Thinking about thinking aloud

Krahmer, E.J.; Ummelen, N.

Published in:
IEEE Transactions on Professional Communication

Publication date:
2004

Thinking About Thinking Aloud

A comparison of two verbal protocols for usability testing

Emiel Krahmer and Nicole Ummelen

Communication and Cognition
Faculty of Arts
Tilburg University
The Netherlands

e.j.krahmer@uvt.nl / ummelen@uvt.nl

Abstract

We report on an exploratory experimental comparison of two different thinking aloud approaches in a usability test that focussed on navigation problems in a highly non-standard website. One approach is a rigid application of Ericsson and Simon's (1993) procedure, the other is derived from a recent proposal by Boren and Ramey (2000), based on speech communication. The latter approach differs from the former in that the experimenter has more room for acknowledging (mm-hmm) contributions from subjects and has the possibility of asking for clarifications and offering encouragement. Comparing the verbal reports obtained with these two methods, we find that the *process* of thinking aloud while carrying out tasks is not affected by the type of approach that was used. The *task performance* does differ. More tasks were completed in the Boren & Ramey condition, and subjects were less lost. Nevertheless, subjects' evaluations of the website quality did not differ, nor did the number of different navigation problems that were detected.

Index terms Thinking aloud, usability testing, usability research, lostness.

Introduction

Over the last three decades, the thinking aloud method has been a widely-used instrument to study cognitive processes, such as problem solving, human-computer interaction, reading and writing. Participants in a thinking-aloud study are asked to carry out a task, while verbalizing their thoughts. The researchers record all verbalizations, write them down in a verbal report, and then analyze these in a way that depends on the research questions. Researchers may for instance pay attention to utterances that reflect a certain state of mind (annoyance, confusion), or to delays between utterances. Data can be analyzed in either a qualitative or a quantitative way.

The thinking aloud method has been used for three types of goals:

1. *To find evidence for models and theories of cognitive processes*: Newell & Simon for instance, developed a theory of human problem solving, and thinking aloud was an important instrument for collecting relevant data to support it [1]. Also, several researchers have been engaged in the development of a model of cognitive processes involved in writing. A well-known example is the model proposed by Flower & Hayes back in 1981, which was the starting point for a discussion about thinking aloud studies in writing research [2].
2. *To discover and understand general patterns of behavior in the interaction with documents or applications, in order to create a scientific basis for designing them*: Carroll et al. used thinking aloud studies to investigate how learners interacted with new software [3]. They found that learners

were annoyed by the large amounts of irrelevant information they encountered in tutorial manuals. Information in the manual appeared not to match with the users' goals and questions. Based on the analyses of many verbal reports, they described how software users learn to work with a new system and they accordingly developed a new design for software manuals: the minimal manual.

3. *To test specific new documents or applications in order to trouble-shoot and revise (usability testing, or pretesting, or formative testing)*: Schriver [4] [5:474] and Nielsen [6] among many others promoted verbal protocols as an instrument for testing and revising functional documents such as manuals and websites. The primary goal is to gather user information to support the design of a specific product.

Ericsson & Simon, whose research matches the first category, developed a theoretical framework and accordingly, a procedure for collecting valid and reliable thinking aloud data [7]. The procedure they proposed should prevent researchers from interfering with the subjects' cognitive processes and thus prevent them from affecting the research outcome. Their approach has been the background for many thinking aloud studies since the early 1980's. However, despite this framework and despite the large amount of thinking aloud studies, the method has been criticized more than once. Continuous doubts have been raised about validity and reliability of the method. Specifically, researchers have been investigating to which extent the thinking aloud method may affect the processes being investigated (e.g. [8] [9] [10] [11]). The results are not yet conclusive. Most results (but not all of them) show longer performance times for thinking aloud subjects, but no altered task outcome. This line of critical research focuses mainly on thinking aloud studies that contribute to the first two types of research goals: modelling cognitive processes, and creating design rules on the basis of behavioral user patterns.

Boren & Ramey, on the other hand, focus on the usage of thinking aloud for the third goal: *usability testing*. They recently pointed to a different and interesting type of problem [12]. Usability testers who work with the thinking aloud method, tend to refer to the Ericsson & Simon framework to account for their methodological choices. In reality however, usability testers do not meet the requirements imposed by the Ericsson & Simon framework. Apparently, they feel that the setting is too unnatural and that they may miss relevant data if they do not interfere during the test phase. Ericsson & Simon would claim that any data collected in interaction with the experimenter is not reliable. Boren & Ramey raise the question whether Ericsson & Simon are right in the context of usability testing. In other words: wouldn't a new theoretical framework, that accounts for the presence of an active listener (the experimenter), work better there? Boren & Ramey propose such a framework —based on speech communication theory— and they argue that the use of this new framework better matches usability practice.

The new framework raises various questions. Can it be used without precautions? Do the data differ from the data collected by means of the Ericsson & Simon approach, both in number and quality of the utterances, as is assumed? What happens to other performance measures that should not be affected by the observational approach? These are important questions, which have so far received little or no attention in the literature. With this paper, we would like to contribute to a discussion on the foundations of thinking aloud for usability testing. In order to do so, we first describe the theoretical foundation of thinking aloud as put forward by Ericsson and Simon. We then describe how thinking aloud is applied to usability testing, and which problems this created according to Boren and Ramey. We describe the discrepancies between theory and practice, and discuss the ramifications of these differences as well as Boren and Ramey's proposal for an alternative foundation. Then we describe a thinking aloud usability study that compares two groups: one following the Ericsson & Simon rules, the other one following the Boren & Ramey rules. We describe the differences and discuss their relevance for both usability practice and document or interface design research.

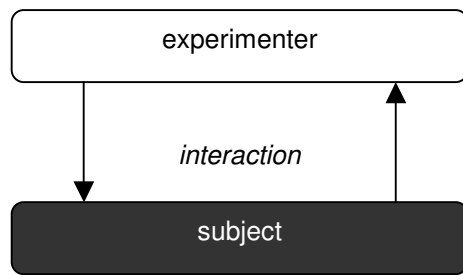## Thinking aloud: theoretical foundations (Ericsson and Simon)

Figure 1   Schematic experimental set-up for a thinking aloud study of cognitive processes.

Thinking aloud as a method for scientific research rests on a solid scientific foundation in cognitive psychology ([1] [7] [8] [13] [14]). It was originally developed for the study of short term memory (STM) processes of subjects (what they attend to and in what order), for instance, when performing problem solving tasks. Figure 1 provides the schematic setting of thinking aloud experiments in cognitive psychology. The focus of attention (black) is the subject's STM. Subjects are instructed to continuously verbalize their thoughts, which in effect means that there is some form of communication between subject and experimenter. In the ideal situation, this communication is purely single-directional: the subject continuously verbalizes his thoughts ("as if alone in the room") and the experimenter only listens.

It is important that the experimenter just listens; the subject's mental processes should not be influenced in any way. When the experimenter interferes, we can no longer be sure that the subjects' verbal data reflect the contents of their short term memory. After all, these contents may be affected by the experimenter's intervention either directly or indirectly, because the intervention causes subjects to elaborate where they would not do so otherwise. Hence, Ericsson and Simon argue, verbalizations that follow an intervention have a higher risk of being unreliable. Unreliable verbal data (or *level 3 data* in the terminology of Ericsson and Simon) should not be used for further analysis. There is one other cause for concern: if the subject keeps silent for a longer period of time, the verbalization will also become unusable, because significant parts of the cognitive process in STM may not be tracked down. To avoid this, the experimenter under the Ericsson and Simon regime is allowed to remind the subject to think aloud if (s)he falls silent. This reminder should be short and non-intrusive, in order to minimize the risk of ending up with level 3 data. Ericsson and Simon propose to only use the phrase "Keep talking."

Thinking aloud is unnatural. Therefore, Ericsson & Simon recommend an initial practice session in which subjects are 'taught' to verbalize their thoughts. "During warmup, the experimenter feels free to interfere with and disrupt the subject, while during the experiment, he should be very concerned not to interfere" [14:82]. In addition, subjects should learn the difference between describing what they are doing ("I now move a disk from here to there.") and thinking aloud ("Since this disk is smaller than that one, I put it on a another pin first.").

In sum, performing a thinking aloud experiment according to the Ericsson and Simon method implies the following rules:

(1) When the subject keeps silent for a longer period of time, the experimenter should provide a reminder ("Keep talking").
(2) Other than that, the experimenter should not interfere during the thinking aloud process.
(3) To make subjects familiar with the thinking aloud method they should be trained in advance.

Thinking aloud for usability testing

Thinking aloud is a popular and effective method for usability testing (see e.g. [6]). It may provide us with useful information about users who are interacting with a certain application. The ―mainly qualitative― data provide indications of application areas that cause user problems, and these data may be used for further development. In addition, actually seeing users struggle with parts of an application is usually a compelling argument for developers to further improve their product.
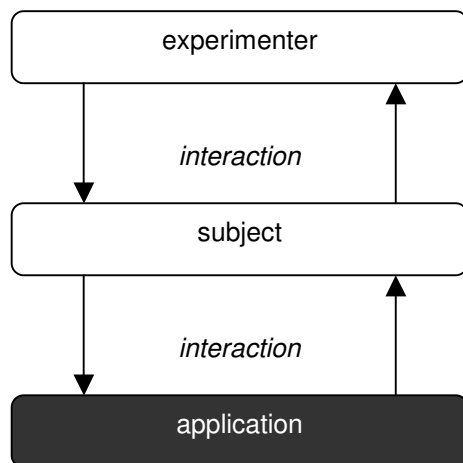
```
                    ┌─────────────────────────┐
                    │     experimenter        │
                    └─────────────────────────┘
                      │                    ▲
                      │    interaction     │
                      ▼                    │
                    ┌─────────────────────────┐
                    │       subject           │
                    └─────────────────────────┘
                      │                    ▲
                      │    interaction     │
                      ▼                    │
                    ┌─────────────────────────┐
                    │     application         │
                    └─────────────────────────┘
```

Figure 2   Schematic experimental set-up for a thinking aloud study of usability.

Thinking aloud as a usability test is different from thinking aloud as a research method. When thinking aloud is employed in the context of a usability test (see Figure 2) the schematic setting is subtly different from that discussed above (cf. Figure 1). The focus of attention (black) is not so much the subject as the *application* the subject interacts with. So, essentially, there are *two* interactions going on: one between subject and application and one between experimenter and subject. Naturally, this by itself does not mean that a usability practitioner could not adhere to the Ericsson and Simon method, and in fact most refer to Ericsson and Simon as the theoretic foundation of their approach (that is, if they provide a source at all).

Yet, Boren and Ramey pointed out that "theory and practice are out of sync" when it comes to thinking aloud as a usability method. On the basis of a literature study and field observations, Boren and Ramey [12:264-265] give four examples of discrepancies between theory and usability practice.

(1) Thinking aloud instructions are often not given in the prescribed manner. For instance, the difference between thinking aloud and explanation is not always made clear, and usually subjects are not offered the possibility to practice thinking aloud.
(2) Reminders often diverge from the prescribed "Keep talking". For example Dumas and Redish [15, page 281] use "John, could you tell us why you pressed the enter key". Clearly, a reminder like this is very likely to interrupt the subject's thinking as it redirects his attention to the enter key and requests an explanation. In general, a large variety of experimenter contributions is encountered in practice, both in phrasing and timing.
(3) When not reminding, experimenters intervene in other ways that are inconsistent with the theory. They "probe" for more information, they "answer a question with a question" [6], they provide clarification when the subject is stuck, while the Ericsson and Simon method prohibits all interventions from the experimenter except for the "keep talking" reminder.
(4) Many practitioners do not use the verbalizations as "hard" data. They appear to prefer explanations, evaluative remarks and suggestions for improvement or alternative designs more than information about the cognitive processes of the subject. This suggests that level 3 data are considered to be more useful for the usability practitioner than "hard" data.

Does the mismatch between theory and practice matter?

According to Boren and Ramey [12] the tried and tested Ericsson and Simon approach to thinking aloud is claimed to be the foundation of thinking aloud for usability testing, while in fact few usability-testing handbooks and usability testing professionals actually follow it. But if one departs from the Ericsson and Simon protocol, there is no longer a solid theoretical framework to motivate usability practice, and it becomes difficult to distinguish correct from incorrect test performance ([12, page 261]. Moreover, if the method used by practitioners is not theoretically motivated and shown to be reliable and valid, it becomes difficult to compare or replicate studies ([12, page 266) or to show that a redesign is an improvement over the version. Several other researchers in the field of ergonomics and document design have stressed the importance of reliable and valid usability testing methods (e.g., [23] [24] [25]) on the basis of comparative analyses of evaluation studies and discussions with experts.

To see why a departure from the Ericsson and Simon approach to thinking aloud may potentially cause problems, suppose that we perform a usability test with the sole purpose of evaluating the usability of the target application in order to redesign it (i.e., the third type of goal mentioned in the introduction). Then perhaps we may not need to adhere strictly to the Ericsson & Simon method, because we need not worry about 'polluting' the verbal representation of cognitive processes like we would do if we were studying the *interaction* between subject and application (the second type of goal) or the way subjects construct a mental model (the first type of goal). Yet, we should wonder whether or not we would gather valid and reliable usability data if we perform the thinking aloud test in this way. It still seems important to find real and relevant user problems, not ones that might be evoked by the test setting. We will give an example to illustrate this. An experimenter asking "Why did you press the enter key?" may lead the participant to return his attention to the enter key. He may then elaborate on the reasons for the decision to press 'Enter', which in turn may result in reflections about the logic of the document or the system that would not have occurred without interventions by the experimenter. These reflections may in turn lead to different search patterns and decisions in the remaining part of the usability test: user problems may be missed that would have occurred otherwise, and user problems that do occur might not have occurred otherwise. In this way, the experimenter's interruption may influence the results of the usability test (in terms of the detected usability problems). In our opinion, this would run counter to the idea of using thinking aloud to gain insight in what users think of a certain application. And to the extent one considers this a problem, validity and reliability are arguably just as important in usability tests as they are in experimental research.

One possible alternative for the 'out-of-sync-practice' would be to apply the Ericsson and Simon framework as it was intended, and create 'work-arounds' to avoid the practical problems Boren and Ramey pointed to, such as how to avoid experimenter interference when software or prototype bugs occur. A strict interpretation of the Ericsson & Simon framework would imply that the data *following* a necessary intervention may not be reliable and should thus be discarded. Other data, such as the preceding verbalizations that pointed to the usability problem, may still be used. As long as this rule is obeyed, validity problems should not occur. In a similar vein, the fact that practitioners actually *value* level 3 data (suggestions for improvement, evaluation, etc.) is not necessarily incompatible with the theory either, provided that detailed questioning of the subject takes place *after* completion of the thinking aloud test. Would the Ericsson & Simon approach with these work-arounds be feasible in usability testing, or would we still need an alternative approach?

An alternative proposal

Boren and Ramey argue that the Ericsson & Simon framework, whether applied strictly or not, does not work adequately in the context of usability testing, as the goals and circumstances in usability testing differ too much from those in Ericsson & Simon's research. Therefore, they propose an alternative theoretical foundation for thinking aloud in a usability context: *speech communication.*

Their central observation is that the interaction between subject and experimenter consists of real-life spoken communication between two people: one who does most of the talking (the subject) and one who primarily listens (the experimenter). Boren and Ramey do not consider the subjects' 'verbalizations' as a monologue, but rather as part of a dialogue. The main difference is of course that this entitles the person who does most of the listening to occasionally *contribute* to the dialogue as well.

According to Boren and Ramey, conducting a thinking aloud test based on a speech communication framework, could be helpful in a number of ways: (1) in *setting the stage*, (2) in *keeping verbalizations continuous*, (3) in *dealing with technical problems* and (4) in *proactive elicitation*, where each step takes us further away from the principles laid down by Ericsson & Simon. We now take a closer look at each of these four aspects.

(1) In the preparation phase of the experiment, a speech communication perspective may help in setting the stage. The various roles can be clearly defined: the application is the *topic* of the communication, the participant is the 'work domain expert' and the main *speaker*, and the experimenter is the 'learner' and the primary *listener*. Note that this is rather different from the Ericsson and Simon instruction that the subject should be "speaking as if alone in the room".

(2) From a communication perspective, "keep talking" is not a very natural contribution. Boren and Ramey propose using the nature of communication to keep a continuous flow of verbalization. They consider information exchange to be not just a speaker's task, but a shared task for speaker and hearer. Similar ideas have been expressed by Clark and Schaefer [16], among others. Clark and Schaefer model information exchange as a two-step procedure: first the speaker has to send the information and second the listener should acknowledge the receipt. In other words, "the listener **must** respond" (as Boren and Ramey state [12, page 267]). Boren and Ramey argue that *acknowledgement tokens* such as "mm-hmm" (a.k.a. conversational grunts, [17]) are well-suited for this. They point out that intonation (or speech melody) is an important cue for both the timing and the production of these tokens: subjects "ending a statement with a stretched, interrogative intonation contour creates a slot for an acknowledgement token" and if the corresponding slot-filler from the experimenter "also displays an interrogative intonation, another slot is opened up for the previous speaker to retain speakership." [12:270]. For example:

S   (…) and then I move this disk here?
E   Mm-hmm?
S   Now I (…)

(3) Furthermore, speech communication may be useful in those situations where an intervention from the experimenter is simply required, for instance, in case of a breakdown of the application which is the topic of the conversation. In such cases, the roles of speaker and listener may be temporarily reversed; from an asymmetric dialogue with the subject doing most of the talking to a symmetric dialogue in which the experimenter solves the problem, and back to an asymmetric dialogue after the problem has been solved.

(4) Finally, and in a similar vein, if the experimenter would like the subject to clarify a remark or would like to redirect the attention of the subject, he can simply do so given his role in the dialogue.

Speech communication: discussion

Boren and Ramey set out to reconcile theory and practice. Obviously, stating that the experimenter plays a predefined and active role in a communication process between subject and experimenter, provides a better fit of current practice than Ericsson and Simon's method.
The first contribution of Boren and Ramey's speech communication approach is a different way to set the stage and prepare the subject for the experiment. The roles are defined of the subject as the primary speaker, the experimenter as listener and learner and the application as the object of interest. In contrast, Ericsson and

Simon stress that the experimenter should not play an active role and should be unnoticed by the subject during the experiment.

The second contribution of speech communication (the use of 'mm-hmm' instead of 'keep talking') is a different way of reminding subjects to think aloud when they fall silent for a longer period of time. The use of acknowledgement tokens confirms a two-way interaction between experimenter and subject. In this respect it is different from Ericsson and Simon's approach, that only allows a short, non-intrusive utterance such as 'keep talking', assuming that this is an extraneous intervention. Ericsson & Simon emphasize that it is important to let the subject talk 'as if alone in the room', and feedback cues are not compatible with this requirement: they may be interpreted by the subject as an implicit confirmation (cf. the example given above).

In spite of these considerations, the differences between the two thinking aloud approaches are still relatively small. It is worth pointing out that Ericsson and Simon's theory is not necessarily incompatible with the idea that a listener must respond. A feeling of 'dialogue' might also be evoked by a person saying 'keep talking', in spite of the precautions that Ericsson and Simon prescribe. Thus, it is not uncommon for a subject to respond to an interruption like "keep talking", as is illustrated in the following exchange.

> S   (…) uhm [silence for 20s]
> E   Keep talking.
> S   Keep talking he says! As if (…)

Arguably, such an 'angry' subject response (which arguably interrupts the 'natural' flow of information) is less likely following a back channel cue.

The remaining suggestions by Boren and Ramey —dealing with necessary interruptions and probes for clarification or further information— do not fit into Ericsson and Simon's approach. These additions may be useful for the usability practitioner and account for their occurrence in usability practice, but it is unclear whether the results are reliable and valid in the sense explicated above.

Research question

In summary, usability practitioners have three options, each with some (expected) advantages and disadvantages:
(1) Apply the Ericsson & Simon framework with several unspecified deviations from the strict original procedure. This is what Nielsen [26] has dubbed "simplified thinking aloud".These deviations are expected to yield more and better usability data than the original Ericsson & Simon framework, but they may also cause validity or reliability problems, even though the data will only be used to troubleshoot a specific product.
(2) Apply the Ericsson & Simon framework in its original strict way; this should guarantee validity and reliability in any kind of research context, but practitioners fear that it may yield fewer and less informative data in the context of usability testing;
(3) Apply the new Boren & Ramey framework of speech communication; the dialogue between experimenter and subject is expected to yield a valuable amount of relevant usability data for product testing contexts.

The question is whether or not the detected problems for the third option (Boren and Ramey) are different from those detected when using the second option (Ericsson and Simon). In other words: to what extent do the different setting and the different intervention types affect the number and type of problems that are detected? In fact, Boren and Ramey state that "(…) further research should be done to determine the extent to which such probes (…) affect subsequent verbalization and task performance" [12, page 275].

In order to further this discussion, we conducted an explorative study to examine the differences between the approaches. We focused on two actual frameworks: the original Ericsson & Simon approach on the one hand, and the Boren & Ramey approach on the other hand. Based on the expectations of Boren & Ramey we were especially interested in differences in the number of (relevant) utterances and differences in the quality

of the utterances for usability practice. Furthermore, we looked for differences in effects of the approach on other performance measures (reactivity issues).

An explorative experimental comparison of two usability tests

Introduction

For this explorative study, we conducted two variants of a usability tests under controlled circumstrances, one variant based on the Ericsson and Simon protocol, the other on Boren and Ramey's proposal. Our target application was *Het Nieuwe Lichaam* (translated: The New Body), the official web site dedicated to the Dutch writer Harry Mulisch (`www.mulisch.nl`). Mulisch (born 1927) is generally considered to be one of the three most important post-1945 Dutch literary writers. His web site is highly unconventional and technologically advanced. At the time of writing this article (late 2003), the site was visited by more than 750.000 times since it was launched on March 14th, 2000. The site was designed to be a journey through Mulisch's mind; it was supposed to be a metaphor for the way he thinks [18]. This is symbolized on the home page by a dark castle (which has been likened to the works of Piranesi (1720-1778), the Italian engraver and architect) with various entrances, corridors and windows (Figure 3). Each of these features represents one of the themes of Mulisch' work, and visitors can explore these themes by navigating through the relevant parts of the castle. Standard hyperlinks are absent. Instead visitors have to manipulate virtual objects to navigate.

Figure 3 Two screens from the Mulisch-website: the homepage (left) and a subpage ('time machine').

The usability research question we addressed in this study was the following: what kind of problems, if any, do subjects experience in the navigation structure of the Mulisch site? Do they experience the site as the inspiring voyage the makers intended it to be, or as an information source in which they become lost? And, if so, how lost will subjects become in this navigation structure?
We designed a thinking aloud study in order to find answers to these usability questions. To explore the differences between thinking aloud approaches, we created two groups of subjects, each following a different thinking aloud procedure. One group followed the Ericsson & Simon rules, the other group followed the Boren & Ramey rules. By comparing these two groups we will attempt to find out if there is any influence of the verbal protocol on task performance, on the amount of thinking aloud data and on the quality of the data. These issues are addressed below.

Method

Tasks

A preliminary expert evaluation was performed to determine potential usability bottlenecks in the Mulisch site. The tasks used in the thinking aloud study were formulated in such a way that subjects were likely to encounter these bottlenecks during the navigation, in order to see if they were really usability problems. These were the tasks:

- What are the eight themes of the Mulisch web site?
- What are the names of the two reviewers for "De ontdekking van de hemel" (The discovery of heaven) quoted on the site?
- In which year was the book "Zielespiegel" (Soul's mirror) published?
- Which two short novels are part of "Chantage op het leven" (Blackmail on life), published in 1953?
- Try to obtain a free ticket for the movie "The discovery of heaven".

All subjects received a booklet with these 5 tasks. Each task was placed on a new page, together with a designated space where the subject could write down the answer. In this way, it was clear for subjects when a particular task was completed.

Experimental setting: the protocols

The E&S protocol was defined in accordance with the rules put forward by Ericsson and Simon [14] [12:263]. Before the actual experiment subjects were trained to think aloud by letting them verbalize their thoughts while solving a towers of Hanoi problem via a computer simulation (cf. [19]). Subjects were then given the booklet containing the tasks, with explicit indications of when a particular task was carried out successfully. When a subject did not make any progress for 5 minutes, (s)he was allowed to move on to the next task. When a subject fell silent for 15 to 20 seconds, the experimenter used the typical Ericsson and Simon reminder "Blijf praten" (Dutch for keep talking). Other than that, there was no interaction between the subject and the experimenter.

The B&R protocol was defined in accordance with Boren and Ramey's framework [12:266ff]. First, the setting was defined: the Mulisch site was explicitly defined as the topic (the system being tested), the subject as the application expert and primary speaker and the experimenter as the learner and primary listener. Then, before the actual experiment, subjects were trained to think aloud in the same way as subjects in the other experimental group. Subjects were also given the booklet with 5 tasks, and when a subject did not make any progress on a given task for 5 minutes, (s)he was allowed to move on to the next one. Each time subjects ended a statement with "a stretched, interrogative intonation contour" thereby creating "a 'slot' for an acknowledgement token" [19 cited in [12]], the experimenter used the typical Boren and Ramey reminder "mm-hmm" with a rising intonation [12:270]. This token was also produced when a subject fell silent for 15 to 20 seconds, according to Boren & Ramey's suggestion for extending the use of "mm-hmm" to situations where there is 'nothing to acknowledge' [12:271]. In addition, the experimenter was allowed to make two other kinds of contributions to the dialogue. When a subject was unclear, the experimenter was allowed to ask for clarification, by repeating "a single word with the proper inflection" [12:275]. And when a subject was really stuck, the experimenter was allowed to "encourage the participant to continue" [12:274] and offer indirect suggestions of how to do so. These suggestions never directly contributed to solving the task. They only put the subject back on the 'right' track.

The respective protocols were administered by two different experimenters. Both were trained by the authors in doing thinking aloud tests and had exactly the same experience with and knowledge of administrating such tests.

Except for these, there were no differences between the two conditions.

Subjects

Ten subjects participated in the experiment; five for each protocol. Five subjects is a more or less standard number for a thinking aloud usability test, and in this sense the results are representative for actual think aloud usability tests (cf. [23, page 197]). Of course, this number is relatively small for an experimental study, but we expected that the data collected in this explorative study could still be informative and further our thinking. None of the subjects had any former experience with thinking aloud studies. All subjects had substantial internet experience. They were between 20 and 25 years old, native speakers of Dutch and they read at least one Dutch literary work per month. The interest in Dutch literature was added to make sure that subjects are likely to take a natural interest in the Mulisch site. Subjects were randomly assigned to one of the two thinking aloud approaches.

Data processing

The utterances by both the experimenter and the subjects were recorded and their verbalizations were transcribed in verbal reports, and combined with a registration of mouse clicks.

Navigation usability measures

To gain insight in the usability issues (i.e. in navigation problems) we derived three measures from the transcriptions. In general, the kind of verbal protocol should not influence the scores related to the issue being tested: there should be no differences in navigation usability between the E&S protocol and the B&R protocol. The three usability measures are:

(1) Utterances conveying the subjects' ideas or experiences in relation to the navigation structure. This qualitative analysis should not yield any differences in judgements of the website being studied.
(2) Number of mouse clicks necessary to complete the task. We measured and compared subjects' navigation in terms of the total amount of clicks that were necessary for the tasks.
(3) Finally, a quantitative measure for lostness was used: Smith's L formula [21]:[1]

$$L = (N/S - 1)^2 + (R/N - 1)^2$$

N is the number of different nodes encountered when searching, S is the total number of nodes visited during the search and R is the required, minimal number of nodes. Theoretically, the L formula ranges from 0 (not lost at all: N = S = R, i.e., no divergence from the minimal path) to 2 (completely lost). In practice, an L value of 2 will never occur, since the total number of links in the site provides an upper bound to N (a subject can never see more different links than there are in the site). According to Smith's observations, subjects are lost for values of L higher than 0.42.[2] Subjects' lostness should not depend on the thinking aloud protocol.

Protocol measures

In order to compare the two thinking aloud groups, we derived four measures from the verbal reports. For some of these measures we expect differences between the two protocols (since the role of the experimenter is different), but, as above, these differences should not result in different user behavior related to the task.

---

[1] Otter and Johnson [22] criticize Smith's lostness measure in that it does not take the nature of the hyperlink into account. For our purposes, this is not a problem since all links in the Mulisch site are of the same type.

[2] One problem with the L measure is that it does not take task failure into account. We assume that a subject who fails to carry out a task, is lost. However, if we simply count the number of nodes visited until the subject gave up and take this as our S value, the lostness is not accounted for. Hence, in the case of task failure we let S range to infinity (it would take an infinite number of nodes to fulfil the task) and compute the limit of the L formula.

(1) Number of utterances by the experimenter. This is a manipulation check rather than a comparison measure, as the difference between the method condition already implies that the experimenter in the B&R protocol will speak more frequently than the experimenter in the E&S protocol. In addition, and to gain more insight into the type of interventions, we classified the utterances in reports obtained with the B&R protocol into different types and looked at the distribution of the utterance types by the experimenter.

(2) Number of successfully executed tasks (correctly answered search questions) by subjects. If task success for the B&R protocol is higher, there is reason to believe that this approach causes validity problems.

(3) Number of words uttered by subjects. Boren & Ramey assume that the speech communication approach naturally leads to more thinking aloud data, which would be beneficial for usability testers.

(4) Quality of subjects' utterances for usability research. Boren & Ramey assume that the subjects' utterances in the speech communication approach will be more relevant for usability testers than in the Ericsson & Simon approach. In this study, usability testers collected data on navigation usability. Therefore, we counted the numbers of different types of navigation problems, mentioned in the verbal reports. First, the reports were analyzed to find all utterances referring to usability problems. Then these problems were divided into seven categories. For a description, see Table 1. If more utterances by one subject referred to the same usability problem in a certain task, the problem was only counted as one, unless a new aspect of the same problem was introduced (e.g. a different reason, different interpretation); then the problem was counted as a separate problem. If the same problem occurred twice in the same task, it was counted as one. If different problem types occurred while navigating in the same screen parts or handling the same screen objects, the problems were counted separately.

Table 1
Classification of the navigation problems encountered in the verbal reports; examples are typical extracts from the verbal reports gathered for this study.

Problem type	Description	Typical protocol-items signalling the problem
Uncertainty about action planning	Subjects do not see where they possibly could go (click) next, or they see several possibilities but haven't got a clue about which one to select	I see books, I see a lot of things...but where should I go now? Can I click this at all? OK, let me see of something is clickable here.
Orientation	Subjects do not understand where they are or cannot interpret their current location in the context of other locations in the website	What does this mean? Where am I now? I guess I should be in a completely different part of the website?
Stuck in loops	Subjects think they move on to a different location in the website but appear to (repeatedly) return to where they came from	… hey, I've been here before! **, again this stupid page.
Unexpected result	Subjects expect a certain result after clicking a link, but this result does not occur	…and then I click this and hope something happens…but it doesn't.
Failed repetition of actions	Subjects think they remember how to navigate because they feel they did that before, but the (assumed) repetition fails	Oh … yes…and now I should be able to select a title here…. Oh no, apparently I can't.
Reasoning about navigation logic	Subjects start reasoning about why the interface makes them think they can find certain information behind a link	A clock that is ticking… ah, perhaps that's a time machine. Let me try that….
Interface manipulation problem	Subject have problems handling certain objects in the interface (e.g. dragging a pointer over a calendar in order to get to information about a certain year)	There's a calendar here, but when I click it nothing happens… how can I do this?

Statistical analyses
To test for statistically significant differences between the means of the two groups (for subjects under the E&S and the B&R protocol respectively), we used one way analysis of variance (ANOVA) tests.

Results

Navigation usability

We will first discuss the usability research question in relation to the thinking aloud groups: how do subjects experience the navigation structure of the Mulisch site and is this in accordance with the premises of the makers? The results are unequivocal: most subjects (irrespective of protocol type) perceive the navigation as a voyage of discovery, but *not* in the positive sense. Subjects do not consider the search to be an "intellectual and literary adventure" (as the makers intended) but a frustrating and complicated affair instead. The following representative quotes from the transcriptions may illustrate this (translations by the authors). First a number of quotes from verbal reports obtained with the E&S protocol:

> [**E&S subject 1**]
> I feel a little as if what I'm doing is hopeless.
> [**E&S subject 2**]
> Well, it must be a mess in his head, because there is
> nothing clear here.
> [**E&S subject 3**]
> You would think that it could be found here. Uh, that's weird. OK start
> Again.
> [**E&S subject 4**]
> Uh sometimes I don't know whether some areas on that picture (...)
> whether it is the same link or whether it is a different one.
> [**E&S subject 5**]
> I am just doing something (...) because I don't really know where to
> look.

And here are examples from the verbal reports obtained with the B&R protocol:

> [**B&R subject 1**]
> I cannot imagine that someone would do this for fun.
> [**B&R subject 2**]
> I miss a search facility (...) it is very unclear.
> [**B&R subject 3**]
> Very incoherent
> [**B&R subject 4**]
> Yes. It's very unclear. Everything looks the same each time.
> [**B&R subject 5**]
> It is disgusting. Some structure would be handy.

Only two subjects (one for each method) indicated that they thought the site was fun and that they appreciated the design. The others were primarily hindered by the highly unconventional set up of the site.

Table 2 presents the average number of mouse clicks the subjects needed per task.

Table 2
Average number of mouse clicks in the two thinking aloud conditions per
subject and per task (Standard deviations between brackets).

	E&S PROTOCOL	B&R PROTOCOL
# Clicks per subject (5 tasks)	231.6 (80.7)	202.4 (69.5)
# Clicks per subject per task	46.3 (16.1)	40.5 (13.9)

The average number of mouseclicks needed per task illustrates how difficult it was for subjects to find the answers to the question; subjects needed 40.5 to 46.3 clicks on average for answering one single question. Yet, these numbers did not differ significantly between the thinking aloud conditions (*F(1,8)<1*).

Finally, we measured lostness (in terms of Smith's L-measure) in both thinking aloud conditions in order to gain insight into navigation usability. Table 3 presents the average lostness subjects experienced while being confronted with the Ericsson and Simon or the Boren and Ramey method.

Table 3
Average lostness (Smith's L measure) in the two thinking aloud conditions
(min 0, max 2; standard deviations between brackets).

	E&S PROTOCOL	B&R PROTOCOL
Lostness (L)	1.22 (0.2)	0.75 (0.1)

Overall, the quantitative lostness-measure shows that all participants were lost to some extent (the outcome would be zero if no lostness had occurred). However, subjects in the Ericsson & Simon conditions appeared to be significantly more lost than subjects in the Boren & Ramey condition (*F(1,8)= 21.92, p<.01*).

Protocol measures

Table 4
Average number of interventions by the experimenter per subject in the two
thinking aloud conditions (Standard deviations between brackets).

	E&S PROTOCOL	B&R PROTOCOL
# Interventions	15.8 (10.9)	133.0 (22.1)

We first checked to see if there were any differences in the number of interventions by the experimenter, as we would expect. Table 4 gives the number of interventions by the experimenter as a function of method. The number of times the experimenter said something is significantly higher for the B&R protocol than for the E&S protocol (*F(1,8)=113.48, p <.001*).

Table 5
Distribution of the total number of interventions by the experimenter in the B&R protocol.

	BACK CHANNELS	CLARIFICATIONS	SUGGESTIONS	OTHER	TOTAL
n	477	71	77	40	665

We then took a closer look at the *kind* of interventions (Table 5) offered by the experimenter in the Boren & Ramey condition. The vast majority of the experimenter's utterances appear to be back channel cues ('mm-hmm') (477, or 71.7%). In 10.7% of the cases the experimenter prompts for clarification by repeating a

word. 11.6% of the interventions consist of suggestions to direct the subject back on the right track. The remaining 6% of the interventions (40) perform a combination of the various functions.

Table 6 presents an overview of task success in the two thinking aloud groups.

Table 6
Average number of tasks successfully executed by the subjects in the two thinking aloud conditions (min 0, max 5; Standard deviations between brackets).

	E&S PROTOCOL	B&R PROTOCOL
Task success	1.4 (0.9)	3.8 (0.5)

Clearly, the subjects under the E&S PROTOCOL have much more difficulty in correctly performing the tasks than the subjects under the B&R PROTOCOL. The average task success for the E&S PROTOCOL is 1.4, while for the B&R PROTOCOL it is 3.8. This is a statistically significant difference ($F(1,8)=28.8, p < .01$).

Table 7
Average number of words uttered by subjects, per subject and task in the two thinking aloud conditions (Standard deviations between brackets).

	E&S PROTOCOL	B&R PROTOCOL
# Words in protocols	353 (189)	326 (80)

One might expect that the differences in experimenter contribution also have repercussions on the contributions of the subjects. However, this appears not to be the case. Table 7 shows the number of words uttered by a subject as a function of method. There are substantial differences among the subjects, but on average subjects utter 326.4 words per task under the B&R PROTOCOL subjects, while under the E&S PROTOCOL this is 353.2 words. This difference is not statistically significant ($F(1,8)= 1.24, p= .30$).

Interestingly, there were no significant differences in the numbers of detected usability problems (Table 8).

Table 8
Average number of different navigation problems that were referred to in the respective verbal reports (standard deviation between brackets).

	E&S PROTOCOL		B&R PROTOCOL	
Uncertainty about action planning	16.8	(10.5)	23.2	(6.4)
Orientation	7.6	(3.2)	12.6	(6.5)
Stuck in loops	4.8	(2.9)	1.6	(1.1)
Unexpected result	7.6	(5.2)	6.0	(2.9)
Failed repetition of actions	1.8	(1.3)	1.6	(1.1)
Reasoning about navigation logic	6.4	(6.2)	4.8	(2.6)
Interface manipulation problem	1.8	(1.6)	1.2	(0.8)
Total	46.8	(15.5)	51.0	(5.4)

(Uncertainty: $F(1,8)=1.3, p=.28$; Orientation: $F(1,8)= 2.3, p= .17$; Loops: $F(1,8)=5.1, p=.053$; Unexpected result: $F(1,8) < 1$; Failed repetition: $F(1,8) < 1$; Reasoning about logic $F(1,8) < 1$; Interface manipulation $F(1,8) < 1$).

All problem types occurred about equally often in both thinking aloud conditions. The only difference that approaches significance is being stuck in a loop; subjects in the E&S PROTOCOL are more likely to get stuck in a loop than subjects in the B&R PROTOCOL, which is in accordance with the lostness figures mentioned above. Moreover, the B&R PROTOCOL seems to detect somewhat more orientation problems and more uncertainty about action planning. But overall there are high standard deviations and no significant differences, so that we cannot conclude that the B&R PROTOCOL detects significantly more usability problems than the E&S PROTOCOL.

Discussion and conclusion

Two different verbal protocols were compared for a usability evaluation focussing on navigation problems for a highly non-standard website (Mulisch). The E&S protocol is a rigid application of the theory presented in Ericsson and Simon (1993), the B&R protocol is derived from the recent proposals of Boren and Ramey based on speech communication. The E&S protocol does not allow the experimenter to interact with the subject, except for a "keep talking" reminder when a subject falls silent for more than 15 to 20 seconds. The B&R protocol is different in three respects: (1) the experimenter gives "mm-hmm" tokens as feedback (and as reminders to continue verbalization after 15-20 seconds of silence), (2) the experimenter is allowed to repeat a single word to trigger clarification, and (3) the experimenter is allowed to encourage a user who is stuck by giving a non-directive suggestion.

When we compare the verbal reports obtained with these two methods an interesting picture emerges. The process of thinking aloud did not seem to be affected by the type of approach that was used. In the two conditions, subjects used equal numbers of words (and equal numbers of clicks as well). The number of interventions by the experimenter was much larger for the B&R protocol, as a natural result of the differences in approach. A perhaps more surprising finding was that more tasks were completed in the B&R condition than in the E&S condition. Moreover, subjects were less lost in the B&R condition than in the E&S condition. Nevertheless, the subjects' evaluations of the website quality did not differ for the two protocols (as illustrated by the quotes from the transcripts), nor did the number of different navigation problems that were detected.

We have to be careful in drawing conclusions from these results, since this was a relatively small, explorative study. But, the results suggest that applying the B&R protocol might be a mixed blessing. Subjects are less lost and can successfully complete more tasks (and it is worth stressing that even though the two tests used a limited of number of subjects, these are clear and significant differences). On the one hand this suggests that subjects have an 'easier' time performing the test in the B&R condition. But on the other hand, it also suggests that subjects are influenced on some of the dimensions under investigation (navigation usability). This would imply that the Boren & Ramey approach may cause validity problems in usability studies that use task performance or lostness as usability measures, since apparently these measures are influenced by the experimenter's interventions. This difference is especially interesting in view of the fact that subjects in both conditions had the same subjective user experience (in both groups, four out of five subjects disliked the navigation on the site), and that the transcripts of both groups reveal comparable average numbers of different navigation usability problems (cf. Table 8).

In sum, the reported study suggests that the way a thinking aloud usability test is administered can have an influence on the results. Of course, this finding is based on only one comparative study, so additional research is called for. Even though some of the findings were strong enough to be statistically significant, it would be useful to replicate this finding with a larger set of subjects, and preferably also with a single person administrating both protocols (although we conjecture that the basic picture will not change drastically as a result of this). Another replication study could involve a different kind of application. The Mulisch site is a highly unconventional one, which implies that participants will in general have more difficulty navigating though this site than through a more conventional one. It would be highly interesting to see which, if any, differences between the two protocols would arise when the test object is a more conventional application. We would hypothesize that subjects require less guidance from an experimenter, and hence differences

between the two protocols with respect to task success and lostness would tend to get smaller than in the current study. Besides such relatively straightforward replication studies, it would be interesting to perform some more detailed experiments testing the relative contributions of the speech communication protocol. We would like to know more precisely, for instance, which aspects of the B&R protocol cause the differences between the two protocols. We conjecture that the frequent use of feedback cues (mm-hmm) stimulates users to continue searching while the occasional encouragement or indirect suggestion may put subjects back on the right track again. The verbal reports collected here do not indicate a clear one-to-one correspondence between suggestions and task success, but this may be due to the contextual driven-ness of experimenter contribution. In this respect it would be useful to perform a replication study where the administration of prompts and probes is carefully balanced, so that potential correlates with lostness and task success can be identified. Naturally, prompts and probes form a more dramatic departure from Ericsson and Simon's approach than the use of backchannel cues (mm-hmm). These seem to offer a useful alternative to the unnatural "keep talking", which, as we have seen, may also create the illusion of conversation even though the test administrator is careful to avoid this. In the current comparison study, we used pauses of 15 to 20 seconds before a 'reminder' "mm-hmm" was offered. It would be interesting to perform an experiment in which the length of this time delay is varied. We conjecture that the timing of back-channel cues could even be done automatically, for instance administrating an mm-hmm following low pitch regions of 110ms or longer (Ward and Tsukahare 2000). It would be interesting to investigate the possibility of taking the experimenter "out of the loop" during thinking aloud studies, and use automatic low pitch detection techniques to trigger acknowledgements. This would offer a more natural counterpart to the reminder tone ("beep") that Ericsson and Simon proposed (but, to the best of our knowledge, never pursued).

What, finally, do the current results suggest for the practitioner? We have seen that the B&R protocol leads to a more natural interaction, where participant and test administrator are engaged in a dialogue, albeit an asymmetric one in which the participant does most of the talking. However, there is a downside to this, since the evidence from the current comparative study shows that subjects' performance (in terms of task success and lostness) is actually influenced by the way the thinking aloud protocol is administered, which —as we have argued— is an undesirable side-effect, especially when navigation is a measure of interest. Still, the E&S protocol brought to light essentially the same navigation problems. This suggests that it is perhaps safer to follow the Ericsson and Simon approach, while taking into account the work-arounds described above (such as asking for comments afterwards; avoid using data following system bugs or technical problems), certainly until a proper theoretical foundation and experimental validation of the B&R protocol has been completed. There is one complication, however, and that is the absence of proper descriptions on how to conduct thinking aloud usability tests in terms of the Ericsson and Simon framework (or any other framework for that matter). Descriptions of thinking aloud experiments in usability handbooks tend to be general and broad, and offer little guidance on the proper administration of such tests. Ironically, Boren and Ramey [12] probably contains the most explicit description to date on how to perform an Ericsson and Simon style thinking aloud test. Clearly, a more thorough theoretical underpinning of thinking aloud as a usability test is called for. Such a program could start from the Ericsson and Simon approach to thinking aloud, probably with some adaptations to account for usability practice. Boren and Ramey offer various useful suggestions in this direction, but comparative experiments of the sort described in this paper are needed to get a better understanding of what liberties (if any) a test administrator can take without causing validity and reliability problems.

Acknowledgements

Bibliography

[1]   A. Newell and H. Simon, *Human Problem Solving*, Englewood Cliffs, NJ: Prentice Hall, 1972.

[2]   L. Flower and J. Hayes, "A cognitive process theory of writing", *College composition and communication*, vol. 32, pp. 365-387, 1981.

[3]   J. Carroll, P. Smith-Kerker, J. Ford and S. Mazur-Rimetz, "The minimal manual", *Human Computer Interaction*, vol. 3, pp. 123-153, 1987.

[4]   K.A. Schriver, *Revising computer documentation for comprehension: ten exercises in method-aided revision*, Technical Report, nr. 14, Pittsburgh PA: Carnegie Mellon University, 1984.

[5]   K.A. Schriver, *Dynamics in Document Design: Creating Text for Readers*, New York: Wiley, 1997.

[6]   J. Nielsen, *Usability Engineering*, Cambridge MA: Academic Press, 1993.

[7]   K. Ericsson and H. Simon, "Verbal reports as data", *Psychological Review*, vol. 87, pp. 215-251, 1980.

[8]   J.E. Russo, E.J. Jonson and D.L. Stephens, "The validity of verbal methods", *Memory and Cognition*, vol. 17, no. 6, pp. 759-769, 1989.

[9]   P. Smagorinsky, "The reliability and validity of method analysis", *Written Communication*, vol. 6, pp. 463-479, 1989.

[10] M. Veenman, J. Elshout and M. Groen, "Thinking aloud: does it affect regulatory processes in learning?", *Tijdschrift voor Onderwijsresearch*, vol. 18, no. 6, pp. 322-330, 1993.

[11] N. Ummelen and R. Neutelings, "Measuring reading behaviour in policy documents: a comparison of two instruments", *IEEE Transactions on Professional communication*, vol. 43, no. 3, pp. 292-302, 2000.

[12] M. Boren and J. Ramey, "Thinking aloud: reconciling theory and practice", *IEEE Transactions on Professional Communication*, vol. 43, no. 3, pp. 261-278, 2000.

[13] R.E. Nisbett and T.D. Wilson, "Telling more than we can know: Verbal reports on mental processes", *Psychological Review*, vol. 84, pp. 231-259, 1977.

[14] K. Ericsson and H. Simon, *Protocol Analysis: Verbal Reports as Data*, Cambridge, MA: The MIT Press, 1993.

[15] J. Dumas and J. Redish, *A Practical Guide to Usability Testing*, Norwood NJ: Ablex, 1994.

[16] H. Clark and E. Schaefer, "Contributing to discourse", *Cognitive Science*, vol. 13, pp. 259-294, 1989.

[17] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese", *Journal of Pragmatics*, vol. 23, pp. 1177-1207, 2000.

[18] H. den Hartog Jager, "Dolen door des schrijvers geest" [in Dutch], *NRC Handelsblad*, 03/14/2000.

[19] H.A. Simon, "The functional equivalence of problem solving skills", *Cognitive Psychology*, vol. 7, pp. 268-288, 1975.

[20] K. Drummond and R. Hopper, "Some uses of yeah", *Language and Social Interaction*, vol. 26, no. 2, pp. 203-212, 1993.

[21] P.A. Smith, "Towards a practical measure of hypertext usability", *Interacting with computers*, pp. 365-381, 1996.

[22] M. Otter and H. Johnson, "Lost in hyperspace: metrics and mental models", *Interacting with Computers*, vol. 13, pp. 1-40, 2000.

[23] N.A. Stanton and M.S. Young, "What price ergonomics?" *Nature*, vol. 399, pp. 197-198, 1999.

[24] W.D. Gray and M.C.S. Salzman, "Damaged merchandise? A review of experiments that compare usability evaluation methods", *Human computer interaction*, vol. 13, pp. 203-261, 1998.

[25] M. de Jong and P. J. Schellens, "Toward a document evaluation methodology: what does research tell us about the validity and reliability of evaluation methods?", *IEEE Transactions on professional communication*, vol. 43, no. 3, pp. 242-260, 2000.

[26] J. Nielsen, "Estimating the number of subjects needed for a thinking aloud test," *International Journal of Human-Computer Studies*, vol. 41, pp. 385-397, 1994.