Society.

DOI: 10.1145/2527299