

Exploring Two Methods of Usability Testing: Concurrent versus Retrospective Think-Aloud Protocols

Maaïke J. van den Haak
University of Twente
m.j.vandenhaak@utwente.nl

Menno D.T. de Jong
University of Twente
m.d.t.dejong@utwente.nl

Abstract

Think-aloud protocols are commonly used for the usability testing of instructional documents, web sites and interfaces. This paper addresses the benefits and drawbacks of two think-aloud variations: the traditional concurrent think-aloud method and the less familiar retrospective think-aloud protocols. It also offers an outline of a long-term research project designed to empirically investigate the value of both variants. The results of a first comparative study indicate that, although the two methods have distinct differences, they do seem to produce a similar outcome. A more detailed description of the results will be offered during the presentation. Keywords: usability testing, think-aloud protocols, methodology.

1. Introduction

In its most common form, usability testing represents either a test situation in which participants are observed while working silently with a particular test object, or a situation in which they simultaneously work and verbalize their thoughts. Even though the tasks involved and the laboratory situation are to a certain degree artificial for both methods, the first one mentioned, working silently, comes closest to regular working procedures. At the same time, the second method, concurrent thinking-aloud, has a clear benefit in that it allows insight into the participants' thinking process. This would seem to result in a more complete overview of user problems encountered: in addition to the observable problems (which constitute deviations from the optimum working procedure), the participants' verbalizations may reveal any doubts, irritation, surprise or other feelings that arise during the process. On the other hand, the

very fact that the participants verbalize their thoughts may also cause reactivity, i.e. participants may work differently from usual as a result of their thinking aloud. This difference may lead to a better or a worse performance, neither of which is desirable because in the first case, potential user problems do not come to light, while in the second case false alarms may be generated. According to Ericsson and Simon, who discuss think-aloud protocols for investigating cognitive processes, the risk of reactivity can be largely eliminated, if strict guidelines are observed [1]. However, in the context of usability testing, the potential bias of concurrent thinking aloud has received little scientific attention.

An alternative approach concerns the use of retrospective think-aloud protocols. This method involves participants first carrying out their tasks silently, after which they verbalize their thoughts in retrospect. In some cases, this retrospective verbalization takes place without any stimuli, which is likely to have a negative effect on the exhaustiveness of the comments produced [2-4]. In other cases, however, the retrospective verbalizations are supported by a recording of the performance. Nielsen, for instance, recommends using a video recording [5]; Henderson *et al.* used computer log files [6]. When verbalization is accompanied by stimuli, the retrospective think-aloud method potentially combines the benefits of both working silently and thinking aloud. All the same, it remains to be seen whether participants are indeed able to remember everything they thought during their task performance. What is more, they might actually come up with invented thoughts. Again, however, there is little empirical evidence with regard to the validity of the method in question.

In sum, it is clear that more research into the methodology of usability testing (and formative evaluation in general) is desirable. In an earlier overview of available research, De Jong and

Schellens show that there is a discrepancy between the popularity of think-aloud usability testing and the scientific attention that has been paid to the reliability and validity of the method [7]. To make up for this lack of attention, a research project was initiated which aims to shed light on the benefits and drawbacks of the three methods mentioned, i.e. participants working silently, thinking aloud concurrently, or thinking aloud retrospectively. This paper first offers a brief overview of the available research in the context of usability testing. It then provides an outline of the current research project.

2. Research available

The literature on usability testing tends to describe concurrent and retrospective think-aloud protocols as equal alternatives [5]. So far, only two studies have compared concurrent and retrospective think-aloud protocols.

Hoc and Leplat used the two types of think-aloud protocols to investigate a problem-solving process of participants (they had to order a set of letters on a computer screen using a limited set of commands) [8]. In the retrospective condition, participants were first asked to give an unaided account of their process, and after that had to think aloud while watching all the steps in the process, which had been recorded in a computer log file. They conclude that unaided retrospective accounts should be avoided, because of the distortions and gaps in the protocols, but that the retrospective and concurrent think-aloud protocols produce similar results. It should be noted, however, that both the task given to the participants (which more or less resembled a logical puzzle) and the analysis of the results (focusing more on strategies than on problems encountered) do not correspond to the situation of usability testing.

Bowers and Snyder compared the two think-aloud variations in a usability test focusing on the handling of multiple windows on a computer screen [9]. They found no significant differences regarding task performance and task completion time, but the retrospective think-aloud condition resulted in considerably fewer verbalizations, and these were often of a different type than the concurrent verbalizations, focusing more on explanations and less on procedures. While these results are interesting, the study has a serious drawback in that it does not report on the number and kinds of problems detected by the participants in the two think-aloud conditions. As

problem detection is typically one of the most important functions of usability testing, this meant that a crucial aspect was not included in the comparison of the two methods.

3. Outline of research project

In a long-term research project, the merits and restrictions of the methods as discussed above will be investigated using different test objects.

3.1 Research questions

Several aspects will be taken into account while comparing the three methods. Given the goals of usability testing, the most important aspect is the outcome in terms of problem detections. Both the number of problems and the nature of problems will be considered. To investigate the nature of the problems, a typology of problems will be developed. One issue to consider here is the manner of problem detecting, i.e. by means of observation, verbalization, or both. Another is the kind of problems detected, i.e. relating to layout, terminology, etc.

A second important aspect to consider with regard to the comparison of the methods is the participants' task performance. This is essential for investigating the reactivity of concurrent think-aloud protocols. In an ideal situation, one would expect that participants in a concurrent think-aloud condition are equally successful as participants working silently.

The third aspect for consideration involves participants' experiences during the test, i.e. how did they feel about carrying out the test situation, tasks, and thinking aloud (retrospectively)?

In sum, three research questions will be addressed:

- Do the methods differ in terms of numbers and types of usability problems detected?
- Do the methods differ in terms of task performance?
- Do the methods differ in terms of participant experiences?

3.2 Research design

The research questions will be investigated by means of three test objects: an online library catalogue, a household appliance plus manual,

and a web site. For each test object, a set of realistic user tasks will be formulated, which will be handed out to 20 participants per condition. The participant sessions will be held individually, and will be recorded on video tape. Each session ends with a questionnaire containing questions on participant experience. Once the sessions are over, all recordings will be analyzed with a view to problems detected and overall successful task performance.

3.3 A first study

A first study involved the testing of an online library catalogue. Its results indicated that the participants' verbalizations indeed resulted in more problem detections, compared to participants working silently. The extent to which these verbalizations complement the observable usability problems differs between the concurrent and the retrospective think-aloud condition. The added value of the verbalizations was more substantial in the retrospective think-aloud method. Overall, the two think-aloud methods resulted in similar numbers and types of problems.

One of the most striking results of our first study is that the participants in the concurrent think-aloud condition performed less successful than the participants who worked silently and verbalized in retrospect. This result is not only reflected in the number of observable problems per participant, but also in the overall success rate for the tasks. This may point to a certain degree of reactivity within the concurrent think-aloud method. A possible explanation lies in the workload of the participants, which together with the requirement to think aloud may have had a negative effect on task performance.

4. References

- [1] Ericsson, K.A., and H.A. Simon, *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, MA, 1993.
- [2] Branch, J.L. Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research*. 22: 371-392, 2000.
- [3] Kuusela, H., and P. Paul. A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology*. 113: 387-404, 2000.
- [4] Taylor, K.L., and J.P. Dionne. Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*. 29:413-425, 2000.
- [5] Nielsen, J., *Usability Engineering*, Academic Press, Boston, MA, 1993.
- [6] Henderson, R.D., et al. A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics*. 38: 2030-2044.
- [7] De Jong, M., and P.J. Schellens. Toward a document evaluation methodology: what does research tell us about the validity and reliability of methods? *IEEE Transactions on Professional Communication*. 43: 242-260, 2000.
- [8] Hoc, J.M, and J. Leplat. Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies*. 18: 283-306, 1983.
- [9] Bowers, V.A., and H.L. Snyder. Concurrent versus retrospective verbal protocols for comparing window usability. *Human Factors Society 34th Meeting, 8-12 October 1990* (Santa Monica: HFES), 1270-1274, 1990.

About the Authors

Maaïke van den Haak is a part-time PhD candidate at the University of Twente (The Netherlands). Her PhD research focuses on the merits and drawbacks of (variants of) the think-aloud method as an evaluation tool for instructive communication. Apart from her position at the University of Twente, she is also a part-time teacher of English at the Vrije Universiteit, Amsterdam (The Netherlands).

Menno de Jong is an associate professor of Communication Studies at the University of Twente (The Netherlands). His research interests concern the use and methodology of applied research to optimize communication. He has published about text and web evaluation, usability, and document design. He was co-editor of a special issue of *IEEE Transactions on Professional Communication* on document evaluation and usability testing.