Does Think Aloud Work? How Do We Know?

Judith Ramey (panel moderator)
Laboratory for Usability Testing and Evaluation (LUTE)
Dept. of Technical Communication
Univ. of Washington, Box 352195
Seattle, WA 98195 USA
jramey@u.washington.edu

Ted Boren
Audiovisual Department
The Church of Jesus Christ of Latter-day Saints
Salt Lake City, UT 84150, USA
borenmt@ldschurch.org

Elisabeth Cuddihy
Dept. of Technical Communication
Univ. of Washington, Box 352195
Seattle, WA 98195 USA
ecuddihy@u.washington.edu

Joe Dumas
Senior Human Factors Specialist
Design and Usability Center
Bentley College
jdumas@bentley.edu

Zhiwei Guan
Dept. of Technical Communication
Univ. of Washington, Box 352195
Seattle, WA 98195 USA
zguan@u.washington.edu

Maaike J. van den Haak
University of Twente
Faculty of Behavioral Sciences
Depart. of Communication Studies
P.O. Box 217, 7500 AE Enschede
The Netherlands
m. j.vandenhaak@utwente.nl

Menno D.T. De Jong
University of Twente
Faculty of Behavioral Sciences
Depart. of Communication Studies
P.O. Box 217, 7500 AE Enschede
The Netherlands
M.D.T.deJong@utwente.nl

CHI 2006, April 22–27, 2006, Montréal, Québec, Canada.
ACM 1-59593-298-4/06/0004.

Abstract

The think aloud method is widely used in usability research to collect user's reports of the experience of interacting with a design so that usability evaluators can find the underlying usability problems. However, concerns remain about the validity and usefulness of think aloud in usability studies. In this panel we will present current studies of the think aloud method, examine and question its usage in the field, discuss the possible pitfalls that may threaten the validity of the method, and provide comments/suggestions on the application of the method. Panel participants will discuss results drawn from both applied research and basic research.

We believe that this panel discussion will be useful for HCI designers and usability practitioners in that it will acquaint them with concerns that people have about the think aloud method and provide them with suggestions for improved use of the method. For HCI or usability researchers, this panel discussion will address the importance of formally investigating currently used or newly designed usability methods.

Keywords

Think aloud method, methodology, validity, reliability, usability study, applied research, scientific research, practical guidelines.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Evaluation/Methodology

Overview of Panel Topics

Think aloud (TA) is widely used in usability studies to gain insight into how people work with a product or interface. The basic principle of TA is to ask users to work on typical tasks and to verbalize their task performance and thought process. In the most commonly used approach, concurrent think aloud (CTA), users' verbalization takes place simultaneously with their task performance. The verbalization can also take place after users have completed the tasks, in which case this method is variously called retrospective think aloud (RTA), post task testing, retrospective protocol, retrospective report, think after, etc. By collecting user's verbal reports on their task performance, usability practitioners hope to gain insight into how users interact with the product and identify the barriers that hinder users' interaction.

Although the TA method has been widely used to study various materials from webpages to end-user products, and in various settings from the laboratory to the users' home or workplace, there has been little *formal* investigation in the HCI field of the validity of this method and several questions about its use in usability studies remain unsolved.

First, when the concurrent think aloud method is used in a study, people have concerns about its reactivity—the possibility that the act of speaking concurrently may influence users' task performance. The effort that users make to verbalize information while performing tasks might distract their attention and concentration, and the effort to fully verbalize the steps in the work might change the ways that users attend to the task components.

Although retrospective think aloud (RTA) was proposed to avoid the problems of concurrent think aloud, there has been little work done to scientifically investigate the validity and reliability of RTA. Most of the research to date on RTA is applied research that has focused on comparing RTA to other methods (e.g., CTA) in specific task domains. More scientific studies of the validity and reliability of think aloud methods are required.

Second, the actual use of the think aloud method in usability studies is significantly different from the procedures used by Ericsson and Simon for doing cognitive science study from which the method is borrowed; use also varies widely among practitioners. Boren and Ramey indicate that these variances make it difficult to "compare or replicate studies, vouch for the validity of the results, or teach a standard of practice to newcomers to the field" .

Third, few empirical studies have formally investigated whether the think aloud method works as people expect in usability studies. Most of the past research on think aloud methods was based on user testing rather than experimental study, which arguably undermines the validity and generalizability of the conclusions drawn.

All the issues discussed above raise concerns about the validity of the think aloud method. To address these concerns, the panelists will discuss specific issues: how people use thinking-aloud methods, what they expect to get from the results, what their concerns are about the methods, and how these concerns are addressed or partially addressed by research and experience.

Proposed Panel Format

The format of the panel will be hybrid and engaging: (1) a brief introduction of the panel by the moderator, (2) short position statements by each of the panelists, (3) quick audience-participation "pop-up surveys" on major issues about the think aloud method, such as whether usability practitioners should strictly follow Ericsson and Simon's instructions about how to collect verbal data, (4) video demonstration of instructions for a usability test that includes both modeling and training of the think aloud methods following Ericsson and Simon's protocol and Boren and Ramey's protocol, after which the panelists will answer questions on the two approaches, and (5) panelist discussion of issues submitted by the audience..

Panelist's Contribution

Judy Ramey, Zhiwei Guan, and Elisabeth Cuddihy's Contribution

While many practitioners have employed think aloud methods in their usability studies, there are very few standards for procedures, data analysis, or reporting. And, there are few studies indicating whether the think aloud method is valid, how effective it is for finding usability problems, and the degree to which it yields useful data suggesting reasonable design revisions.

Every discipline needs to demonstrate the appropriateness of its methods. In our field, we believe that it is the responsibility of academic researchers to scientifically study common usability methods in order to determine their validity, to suggest best practices for administering the methods, to determine how to best analyze resultant data, and to determine how usability testing data can best inform redesign.

The Laboratory for Usability Testing and Evaluation (LUTE) in the Department of Technical Communication at the University of Washington, has begun such work by hosting one study that examined the validity of concurrent think aloud and running a second study that examined stimulated retrospective think aloud. We will report on the scientific questions that have been raised during these two studies, the difficulties involved in answering these questions, and how scientific study of such questions can improve usability practice.

Joe Dumas's Contribution

Think aloud is the most important method we have in the toolkit of usability evaluation. It uncovers more problems than any other measure. Without it usability testing would never have gained the popularity it has and without it many developers and managers would never been converted to user-centered design. But the think aloud protocol is one part reality and one part illusion. Test moderators are not consistent in the way they train participants to think aloud and in the way they intervene during a session. The typical participants' protocol is a mixture of reports of actions and statements that may have as much to do with the role that participants think they should be playing as with any interior dialog. As practitioners we selectively pick what we want from this protocol, and we choose to let developers who watch it believe that the protocol presents an unedited record of interior dialog. We have also chosen to ignore the validity and reliability of think aloud because it has been so useful for convincing developers that problems exist. It's time we examined our motives and took a hard look at this method.

Menno De Jong and Maaike Van den Haak's Contribution

Literature on think-aloud protocols often has an ambivalent status among usability professionals. On the one hand, they tend to defend their usability test practices by referring to the theoretical and empirical literature which justifies the use of verbal protocols as a valid method for uncovering mental processes. On the other hand, they increasingly object to the straightforward adoption of the guidelines that are proposed in the same literature. In view of this ambivalence, it would seem justified to reexamine the available literature in search of the validity of think-aloud protocols as a professional research method to uncover user problems in interfaces and websites.

An additional motivation for such a reexamination lies in the fact that our research comparing concurrent think-aloud protocols to retrospective think-aloud protocols and constructive interaction so far has revealed inconsistent results, depending on the test object that we used. This makes it interesting to focus on the influence of different task types on the reactivity and the exhaustiveness of think-aloud protocols.

We will report on the results of an ongoing study, in which we differentiated between those sub-processes that are part of using an interface (e.g. an online catalogue) and the sub-processes that are part of using a website (e.g. a municipal website). For each of these sub-processes, we will discuss relevant findings from the literature available.

Ted Boren's Contribution

One often encounters two diametrically opposed approaches to conducting a think-aloud protocol. One approach is that usability research methods (including "think aloud") are "quick and dirty" anyway, so it really makes no difference if the techniques we employ are methodologically sound. We just need to get our answers, regardless of whether the data is "100% pure." The contrasting approach is to apply "experimental rigor" in a usability testing context. Some usability professionals describe usability tests and results in terms of "experimental design," "subjects," "significance," "samples," "variability," "outliers," etc.—despite the fact that most formative usability tests do not have truly experimental goals and produce data that is not generally amenable to statistical analysis. In terms of verbal protocols, those with this experimental mindset balk at any intervention other than perhaps reminding participants to "keep talking," for fear that intervention may de-rail "the study." (In practice, however, more people seem to "talk this way" than "walk this way.")

In the middle are some people taking what I think is the right approach. They realize that how we conduct usability sessions can indeed affect the kind, amount, and reliability of data we collect. At the same time, they acknowledge that in certain situations, practitioners must intervene to collect crucial information, even if this has some impact on task performance. This is the area I would like to explore—how to make those crucial interventions, how to identify when you are at a point that requires one, and when it is best to just keep quiet. Until there is some agreement on these points, it is difficult to consider think Aloud as a standard usability "method."