This forum addresses conceptual, methodological, and professional issues that arise in the UX field's continuing effort to contribute robust information about users to product planning and design. **— David Siegel and Susan Dray, Editors**

A Usability Test Is Not an Interview

Morten Hertzum, University of Copenhagen

Usability tests are conducted to gauge users' experience with a system, preferably before it is released for real use, and thereby find any problems that prevent users from completing their tasks, slow them down, or otherwise degrade their user experience. Such tests are important to successful systems development, yet test procedures vary and the quality of test results is sometimes contested. While there is no single accepted procedure for usability specialists to follow when conducting usability tests, these tests normally involve users who think out loud while using a system and an evaluator who observes the users' behavior and listens in on their thoughts. This common core of usability tests is illustrated in Figure 1. The possible variations include, for example, whether the users work individually or in pairs, whether the evaluator is in the room with the user or in an adjoining room, whether use of the system consists of solving preset tasks or exploring the system more freely, and whether the interaction between the evaluator and user is kept to an absolute minimum or involves frequent prompts for reflections and experiences.

Here, I focus on the interaction between the evaluator and the user. Studies in cognitive psychology, particularly K. Anders Ericsson and Herbert Simon's seminal work on verbal reports [1], prescribe that the interaction between evaluator and user should be restricted to a simple reminder to think aloud if the user falls silent: "Keep talking." While claims about the validity of user verbalizations are frequently adopted from these studies, usability practitioners tend not to follow the prescriptions of what could be termed classic thinking aloud. Instead, many usability professionals relax the protocol for thinking aloud to get richer verbalizations. The primary characteristic of such relaxed thinking aloud is that the evaluator prompts the user more frequently and in more detail. Examples of such prompts from handbooks about usability testing include: "You're frowning. Tell me what is happening," "What were you looking for in the index?" and "If this was your first time here, what would you do next?" By prompting the user in this way, relaxed thinking aloud yields verbalizations about the user's feelings, expectations, and reflections, whereas classic thinking aloud mainly yields verbalizations that describe what users are doing and how they do it.

Insights

- → Usability tests may have a conversational element but their defining characteristic is concrete system use.
- → Evaluators in usability tests observe users' behavior and probe users for verbalizations about that which cannot be observed.
- → Every usability test must strike a balance between undistorted use and rich verbalizations.

A CONVERSATIONAL ELEMENT BUT . . .

There is a conversational element in relaxed thinking aloud (a.k.a., interactive thinking aloud). A recent study finds that the users spoke an average of 110 words per minute during a test session and the evaluator who moderated the sessions spoke an average of 26 words per minute [2]. It may, therefore, be tempting to construe a usability test with relaxed thinking aloud as a kind of interview. Indeed, Goodman et al. do exactly that [3]. In my opinion, it is a mistake.

In a usability test, the users interact with the system that is being tested. Their behavior and verbalizations relate to their concrete use of the system for solving the test tasks. Conversely, interviewees interact with the interviewer and their verbalizations relate to the interviewer's questions. Interviewees may talk about a system and reflect on their experiences with it, but in the interview situation, the talking is detached from concrete use of the system. In the same way, the usability-test evaluator observes users' task performance directly and need only probe users for verbalizations about matters that cannot be observed. The resulting data consists of observations as well as user verbalizations and provides the additional opportunity for comparing and contrasting these two sources of data. Conversely, an interviewer does not have direct access to users' task performance and must obtain data about it in the same way as data about the users' experience. That is, by asking users to

describe and explain verbally.

Table 1 summarizes how usability tests are defined by concrete system use and contrasts them with interviews.

Obviously, there are situations in which the appropriate method to use is the interview, not a usability test. Examples include but are not restricted to capturing requirements prior to the existence of a testable prototype and evaluations aimed at aggregating the users' experience of a long-used system. In these situations, usability specialists should conduct interviews rather than transmute usability tests into something they are not.

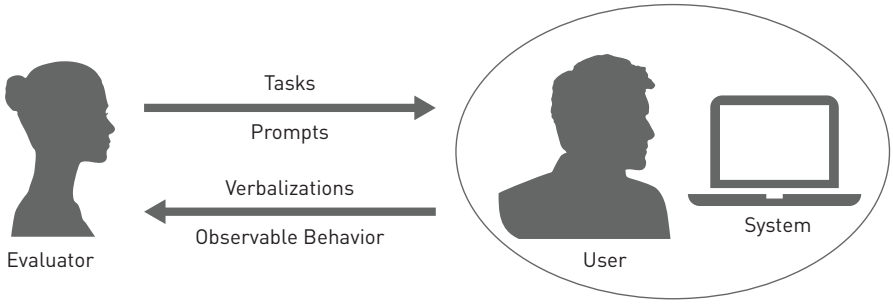

Figure 1. A usability test.

THE TRADE-OFF

Just as usability tests can be conducted in numerous ways, there are also many ways to conduct interviews. Interviews can, for example, be set in a location away from other activities to avoid disruptions and provide for thoughtfulness, or they can be conducted in situ to be close to the objects and activities the interviewee is talking about. The closeness allows the interviewee to show some of the things that would otherwise have to be told, and it may trigger verbalizations that more closely explain how objects are used and activities performed. While in-situ interviews are conducted in the setting the interviewee is talking about, the interviewee is not performing the activities but rather giving a guided tour of the setting. Thus, even when interviews are conducted in situ, use features differently in interviews than in usability tests. Figure 2 illustrates how usability tests may differ in their inclusion of conversational elements but share concrete system use as their defining characteristic. Interviews may differ in their closeness to use but share conversation as their defining characteristic.

It is important to maintain the distinction between usability tests and interviews because it creates clarity about what each method aims to accomplish. Usability tests aim to represent the use of the tested system sufficiently undistorted for users' behavior and experience during the test to mimic their behavior and experience when using the system outside the test. How different variants of the usability test succeed in achieving this aim is an important concern. For example, Hertzum et al. express concern about their findings that relaxed thinking aloud affects user behavior in multiple ways [4]. During relaxed thinking aloud, users took longer to solve tasks, navigated more from one page to another on the tested website, scrolled more on the individual pages, spent a larger part of tasks on general distributed visual behavior, and experienced higher mental workload. These findings show that the richer verbalizations of relaxed thinking aloud come at an unwelcome cost and, thereby, highlight a trade-off between undistorted use and rich verbalizations. This trade-off is central to usability testing but does not feature in interviews.

	Usability test	Interview
User (interviewee)	Exercises the system by attempting to use it and comments on the process	Talks about the system by reflecting on user experiences with it
	Interacts with the system, which is the main focus of the user's attention	Interacts with the interviewer, who is the target of the user's verbalizations
	Is confronted with his or her task performance	May be in touch with or remain detached from his or her task performance
Evaluator (interviewer)	Observes users' behavior and task performance directly	Can access behavior and task performance only through the users' verbalizations
	Probes users for verbalizations about that which cannot be observed	Relies on verbalization for descriptions of users' experience as well as their behavior
	Can compare and contrast users' verbalizations with their task behavior	Must take users' verbalizations at face value or probe for examples and elaborations
Setting	The tasks stipulate, strongly or loosely, what the users should do	The interview guide stipulates, strongly or loosely, what the users should talk about
	The system impacts what users can do, thereby triggering behavior and verbalizations	The probing impacts what users say, but they may stick to or drift away from the probes
	The setting supports a sustained focus on concrete matters	The focus may alternate between concrete and abstract matters

→ **Table 1.** Contrasting usability tests and interviews.

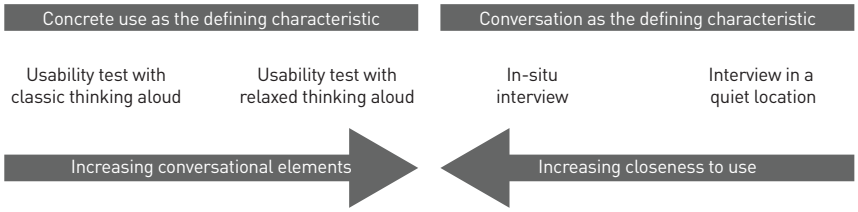

Figure 2. The different roles of concrete use and conversation in defining usability tests and interviews.

STRIKING A BALANCE

The proper balance between undistorted use and rich verbalizations varies with the situation. Thus, a decision about how to strike the balance should be part of the planning of every usability test. For example, a use situation consisting of many brief and fast-paced interactions—such as a ticket vending machine in a public place—suggests a test that gives priority to undistorted use. Conversely, a use situation aimed at capturing the users' interest and getting them to browse for a purchase—such as an e-commerce site for fashion products—suggests a test that focuses more on eliciting rich verbalizations.

Undistorted use can largely be achieved by restricting users' verbalizations to classic thinking aloud. Ericsson and Simon assert that classic thinking aloud distorts use in no other way than by prolonging task completion [1]. That is, users follow the same strategies, perform the same activities, and reason about tasks in the same way as when they are not thinking aloud. They just take longer to do it because thinking out loud is a slower process than thinking without the verbalization. Some studies indicate, however, that classic thinking aloud influences use in subtle ways. For example, Hertzum and Holmegaard find that classic thinking aloud affects users' time perception [5]. Possible explanations for this finding include that classic thinking aloud requires attention or increases mental workload. Furthermore, explicit instruction is necessary to train users in classic thinking aloud. Otherwise, they will normally construe thinking aloud as a conversation with the evaluator who moderates the test.

Rich verbalizations can be achieved by employing relaxed thinking aloud. Studies find that relaxed thinking aloud includes verbalizations of user experience, redesign proposals, and explanations of behavior [2]. These verbalizations were found to be more relevant to the identification of usability problems than verbalizations describing the users' actions—a frequent type of verbalization during both relaxed and classic thinking aloud. Ted Boren and Judith Ramey are, however, concerned that some of the prompts used in relaxed thinking aloud, including the examples given earlier in this article, go too far [6]. They propose a middle ground between classic and relaxed thinking aloud to acknowledge the value of a conversational element in usability tests but, at the same time, constrain it to maintain the focus on concrete use. Their proposal includes prompts such as "Mm hmm" and "Uh huh," which encourage the user to elaborate without suggesting that the evaluator will assume speakership and without directing the user's attention to specific actions and interface objects.

Finally, it may be noted that crowdsourced usability tests, which are gaining popularity, add reality to the standard recommendation in classic thinking aloud of instructing the user to "act as if you are alone in the room speaking to yourself" [1]. In crowdsourced usability tests, the users run the test session themselves, often in their homes, while their behavior and verbalizations are video-recorded for subsequent analysis by usability specialists. This setup prevents construing the test as an interview. To the extent that the absence of an evaluator moderating the test session and prompting the user promotes classic thinking aloud, these tests may produce more undistorted records of use. At the same time, studies show that users in crowdsourced usability tests make frequent verbalizations, though fewer than during relaxed thinking aloud with the user and evaluator in the same place [2]. Future research should explore crowdsourced usability tests in more detail to determine how they balance undistorted use and rich verbalizations.

Deciding how to strike the balance between undistorted use and rich verbalizations should be part of the planning of every usability test.

ENDNOTES

1. Ericsson, K.A. and Simon, H.A. *Protocol Analysis: Verbal Reports as Data*. Revised Edition. MIT Press, Cambridge, MA, 1993.
2. Hertzum, M., Borlund, P., and Kristoffersen, K.B. What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction 31*, 9 (2015), 557–570.
3. Goodman, E., Kuniavsky, M., and Moed, A. *Observing the User Experience: A Practitioner's Guide to User Research. Second Edition*. Morgan Kaufmann, Amsterdam, 2012.
4. Hertzum, M., Hansen, K.D., and Andersen, H.H.K. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology 28*, 2 (2009), 165–181.
5. Hertzum, M. and Holmegaard, K.D. Thinking aloud influences perceived time. *Human Factors 57*, 1 (2015), 101–109.
6. Boren, T. and Ramey, J. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication 43*, 3 (2000), 261–278.

Morten Hertzum is a professor of information science at the Royal School of Library and Information Science at the University of Copenhagen. His research interests are in human-computer interaction, computer-supported cooperative work, information seeking, and healthcare informatics.
→ hertzum@hum.ku.dk

DOI: 10.1145/2875462