Journal of Usability Studies
Vol. 11, Issue 3, May 2016 pp. 83–88

Usability Testing: Too Early? Too Much Talking? Too Many Problems?

Morten Hertzum
University of Copenhagen
Denmark
hertzum@hum.ku.dk

Abstract

Usability testing has evolved in response to a search for tests that are cheap, early, easy, and fast. In addition, it accords with a situational definition of usability, such as the one propounded by ISO. By approaching usability from an organizational perspective, this author argues that usability should (also) be evaluated late when the system is ready for field use, that usability professionals should be wary of using the thinking-aloud protocol, and that they should focus more on the achievement of effects than on problem detection.

Introduction

Usability is often defined by quoting the ISO definition that states usability is the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241, 2010, p. 3). However, a study of how usability professionals construe usability found that the ISO definition captured only 53% of the constructs usability professionals used in talking about the usability of information systems (Hertzum & Clemmensen, 2012). This discrepancy between the ISO definition of usability and the thinking of the professionals concerned with delivering usability suggests the existence of additional conceptions of usability. Hertzum (2010) described six images of usability—universal, situational, perceived, hedonic, organizational, and cultural—and argued that in spite of a shared essence they differ in focus, scope, mindset, and perspective. Each image provides a partial view of usability in that it emphasizes some issues and renders other issues invisible.

The partiality of the individual images of usability becomes practically important when the focus is shifted from the definitions of the usability concept to the methods used in evaluating usability. The ISO definition is a prominent expression of situational usability, which lends itself to usability testing. In usability testing (e.g., Dumas & Loring, 2008; Rubin & Chisnell, 2008) it is common for users to be asked to think aloud while using a system to solve predefined test tasks while an evaluator observes the users' behavior and listens in on their "thoughts." Much of the thinking about how to conduct usability tests has been driven by a concern that they must be early in the design process, fast to perform, cheap in resource requirements, and easy to apply for the usability professional. This discount thinking has previously been challenged by, for example, Cockton and Woolrych (2002). In this paper, I critically discuss three aspects of usability tests by looking at them from the perspective of organizational usability.

Test Late

Early testing is a longstanding principle in user-centered design (Gould & Lewis, 1985), and it is frequently argued that unless usability problems are discovered early they will be too costly to fix and remain unaddressed. For example, Hertzum (2006) found that the percentage of fixed problems dropped from 73% of the 38 problems discovered in the first test to 0% of the 11 problems discovered in the last test. The first test was a usability test conducted in the lab; the last test was a pilot implementation conducted in the field. Usability testing is, partly, done in the lab to enable testing of system prototypes that are not yet sufficiently functional to be tested in the field, that is, to test earlier than it would otherwise be possible. However, substituting organizational for situational usability changes the priorities.

Elliott and Kling (1997) defined organizational usability as "the match between a computer system and the structure and practices of an organization, such that the system can be effectively integrated into the work practices of the organization's members" (p. 1024). By explicitly setting usability in a context of organizations with a structure and collaborative work practices, Elliott and Kling bring issues to the forefront that are crucial to the usability of many systems but deemphasized in the ISO definition and foreign to usability testing in the lab. To address organizational usability, an evaluation must await that the system is sufficiently functional and robust to be tested in the field. That is, organizational usability lends itself to pilot implementation (Hertzum, Bansler, Havn, & Simonsen, 2012) in which a system is introduced in its intended environment and used for real work by prospective users for a restricted period of time. In addition, Wagner and Piccoli (2007) contended that a system does not really become salient to users until it starts to affect their work and require them to change their work practices. It is at this point most users start reacting to the system and become motivated to influence its design.

In summary:

- The effects of a system on the users' work practices are not exercised in a usability test because it is conducted away from their real work.
- Problems relating to mismatches between the system and its organizational context can be uncovered in pilot implementations.
- To be truly user centered, usability work must involve users when they are motivated to take part.

A consequence of these summary points is that systems must (also) be tested late and that there is a need for methods aimed at evaluating the usability of a system in the transition from design process to system use. Pilot implementation is one such method.

Be Wary of Thinking Aloud

Usability testing normally involves that the users are asked to think aloud while they use the system. Thinking aloud is likely to be prohibitively unnatural and taxing if tests are performed in the field, say in a public or collaborative context, or for prolonged periods of time. That is, thinking aloud appears confined to lab settings. In addition, it is often not clear what the users are concretely instructed to do when they are asked to think aloud. Many usability professionals appear to relax the prescriptions of the classic thinking-aloud protocol by asking users to verbalize their feelings, expectations, reflections, and proposals for redesigns (Boren & Ramey,

2000; Nørgaard & Hornbæk, 2006). It is well-established that such relaxed thinking aloud affects behavior (Ericsson & Simon, 1993; Hertzum, Hansen, & Andersen, 2009). For example, Hertzum et al. (2009) found that during relaxed thinking-aloud users took longer to solve tasks, spent a larger part of tasks on general distributed visual behavior, navigated more from one page to another on the websites used in the experiment, scrolled more within pages, and experienced a higher mental workload. It is, however, debated whether the additional information that usability professionals get from relaxed thinking aloud outweighs its effects on behavior. For example, Goodman, Kuniavsky, and Moed (2012) considered the additional information valuable whereas Boren and Ramey (2000) recommended restricting relaxed thinking aloud to curb its effects on behavior.

It has also been debated whether complying with the classic thinking-aloud protocol produces verbalizations without affecting behavior. A meta-analysis (Fox, Ericsson, & Best, 2011) found that behavior was unaffected, except for prolonged task completion times. However, some studies indicate that classic thinking aloud is reactive in the presence of interruptions (Hertzum & Holmegaard, 2013), impairs users' performance of spatial tasks (Gilhooly, Fioratou, & Henretty, 2010), and influences perceived time (Hertzum & Holmegaard, 2015). That is, it may not be enough to avoid relaxed thinking aloud; classic thinking aloud may also affect users' behavior. Because part of the mental activity that is of interest in usability evaluations is manifest through the users' observable behavior as well as through their verbalizations, some studies found that classic thinking aloud adds little value to usability tests (Lesaigle & Biers, 2000; van den Haak, de Jong, & Schellens, 2004). These findings may suggest abandoning thinking aloud in favor of having users perform in silence. Other studies found that thinking aloud, especially relaxed thinking aloud, adds considerable value. For example, Hertzum, Borlund, and Kristoffersen (2015) found that 38–44% of the verbalizations made during relaxed thinking aloud were of medium or high relevance to the identification of usability problems.

In summary:

- Classic thinking aloud may not affect behavior but may also add little value to usability tests beyond what can be derived from users' observable behavior.
- Relaxed thinking aloud affects behavior but appears to add value to usability tests beyond what can be derived from users' observable behavior.
- The only way to avoid that thinking aloud may affect behavior is to abandon thinking aloud concurrently with the behavior.

Retrospective thinking aloud or a retrospective interview, possibly supported by a video recording of the session, may provide a cost-effective separation between performing with the system and commenting on it. This separation also provides for using the system in the field but moving to the lab for the retrospective part.

Focus on Effects, Not Problems

Usability tests are conducted to gauge users' experience with a system and, thereby, find any problems that prevent users from completing their tasks, slow them down, or otherwise degrade their user experience. The focus of usability tests on identifying problems has caused its own problems because evaluators who analyze the same usability test sessions have been found to identify substantially different sets of usability problems (Hertzum & Jacobsen, 2003; Hertzum, Molich, & Jacobsen, 2014). Evidence of this evaluator effect exists for multiple usability evaluation methods, for novice and experienced evaluators, for simple and complex systems, for minor and severe problems, and for problem detection and ratings of problem severity. Limited agreement about what the problems are complicates directed efforts to improve systems by addressing their problems. In addition, avoiding usability problems is not the same as achieving good usability—just as the absence of dissatisfaction does not equal the presence of satisfaction (Tuch & Hornbæk, 2015).

An information system is a means to an end. Good usability implies that the system supports its users in obtaining a desired effect. For many systems the effects concern the performance of the organization in which the system will be used. For example, the effect sought from the medication module of an electronic patient record in a hospital may be that the right patient will get the right medication at the right time. Such effects can be assessed in pilot implementations

and provide a meaningful test of the organizational usability of the system. Somewhat surprisingly this kind of test is rarely conducted, with the result that it often remains unknown whether adjustments to a system and the associated work practices are needed to achieve the effect that motivated the investment in the system. Usability professionals can provide a good service by shifting part of their focus from the detection of usability problems to the evaluation of whether the intended effects of introducing the system are achieved. This proposal resembles the usability specifications described by Whiteside, Bennett, and Holtzblatt (1988) but is directed at the effects decisive to organizational usability.

While some of the effects sought from a system can be specified up front and function as an instrument for managing the design process, others will emerge during pilot implementations and must be added in an opportunity-based fashion (Hertzum & Simonsen, 2011; Simonsen & Hertzum, 2008). The emergent effects contribute importantly to the user experience, positively or negatively, and because they grow out of practice they cannot be experienced or evaluated unless the system is tested in the field during real work. In addition, the test must be long enough for emergent effects to surface.

In summary:

- Usability tests aim to detect problems in the design of systems, but the evaluator effect shows that the reliability of problem detection is limited.
- The effects that motivate the investment in systems are often not systematically pursued through evaluation and redesign.
- The meeting between a system and its environment will also result in emergent effects that may or may not be desired.

The achievement of planned effects and the opportunity-based response to emergent effects are central to organizational usability. Working with the achievement of effects involves other activities and, possibly, another mindset than working with problem detection.

Conclusion

The meeting between a system and its organizational context is a sociotechnical process that evolves over time and cannot be fully anticipated ahead of time. To facilitate this process, usability professionals need to feed experience from real use of the system back into the design activities. This suggests moving usability tests to later stages of the design process, replacing the lab with the field, and shifting the focus from problems toward effects. Doing so will change usability tests, for example, by boosting retrospective thinking aloud and by making pilot implementations the setting for many usability tests.

References

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261–278.

Cockton, G., & Woolrych, A. (2002). Sale must end: Should discount methods be cleared off HCI's shelves? *ACM Interactions, 9*(5), 13–18.

Dumas, J. S., & Loring, B. (2008). *Moderating usability tests: Principles & practices for interacting*. Burlington, MA: Morgan Kaufmann.

Elliott, M., & Kling, R. (1997). Organizational usability of digital libraries: Case study of legal research in civil and criminal courts. *Journal of the American Society for Information Science, 48*(11), 1023–1035.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*(2), 316–344.

Gilhooly, K. J., Fioratou, E., & Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology, 101*(1), 81–93.

Goodman, E., Kuniavsky, M., & Moed, A. (2012). *Observing the user experience: A practitioner's guide to user research* (2nd ed.). Amsterdam: Morgan Kaufmann.

Gould, J. D., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM, 28*(3), 300–311.

Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction, 21*(2), 125–146.

Hertzum, M. (2010). Images of usability. *International Journal of Human-Computer Interaction, 26*(6), 567–600.

Hertzum, M., Bansler, J. P., Havn, E., & Simonsen, J. (2012). Pilot implementation: Learning from field tests in IS development. *Communications of the Association for Information Systems, 30*(1), 313–328.

Hertzum, M., Borlund, P., & Kristoffersen, K. B. (2015). What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction, 31*(9), 557–570.

Hertzum, M., & Clemmensen, T. (2012). How do usability professionals construe usability? *International Journal of Human-Computer Studies, 70*(1), 26–42.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology, 28*(2), 165–181.

Hertzum, M., & Holmegaard, K. D. (2013). Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction, 29*(5), 351–364.

Hertzum, M., & Holmegaard, K. D. (2015). Thinking aloud influences perceived time. *Human Factors, 57*(1), 101–109.

Hertzum, M., & Jacobsen, N. E. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 15*(1), 183–204.

Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology, 33*(2), 143–161.

Hertzum, M., & Simonsen, J. (2011). Effects-driven IT development: Specifying, realizing, and assessing usage effects. *Scandinavian Journal of Information Systems, 23*(1), 3–28.

International Organizational for Standardization (ISO; 2010). ISO 9241: *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems*. Genève, Switzerland: International Organization for Standardization.

Lesaigle, E. M., & Biers, D. W. (2000). Effect of type of information on real time usability evaluation: Implications for remote usability testing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44*(37), 585–588. doi:10.1177/154193120004403710.

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the Sixth DIS Conference on Designing Interactive Systems* (pp. 209–218). New York, NY: ACM Press.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed.). Indianapolis, IN: Wiley.

Simonsen, J., & Hertzum, M. (2008). Participative design and the challenges of large-scale systems: Extending the iterative PD approach. In J. Simonsen, T. Robertson, & D. Hakken (Eds.), *PDC2008: Proceedings of the Tenth Anniversary Conference on Participatory Design* (pp. 1–10). New York, NY: ACM Press.

Tuch, A. N., & Hornbæk, K. (2015). Does Herzberg's notion of hygienes and motivators apply to user experience? *ACM Transactions on Computer-Human Interaction, 22*(4), Article 16.

van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers, 16*(6), 1153–1170.

Wagner, E. L., & Piccoli, G. (2007). Moving beyond user participation to achieve successful IS design. *Communications of the ACM, 50*(12), 51–55.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 791–817). Amsterdam: Elsevier.

About the Author

Morten Hertzum
Dr. Hertzum is a professor of Information Science at University of Copenhagen. His research interests include human-computer interaction, usability, computer supported cooperative work, information seeking, and medical informatics. He is co-editor of the book *Situated Design Methods* (MIT Press, 2014) and has published a series of papers about usability evaluation methods.