Changing Perspectives on Evaluation in HCI: Past, Present, and Future

Craig M. MacDonald
Pratt Institute
School of Information and Library Science
144 West 14th Street, 6th Floor
New York, NY 10011 USA
cmacdona@pratt.edu

Michael Atwood
Drexel University
College of Information Science and Technology
3141 Chestnut Street
Philadelphia, PA 19104 USA
mea23@drexel.edu

Abstract
Evaluation has been a dominant theme in HCI for decades, but it is far from being a solved problem. As interactive systems and their uses change, the nature of evaluation must change as well. In this paper, we outline the challenges our community needs to address to develop adequate methods for evaluating systems in modern (and future) use contexts. We begin by tracing how evaluation efforts have been shaped by a continuous adaptation to technological and cultural changes and conclude by discussing important research directions that will shape evaluation's future.

Author Keywords
Evaluation; usability; user experience

ACM Classification Keywords
H.5.2. User Interfaces: Evaluation/Methodology.

CHI'13, April 27 – May 2, 2013, Paris, France.

Introduction
Evaluating interactive systems has always been central to Human-Computer Interaction (HCI), in that it is one of the first skills we teach students and is a staple for many practitioners. Not only is evaluation listed as one of the fundamental foci (along with design and implementation) of the HCI curriculum [35], it is also one of the three principles of user-centered design [24]. Not surprisingly, evaluation is listed as a core activity in nearly every design model [40], [64], with some even making it the center of the design process [30]. In short, design and evaluation are closely related activities that support and inform each other.

Despite this clear connection, there is some debate (e.g., [14] [49] [73]) about the value and role of evaluation in HCI. The intention of this paper is not to assess the merits of arguments for or against evaluation, but rather to start a meaningful dialogue about how to address the challenges facing evaluation in modern contexts. This will be done by first examining how evaluation has evolved into its current state and then by discussing how we can transform this current state into a more preferable one.

The History of Evaluation
In order to accurately assess the current state of evaluation in HCI and develop a vision for the future,

it's necessary to first understand how and why we evaluate the way we do today. In their review of the history of evaluation in HCI, Kaye and Sengers [43] identified four phases of evaluation by focusing on who was doing the evaluation and what role users played in the evaluation. This approach provided a valuable framework for illuminating the changes to evaluation methods over time, but we hope to build on this framework in two ways. First, we will broaden the discussion by considering how changes in technology and use contexts influenced the choice of evaluation objectives (and, thus, evaluation methods), which will provide greater insight into how our evaluation methods became what they are today and highlight important lessons that may inform how we adapt to future changes. Second, their review was written at a time when experience-focused approaches were just beginning to emerge. Since then, evaluation methods for assessing experiential goals have taken the field in exciting new directions and we will identify a fifth phase of evaluation that will help to put these developments in their appropriate historical context. The resulting five phases are delineated by the primary evaluation objectives. Although these phases are presented as discrete time periods, the years should not be taken literally as there was substantial overlap as evaluators transitioned from one objective to the next.

System Reliability Phase (1950s and before)

Although many machines invented in the early 20th century could be considered "computers" in a technical sense, the modern computing age really began in the 1940s with the development of the ENIAC, which is now widely recognized as the first electronic computer ever invented [22]. The ENIAC and other early computer systems were incredibly large machines whose operation required the complicated manipulation of switches, lights and plugs. Because operating these machines required a high degree of technical expertise, the users (and evaluators) were engineers and other highly trained professionals. Though only a few were in operation, computer systems during this period were used primarily to perform complex calculations on large quantities of data (e.g., calculating ballistic trajectories or processing census data). As one evaluation report noted, the major concerns of evaluation were minimizing system fault time and quickly repairing errors because the computer was "entirely judged by its reliability in operation and the ease of maintenance" [61]. Therefore, evaluation efforts tended to focus on system reliability, mostly in terms of how long it would function without failure [43].

System Performance Phase (1950s-1960s)

Over time, computer systems became increasingly stable and reliable and although they were still relatively large, they were beginning to shrink considerably compared to their predecessors. Researchers and scientists also invented new methods of operation, such as magnetic tape, punch cards, light guns, and, eventually, keyboards, which led to the development of early programming languages (e.g., FORTRAN and COBOL). These languages promised significantly more power and flexibility but only with sufficient training, which meant that users (and evaluators) shifted from engineers to programmers and computer scientists. Since the cost of computer systems was a primary concern, the emphasis shifted away from how long a system would perform to how quickly it would perform [43]. A common form of evaluation studies during this time period were called "acceptance tests" in which evaluators estimated the

minimum performance time necessary for a computer to be "economically competitive with conventional tabulating-card methods" [1]. As a result, the tests were designed to evaluate how long it would take the system to process large amounts of data with minimal down time and minimal errors [53]. Other types of evaluations included simulations that evaluated computer systems for "safety and adequacy of performance" [41] or throughput, turnaround, and availability [43], all of which focused on system speed and system performance.

User Performance Phase (1960s-1970s)

By the late 1960s, large-scale batch-processing machines were being challenged by more expensive but supposedly more efficient time-sharing machines. There was significant debate in computer circles about the efficacy and value of time-sharing systems, which led Grant and Sackman to conduct an evaluation of the two types of systems in what they described as "a pioneering effort in the collection of performance data…under controlled conditions" [25]. Not surprisingly, a focus on users became much more prevalent as time-sharing systems grew in popularity and people started using computers for non-programming tasks (e.g., text editing). The introduction of "non-specialists" into the equation was itself a major shift because it forced evaluators to be more interested in evaluating the speed of the user rather than the speed of the system [43]. It is no coincidence that the field of HCI began to emerge as a discipline during this time period [21].

An increased interest in users and their performance brought a new type of evaluator to the field: experimental psychologists, who further popularized the use of laboratory-based user studies [21] that typically focused on enhancing worker productivity through the use of performance-based metrics. For example, English, Engelbart, and Berman [19] reported on a study evaluating the efficacy of input devices and found that the mouse was the most effective method effective for selecting bits of text on a computer screen[1]. Their preferred metrics were speed, ease of learning, error rate, accuracy, and satisfaction. A few years later, Sime, Green, and Guest [69] reported the results of an evaluation of programmer performance in which they used task completion rate, task completion time, and number of errors. In summary, during this time period there was a significant shift system performance to user performance.

Usability Phase (1980s-2000s)

The 1970s saw tremendous improvements in computer processing speed, innovations like the graphical user interface (GUI) and the WIMP interaction style [50], and drastic reductions in the size and cost of computers. By the early 1980s, the release of powerful software that could be used without extensive training led to a sudden increase in the number of novice users, which challenged designers to develop systems that could be used with minimal training and support. Thus, evaluators started to develop methods for evaluating the usability [24] or "ease of use" of computer systems because "if a system is not easy to learn, it [would] not be used" [23]. Accordingly, evaluation efforts were

[1] The mouse was famously introduced at the 1968 Fall Joint Computer Conference when Douglas Engelbart demonstrated the NLS system (commonly referred to as "the mother of all demos"). Video of the demonstration has been archived by Stanford University and is available at http://sloan.stanford.edu/mousesite/1968Demo.html

expanded to encompass aspects of learnability and ease of use in addition to speed and efficiency. Early on, many of these efforts focused on evaluating the ease of use of text editors (e.g., [18] [63]) but it eventually expanded to include other types of software.

It was also during this time period that GOMS models (goals, operators, methods, and selection) were created to develop models of human performance [12]. Researchers also began formalizing the process of laboratory-based user testing with the "think aloud" method [47]. Quantitative user testing methods were also developed during this time and were based on the assumption that systems were more likely to be usable "if the design objectives are closely tied to empirical definitions of desirable user performance" [11]. The three recommended dimensions of user performance included learnability, throughput, and attitude [5], which eventually evolved into the five metrics commonly used today: time to complete tasks, error rate, accuracy, task completion rate, and satisfaction [38] [67]. Interestingly, these five metrics are nearly identical to the metrics used in the evaluation studies from the 1960s and 1970s (e.g., [19] [69]), which highlights the fact that user performance has always been a core aspect of usability. Questionnaire-based approaches were also popularized [65].

In the 1990s, researchers developed methods such as the heuristic evaluation [56] and the cognitive walkthrough [48] as "discount" methods aimed to replace empirical user observation with the knowledge and expertise of usability professionals. The rise of the Web and the proliferation of web-based interactive systems further increased the visibility of the usability profession and also led to an influx of new professionals interested in improving the usability of the web. Throughout this time period, usability (specifically ease of learning and ease of use with an emphasis on performance) was a core evaluation goal.

User Experience (UX) Phase (2000s-Present)
Over the last decade, personal computing, social computing, mobile computing, and cloud computing have drastically altered the contexts in which people use computers [9], leading to the emergence of user experience (UX) as "a new paradigm" for design and evaluation [4]. As use contexts have broadened and technologies have become more pervasive, designers and evaluators recognized the importance of considering the "non-utilitarian" aspects of using computers, which shifted the focus from task-based performance to user affect and the value of computer interactions in everyday life [46] [71]. To put it simply, a UX perspective emphasizes "designing for pleasure rather than for absence of pain" [33].

A major challenge facing evaluators is the lack of a shared conceptual framework for UX, although several models have been proposed. For example, Hassenzahl [32] argued that a product has pragmatic attributes (e.g., a product's ability to help users achieve behavioral goals) and hedonic attributes (e.g., a product's ability to evoke feelings of pleasure, allow for self-expression, and provoke memories). Similarly, Norman [58] advocated for three levels of "emotional design" that consist of visceral, behavioral, and reflective experiences. Likewise, Forlizzi and Battarbee [20] described three types of user interactions that influence their experience: fluent, cognitive, and expressive. From these models, it seems clear that a UX evaluation approach should address both the

hedonic and pragmatic dimensions of system use. However, it is still common to associate non-instrumental or hedonic goals with UX and instrumental or pragmatic goals with usability [64]. This has led to a situation where most methods that fall under the category of UX evaluation focus solely on hedonic attributes [72], with usability evaluation methods used to capture pragmatic attributes (e.g., performance).

The Present State of Evaluation

In this section, we will more closely examine the current state of evaluation by discussing the different evaluation techniques or approaches in four broad categories: user testing methods, inspection methods, traditional research methods, and field methods [16].

User Testing Methods

User testing is by far the most popular evaluation method. Although traditionally done face-to-face in a controlled setting, a more recent variation includes testing with remote users either asynchronously [3] or synchronously [51]. User testing with think aloud is widely considered the "gold standard" [39] for usability evaluation (i.e., pragmatic goals). As previously mentioned, common approaches include collecting a variety of usability metrics [38], [28] or distributing post-use questionnaires regarding usability attitudes and perceptions (discussed later). More recent variations include using eye-tracking [36] or mouse-tracking [55] technology to highlight "hot spots" or problem areas. There are also a number of variations on user testing that focus on the hedonic dimension. As with usability, a common approach is to distribute a post-use questionnaire but other techniques include "emocards" [17] and personal meaning maps [8]. Other methods require the use of technology and advanced algorithms to automatically measure users' affective states from their facial expressions [70] or physiological measurements [52].

Inspection Methods

Inspection methods aim to replace user involvement with expert judgment and have emerged as a viable alternative to user testing. Early usability inspection methods included guidelines and checklists [62] but the most common methods are expert reviews and walkthroughs [37] such as the heuristic usability evaluation [56] and the cognitive walkthrough [48]. More recent variations include the cognitive walkthrough for the web [7] and the activity walkthrough [6]. Although inspection methods have become a widely used tool for evaluating systems from a pragmatic perspective [16], we are unaware of any inspection methods for measuring hedonic attributes.

Traditional Research Methods

Traditional research methods include surveys, interviews, and focus groups. It should be noted that these methods are typically used in conjunction with user testing (e.g., as a follow-up to a user testing session). As mentioned previously, a common approach to measuring pragmatic attributes is via a post-use questionnaire. Several usability questionnaires have been developed, validated, and popularized, including the Software Usability Measurement Index (SUMI) [44] and the System Usability Scale (SUS) [10]. Interviews and focus groups can be used in place of (or in addition to) questionnaires when the scenarios are more complex or when evaluators are interested in more detailed feedback on specific usability problems [16]. Post-use questionnaires are actually the most the most widely used approach for measuring hedonic and are

used to measure a variety of hedonic attributes [4] including enjoyment [34], engagement [59], and visual aesthetics [45]. Interviews and focus groups can be used to further probe these issues (e.g., [31]).

Field Methods

Field methods represent the tools and techniques for evaluating interactive systems in naturalistic settings. Common field methods for evaluating pragmatic attributes include behavioral observations, collages or artifacts, and log analysis [16]. User diaries are becoming increasingly popular as a way to catalog usability problems and elicit feedback on usability issues based on real world usage [60]. An emerging method is the living laboratory [13], which includes methods like A/B testing [2]. Of course, user diaries have also been used to probe for users' emotional responses to product usage [42] and A/B tests can be conducted to test whether certain aesthetic or experiential elements are more successful at engaging users. Behavioral observations in natural settings can also focus on experiential attributes, such as pleasure, engagement, or fun [15].

The Future of Evaluation

Despite the variety of methods available for evaluating usability and UX, we cannot ignore the fact that many evaluators remain dissatisfied with the methods available today. Critiques of UX evaluation methods seem to be largely a matter of novelty but critiques of usability evaluation are more direct, ranging from questions about their reliability [54] and validity [26] to their potential harmfulness in certain situations [27]. These criticisms are well known and we will not re-hash them here. Instead, we will describe five potential research directions that will help shift our perspective on evaluation and put us in a better position to address current and future challenges.

Create a More Holistic Vision for UX Evaluation

Most frameworks for understanding UX stress the importance of considering both instrumental and non-instrumental goals (e.g., [32]), but few evaluation methods are designed to capture this holistic perspective [4]. In short, we are still striving for evaluation methods that can quickly and accurately assess whether interactive systems are truly "useful, usable, and desirable" [68]. Thus, we believe the exploration of more holistic UX evaluation approaches – which address both pragmatic and hedonic dimensions and are applicable in real world settings – represents a valuable research direction.

Develop Inspection Methods for Hedonic Attributes

While there has been some debate about the value of usability inspection methods [66], they were never intended to be a replacement to user testing but rather a viable alternative that requires less time and fewer resources while still producing meaningful results. The important characteristic of inspection methods is that they do not involve users, but we have yet to uncover any hedonic evaluation methods that offer this feature. We believe this represents an exciting research opportunity. Developing an inspection method aimed at assessing hedonic attributes – such as aesthetics, engagement, or enjoyment – would be a great benefit to the field, particularly in situations where time and financial resources are limited.

Examine the Core Skills of Evaluation

The history of evaluation has repeatedly demonstrated how evaluation skills have changed over time to adapt

to new technologies and use contexts. Thus, there are a number of unanswered questions about the skills and competences required to be an effective evaluator, how those skills/competencies compare to those of design, and, perhaps more importantly, how educators can equip future HCI professionals with these necessary skills. Examining and defining these core skills would not only enhance our understanding of what evaluation is, but also how it should be done. Knowledge gained from these efforts can then be used to reshape education programs and curricula to focus on the requisite skills and competencies for planning, managing, and conducting effective evaluations.

Investigate Informal Approaches to Evaluation
The implied methodological rigidity of evaluation is necessary for evaluation efforts to maintain a semblance of reliability and validity, but it also suggests that "evaluation" refers solely to the formal, structured process of assessing the quality of a design. However, informal evaluation occurs frequently throughout the design process when, for example, bad ideas are abandoned and good ideas are adopted. Thus, we feel it is worth exploring the efficacy of more informal and subjective evaluation approaches that reflect the "fuzziness" of real design problems. Understanding these design decisions from an evaluation perspective may yield important insights into how designs are formed and the role of expertise and judgment in making design decisions.

Learn from Evaluation in Practice
Because it provides a searchable public record, we know quite a bit about evaluation in HCI research. But, with a few exceptions (e.g., [57]), we know very little about how evaluation is conceptualized and conducted by HCI practitioners. In the practitioner world, where time and/or budgetary restrictions are commonplace, it makes sense to gain a richer understanding of the tradeoffs evaluators are forced to make and how these decisions impact their choice of evaluation methods and, ultimately, the effectiveness of those methods. Understanding these processes can help us better understand the goals of evaluation in practical settings, which will also help us focus on designing methods that are more adaptable and applicable in these settings.

Conclusion

There is a long tradition of research on evaluation in HCI. Since the beginning of the computer age, designers and evaluators have continuously been forced to adapt to the rapidly changing use contexts resulting from technological innovations and their related social and cultural expectations. Today, cloud-based systems, social, mobile, and ubiquitous computing, and intelligent, adaptive technologies have once again changed the nature of interactive systems and the ways they are used, which necessarily changes the nature of evaluation. Consequently, despite the plethora of methods and approaches available to us, we simply can't put evaluation into the "solved problems" category.

In this paper, we reviewed our history, examined our present, and outlined a vision for the future of evaluation in HCI. We offered several potential research directions that we believe will give us the tools we need to meet the known challenges of the present and the unknown challenges of the future. But we, as a community, need to work collectively to develop, first, a better understanding of these challenges and, second, a set of reliable methods aimed at addressing

them. Facing these challenges will require a strong, dedicated community of researchers and practitioners working together to create a culture in which evaluation is seen not as a chore but as an integral part of design: a difficult, but rewarding, activity that supports and enhances the process of designing products that adapt and respond to users' pragmatic and hedonic needs.

Acknowledgements

We thank the anonymous reviewers for providing valuable feedback and helping us to clarify some major points and fill in some important gaps.

References

[1] Alexander, S. N., and Elbourn, R. D. Computer Performance Tests Employed by the National Bureau of Standards. In *Proc. AIEE-IRE 1953*, ACM Press (1953), 58-61.

[2] Andersen, E., Liu, Y., Snider, R., Szeto, R., and Popovic, Z. 2011. Placing a value on aesthetics in online casual games. In *Proc. CHI 2011*. ACM Press (2011), 1275-1278.

[3] Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., and Stage, J. What happened to remote usability testing?: an empirical study of three methods. In *Proc. CHI 2007*, ACM Press (2007), 1405-1414.

[4] Bargas-Avila, J. A., & Hornbæk, K. Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Empirical Studies of User Experience. In *Proc. CHI 2011*, ACM Press (2011), 2689-2698.

[5] Bennett, J. Managing to meet usability requirements. In J. Bennett, D. Case, J. Sandelin, & M. Smith (Ed.s) *Visual Display Terminals: Usability Issues and Health Concerns*, Englewood Cliffs, NJ, Prentice-Hall (1984), 161-184.

[6] Bertelsen, O. The Activity Walkthrough: an expert review method based on activity theory. In *Proc. NordiCHI 2004*, ACM Press (2004), 251-254.

[7] Blackmon, M. H., Polson, P. G., Kitajima, M., and Lewis,C. Cognitive Walkthrough for the Web. In *Proc. CHI 2002*, ACM Press (2002), 463-470.

[8] Blythe, M., Robinson, J., and Frohlich, D. Interaction design and the critics: what to make of the "weegie". In *Proc. NordiCHI 2008*, ACM Press (2008), 53-62.

[9] Bødker, S. When Second Wave HCI meets Third Wave Challenges. In *Proc. NordiCHI 2006*, ACM Press (2006), 1-8.

[10] Brooke, J. SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Weerdmeester, A. Thomas, & I. L. McClelland (Eds.) *Usability Evaluation in Industry*, London, UK, Taylor & Francis (1996), 189-194.

[11] Butler, K. A. Connecting theory and practice: A case study of achieving usability goals. In *Proc. CHI 1985*, ACM Press (1985), 85-88.

[12] Card, S. K., Moran, T. P., and Newell, A. Computer Text-Editing: An Information-Processing Analysis of a Routine Cognitive Skill. *Cognitive Psychology 12*, 1 (1980), 32-74.

[13] Chi, E. H. A position paper on 'Living Laboratories': Rethinking ecological designs and experimentation in Human-Computer Interaction. In *Proc. HCII 2009*, Springer (2009), 597-605.

[14] Cockton, G. Make Evaluation Poverty History. In *Ext. Abstracts CHI 2007*, ACM Press (2007).

[15] Costello, B., and Edmonds, E. Directed and emergent play. In *Proc.C&C 2009*, ACM Press (1009), 107–116.

[16] Dumas, J. S., and Salzman, M. C. Usability Assessment Methods. *Reviews of Human Factors and Ergonomics 2*, 1 (2006), 109-140.

[17] Desmet, P.M.A., Overbeeke, C.J. and Tax, S. J. E. T. Designing Products with Added Emotional Value: Development and Application of an Approach for Research through Design. *The Design Journal 4*, 1 (2001), 32-47.

[18] Embley, D. W., and Nagy, G. Behavioral aspects of text editors. *ACM Computing Surveys 13,* 1 (1981), 33-70.

[19] English, W. K., Engelbart, D. C., and Berman, M. L. (1967). Display-Selection Techniques for Text Manipulation. *IEEE Trans. on Human Factors in Electronics HFE-8*, 1 (1967), 5-15.

[20] Forlizzi, J., and Battarbee, K. Understanding Experience in Interactive Systems. In *Proc. DIS 2004*, ACM Press (2004), 261-268.

[21] Gaines, B. R. From Ergonomics to the Fifth Generation: 30 Years of Human-Computer Interaction Studies. *Computer Compacts 2*, 5-6 (1985), 158-161.

[22] Goldstine, H. H. *The Computer: from Pascal to von Neumann*. Princeton University Press, Princeton, New Jersey, 1972.

[23] Good, M. An Ease of Use Evaluation of an Integrated Document Processing System. In *Proc. CHI 1982*, ACM Press (1982), 142-147).

[24] Gould, J.D., and Lewis, C. Designing for Usability: Key Principles and What Designers Think. *Comm. of the ACM 28*, 3 (1985), 300-311.

[25] Grant, E. E., and Sackman, H. An Exploratory Investigation of Programmer Performance Under On-Line and Off-Line Conditions. System Development Corp. Technical Report SP-2581, 1966.

[26] Gray, W. D., and Salzman, M. C. Damaged merchandise? A review of experiments that compare usability methods. *Human-Computer Interaction 13,* 3 (1998), 203-261.

[27] Greenberg, S., and Buxton, B. (2008). Usability evaluation considered harmful (some of the time). In *Proc. CHI 2008*, ACM Press (2008), 111-120.

[28] Grossman, T., Fitzmaurice, G., and Attar, R. A Survey of Software Learnability: Metrics, Methodologies and Guidelines. In *Proc. CHI 2009*, ACM Press (2009), 649-658.

[29] Grudin, J. Utility and usability: research issues and development contexts. *Interacting with Computers 4*, 2 (1992), 209-217.

[30] Hartson, H. R., and Hix, D. *Developing User Interfaces*. John Wiley, New York, USA, 1993.

[31] Hartmann, J., Sutcliffe, A., and De Angeli, A. Towards a theory of user judgment of aesthetics and user interface quality. *ACM TOCHI 15*, 4 (2008), 1–30.

[32] Hassenzahl, M. The Thing and I: Understanding the Relationship Between User and Product. In M. A. Blythe, Monk A. F., Overbeeke, K, & Wright, P. C. (Eds.) *Funology: From Usability to Enjoyment*, Norwell, MA, Kluwer Publishers (2003), 31-42.

[33] Hassenzahl, M., and Tractinsky, N. User experience – a research agenda. *Behaviour & Information Technology 25*, 2 (2006), 91-97.

[34] Hassenzahl, M., and Ullrich, D. To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers 19*, 4 (2007), 429–437.

[35] Hewett, T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., and Verplank, W. *ACM SIGCHI Curricula for Human-Computer Interaction.* (1996). http://sigchi.org/cdg/cdg2.html#2_1

[36] Holland, C., Komogortsev, O., and Tamir, D. Identifying usability issues via algorithmic detection of excessive visual search. In *Proc. CHI 2012*, ACM Press (2012), 2943-2952.

[37] Hollingsed, T., and Novick, D. G. Usability Inspection Methods after 15 Years of Research and Practice. In *Proc. SIGDOC 2007*, ACM Press (2007), 249-255.

[38] Hornbæk, K., and Law, E. L. Meta-analysis of correlations among usability measures. In *Proc. CHI 2007*, ACM Press (2007), 617-626.

[39] Hornbæk, K. Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology 29*, 1 (2010), 97-111.

[40] ISO 13407. Human-Centered Design Process for Interactive Systems (1998).

[41] Israel, D. R. Simulation Techniques for the Test and Evaluation of Real-Time Computer Programs. *Journal of the ACM 4*, 3 (1957), 354-361.

[42] Karapanos, E., Zimmerman, J., Forlizzi, J., and Martens, J. User Experience Over Time: An Initial Framework. In *Proc. CHI 2009*, ACM Press (2009), 729-738.

[43] Kaye, J., & Sengers, P. The Evolution of Evaluation. In *Ext. Abstracts CHI 2007*, ACM Press (2007).

[44] Kirakowski, J., and Corbett, M. Measuring User Satisfaction. In *Proc. HCI 1988*, Cambridge University Press (1988), 329-338.

[45] Lavie, T., and Tractinsky, N. Assessing dimensions of perceived visual aesthetics of web sites. *Int. Journal of Human-Computer Studies 60*, 3 (2004), 269-298.

[46] Law, E. L., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. Understanding, Scoping and Defining User eXperience: A Survey Approach. In *Proc. CHI 2009*, ACM Press (2009), 719-728.

[47] Lewis, C., and Mack, R. Learning to Use a Text Processing System: Evidence from "Thinking Aloud" Protocols. In *Proc. CHI 1982*, ACM Press (1982), 387-392.

[48] Lewis, C., Polson, P., Wharton, C., and Rieman, J. Testing a Walkthrough Methodology for Theory-Based Design of Walk-

Up-And-Use Interfaces. In *Proc. CHI 1990*, ACM Press (1990), 235-242.

[49] Lieberman, H. *The Tyranny of Evaluation*, paper presented at CHI 2003 Fringe. http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html

[50] Lindgaard, G., and Parush, A. Utility and Experience in the Evolution of Usability. In E. Law, E. Hvannberg, and G. Cockton (Eds.), *Maturing Usability: Quality in Software, Interaction and Value*, Berlin, Springer-Verlag (2008), pp. 222-240.

[51] Madathil, K. C., and Greenstein, J. S. Synchronous remote usability testing: a new approach facilitated by virtual worlds. In *Proc. CHI 2011*, ACM Press (2011), 2225-2234.

[52] Mandryk, R., Inkpen, K., and Calvert, T. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology 25*, 2 (2006), 141–158.

[53] McPherson, J. L., and Alexander, S. N. Performance of the census UNIVAC system. In *Proc. AIEE-IRE 1951*, ACM Press (1951), 16-22.

[54] Molich, R., and Dumas, J. S. Comparative usability evaluation (CUE-4). *Behaviour & Information Technology 27*, 3 (2008), 263-281.

[55] Navalpakkam, V., and Churchill, E. Mouse tracking: measuring and predicting users' experience of web-based content. In *Proc. CHI 2012*, ACM Press (2012), 2963-2972.

[56] Nielsen, J., and Molich, R. Heuristic Evaluation of User Interfaces. In *Proc. CHI 1990*, ACM Press (1990), 249-256.

[57] Nørgaard, M. and Hornbæk, K. What Do Usability Evaluators Do in Practice? An Explorative Study of Think-Aloud Testing. In *Proc DIS 2006*, ACM Press (2006), 209-218.

[58] Norman, D. Emotional Design: Why We Love (or Hate) Everyday Things. Basic, New York, 2004.

[59] O'Brien, H., and Toms, E. The development and evaluation of a survey to measure user engagement. *JASIST 61*, 1 (2010), 50–69.

[60] Palen, P., and Salzman, M. Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proc. CSCW 2002*, ACM Press (2002), 87-95.

[61] Pollard, B. W. The Design, Construction, and Performance of a Large-Scale General Purpose Digital Computer. In *Proc. AIEE-IRE 1951*, ACM Press (1951), 62-70.

[62] Ravden, S., and Johnson, G. Evaluating Usability of Human-Computer Interfaces: a Practical Method. New York, NY, Halsted Press, 1989.

[63] Roberts, T. L., and Moran, T. P. Evaluation of Text Editors. In *Proc. CHI 1982*, ACM Press (1982), 136-141).

[64] Rogers, Y., Sharp, H., and Preece, J. *Interaction Design: Beyond Human-Computer Interaction* (3rd ed.). Wiley Publishing, Chichester, UK, 2011.

[65] Root, R. W., and Draper, S. Questionnaires as a software evaluation tool. In *Proc. CHI 1983*, ACM Press (1983), 83-87.

[66] Rosenbaum, S., Rohn, J., and Humburg, J. A toolkit for strategic usability: Results from workshops, panels, and surveys. In *Proc. CHI 2000*, ACM Press (2000), 337–344.

[67] Sauro, J., and Lewis, J. R. Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability. In *Proc. CHI 2009*, ACM Press (2009), 1609-1618.

[68] Sanders, E. N. Converging perspectives: Product development research for the 1990s. *Design Management Journal 3*, 4 (1992), 49-54.

[69] Sime, M. E., Green, T. R. G., and Guest, D. J. Psychological Evaluation of Two Conditional Constructions Used in Computer Languages. *Int. Journal of Man-Machine Studies 5*, 1 (1973), 105-113.

[70] Staiano, J., Menéndez, M., Battocchi, A., De Angeli, A., and Sebe, N. UX_Mate: From facial expressions to UX evaluation. In *Proc. DIS 2012*, ACM Press (2012), 741-750.

[71] Tractinsky, N., Katz, A. S., and Ikar, D. What is beautiful is usable. *Interacting with Computers 13*, 2 (2000), 127-145.

[72] Vermeeren, A., Law, E., Roto, V., Obrist, M., Hoonhout, J., and Väänänen-Vainio-Mattila, K. User experience evaluation methods: current state and development needs. In *Proc. NordiCHI 2010*, ACM Press (2010), 521-530.

[73] Zhai, S. *Evaluation is the worst form of HCI research except all those other forms that have been tried*. http://www.shuminzhai.com/papers/EvaluationDemocracy.htm