What Do Usability Test Moderators Say? 'Mm hm', 'Uh-huh', and Beyond

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

Kristina Bonde Kristoffersen
Peytz & Co
Copenhagen, Denmark
kristinabonde@gmail.com

ABSTRACT
Moderators in usability tests wrestle with the conflicting goals of obtaining relevant information from the users while at the same time avoiding to influence the users in ways that change how they use and feel about the tested system. In this study we investigate what moderators say by categorizing the moderator verbalizations from 12 test sessions. During the test tasks affirmations (38%) were the most common moderator verbalizations, followed by task instructions (32%) and prompts for reflection (16%). In addition, more of the moderator verbalizations during the tasks were closed (31%) than open (14%) and many more were positive (16%) than negative (1%). The moderators verbalized at a lower rate during the tasks than in the part of the sessions before the first task and after the last task. Still, they talked quite a lot. We discuss the content of their verbalizations and the implications of our findings.

Author Keywords
Usability evaluation methods; usability testing; test moderation; test instructions; verbal reports; thinking aloud.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces – Evaluation/methodology.

INTRODUCTION
Designs and design ideas are, at least at the outset, half-baked, incomplete, or otherwise partly flawed. Improving on these flaws is a main driver in innovation [26] and a main activity in design [15]. As a consequence good design presupposes effective methods for identifying flaws. With respect to usability and user experience a well-established method for this purpose is the usability test [e.g., 7, 27]. Usability testing yields feedback to designers about users' experience of ease, difficulty, satisfaction, frustration and the like when they encounter a design.

NordiCHI'18, September 29-October 3, 2018, Oslo, Norway

ACM ISBN 978-1-4503-6437-9/18/09…$15.00
https://doi.org/10.1145/3240167.3240181

In essence a usability test consists of a user who exercises a system while thinking out loud and an evaluator who observes the user and listens in on the user's thoughts. To tailor usability tests to different needs and situations numerous variants of the test have been devised, including some that skip thinking aloud [27] or the presence of an evaluator [3]. However, this study is about usability tests in which the user thinks aloud and an evaluator is present to moderate the test sessions. We focus on the evaluator's role as test moderator and ask: *What do usability test moderators say*? This question is important because test moderation wrestles with conflicting goals [2, 17]. On the one hand, the moderator's prompts serve to obtain information from the users about the reasons for their observable behavior and about their experience from using the system. To pursue this goal the moderator asks the users to reflect on their use of the system. The reflections will help reveal difficulties, misunderstandings, and unmet expectations and, thereby, contribute to pinpointing the flaws in the design. On the other hand, the moderator's prompts may bias the users and create a test situation in which their behavior and experience misrepresent how they would use and feel about the system in a real-life situation. Avoiding such bias and reactivity is also an important goal in test moderation. Ericsson and Simon [10] contend that reflections cause reactivity and should be avoided during thinking aloud. The users should, instead, be instructed to report merely the information they attend to in using the system. This instruction clashes with obtaining rich usability insights from the users.

In the following we review related work on how usability tests are influenced by the verbalizations made by moderators and users during the tests. Notably, this work has analyzed user verbalizations but merely made recommendations about moderator verbalizations. Then we account for the method of our empirical work and present our results. In short, we transcribe the moderator verbalizations from twelve usability test sessions and categorize these verbalizations according to their topic, directedness, tense, and valence. Finally, we discuss the content and implications of moderator verbalizations and the limitations of this study.

RELATED WORK
The purpose of the moderator's verbalizations is, partly, to instruct the users about what to do and, partly, to elicit information from them about their thoughts and activities.

We are primarily concerned with the second of these purposes – supporting the users in thinking aloud.

Classic and Relaxed Thinking Aloud

Thinking aloud ties into discussions about whether and, if so, how users may give a concurrent verbal report of their activities without at the same time changing these activities. Ericsson and Simon's [9, 10] work on verbal reporting provides the primary theoretical framework for informing these discussions. This framework distinguishes between user verbalizations at three levels:

- *Level 1* is the verbalization of information that is in the user's present focus of attention in verbal form. This information can be reported as is; no intermediate processes are necessary to prepare it for reporting.
- *Level 2* is the verbalization of information that is in the user's present focus of attention but in nonverbal form. To report this information it must be recoded into verbal form. The recoding involves additional processing of the information but does not bring new information into the user's focus of attention.
- *Level 3* is the verbalization of information that must first be brought into the user's focus of attention. Reporting this information influences the user's focus of attention by changing it in ways beyond those occasioned by task performance.

Ericsson and Simon [10] contend that verbal reports at Levels 1 and 2 do not change the users' activities because the information to which the users attend is unchanged. Consequently, they recommend restricting thinking aloud to these two levels of verbalization. This recommendation means, however, that users cannot be asked for explanations of their behavior, expressions of their expectations, and reflections on the user experience. Such verbalizations are precluded because they are at Level 3: They change the user's focus of attention from the information involved in performing the test task to the information needed to explain this performance, express expectations, and reflect on the experience. By changing the information to which users attend, verbalizations at Level 3 distort the thought process and change user behavior. There is plenty of evidence of such changes [11, 18, 29]. However, it is also well-recognized that explanations, expectations, and reflections are valuable inputs for evaluating the user experience [2, 23].

In practical usability testing Ericsson and Simon's recommendation is often relaxed [2, 24]. This relaxed thinking aloud includes verbalizations at all three levels of verbalization. The users are requested to provide a running commentary of their actions and, if they fall silent, they are prompted for their current thoughts and for reflections on their actions. In contrast, classic thinking aloud consists of instructing users to restrict their verbalizations to Levels 1 and 2, thereby complying with Ericsson and Simon's [10] recommendation. Most studies find that classic thinking aloud supplies an accurate record of the users' thought process without altering their task performance in any other way than by prolonging it [10, 11]. However, a few studies find that this variant of thinking aloud also distorts the thought process and changes user behavior, for example by influencing the users' perception of time [19] and by deteriorating their performance on spatial tasks [12].

Recommendations for Moderator Verbalizations

It is a standard phrase in classic thinking aloud to instruct the user to "act as if you are alone in the room speaking to yourself" [10, p. 376]. The rationale for this instruction is that the exclusive focus of classic thinking aloud on task information is easier to maintain if the users do not think of themselves as involved in a conversation with the moderator about their task performance. In this approach moderators should remain silent unless the user has not been verbalizing for some time and they should restrict their verbalizations to a neutral 'keep talking'.

Boren and Ramey [2] criticize this approach for its assumption that verbal processes can be divorced from communicative purposes. Instead, they propose to acknowledge that the user and moderator are communicating and, then, utilize insights from speech communication theory to create a highly asymmetrical speaker/listener relationship. Speech communication theory acknowledges the speaker as well as the listener: "talk is not simply a form of action" performed by the speaker alone "but a mode of interaction" between speaker and listener [14, p. 205]. The roles of speaker and listener are continuously negotiated and change dynamically when the communicating parties take turns at speaking and listening. While the speaker/listener relationship may be asymmetrical, speech communication theory asserts that the listener must necessarily respond [6]. A type of responses of specific relevance to usability-test moderators are acknowledgement and continuer tokens, such as 'okay', 'mm hm', and 'uh-huh'. Listeners utter these tokens at the end of a speaker's conversational units to indicate that they follow what the speaker has said so far and wish for the speaker to continue with the next unit [14]. That is, continuer tokens indicate engaged listening while at the same time indicating that the listener foregoes the opportunity to assume speakership. Boren and Ramey [2] propose that by using these tokens the moderator can fulfil his or her communication role without making users overly conscious that they do nearly all the talking.

Textbooks on usability testing acknowledge the need to avoid influencing the user's task performance because that would jeopardize the validity of the test [7, 8, 27]. However, they also recommend moderator verbalizations that go beyond classic thinking aloud. For example, Dumas and Redish [8, p. 281] propose that users are prompted to think out loud by saying something like: "John, could you tell us why you pressed the enter key?" This prompt explicitly invites explanation and asks John to revisit

information that is no longer in his present focus of attention. Rubin and Chisnell [27, p. 208] propose asking: "Exactly how did that differ from what you expected to happen?" They consider this question neutral because it does not imply right or wrong answers. However, to answer the question the user must shift her focus of attention to her expectations and do so to the extent to being able to express them 'exactly'. Dumas and Loring [7, p. 75] find it useful to replace questions with probes such as: "I noticed that you paused before clicking [name of a UI object]. Share with me what you were thinking at that point." Their rationale for avoiding questions is that questions often put users in a defensive position and make them feel challenged. However, the probe invites reflection by requiring that the user's current thoughts and presently attended information are replaced with earlier thoughts and previously attended information. Finally, some textbook authors refer to usability tests as interviews [13], thereby emphasizing the communicative element and deemphasizing an alertness to relaxed thinking aloud as a possible source of reactivity. In all these examples prompts that will foster relaxed thinking aloud are put forward as good practice.

The Effect of Verbalizations on Test Outcomes

Krahmer and Ummelen [21] compared classic thinking aloud with thinking aloud following Boren and Ramey's [2] speech-communication protocol and found that the users in the latter condition were less lost and solved more tasks correctly. The reason for these indications of reactivity could be that 12% of the moderator verbalizations in the speech-communication protocol were "suggestions to direct the subject back on the right track" [21, p. 114]. Such suggestions were not offered during classic thinking aloud. Several other studies [1, 4, 25] have also compared classic thinking aloud with the speech-communication protocol; none of these studies found differences in task completion rate, task completion time, or user satisfaction. Probably, thinking aloud following the speech-communication protocol works quite similarly to classic thinking aloud.

Multiple studies have compared classic and relaxed thinking aloud with a control condition in which the users performed without thinking aloud. For example, Hertzum et al. [18] found that relaxed thinking aloud resulted in longer task completion times, more commands to navigate the website, and higher mental workload than the control condition; classic thinking aloud only prolonged task completion times. Olmsted-Hawala et al. [25] found that relaxed thinking aloud yielded higher satisfaction with the website than classic thinking aloud and higher task completion rates than both classic thinking aloud and the control condition. Bruun and Stage [4] found that classic and relaxed thinking aloud revealed more usability problems than the control condition; there was no difference between the two thinking-aloud conditions. These authors note that thinking aloud "led to data from which we could enrich problem descriptions to include notions of why a particular problem was observed. The opposite was the case for the Silent [i.e., control] condition where we could observe a problem but not interpret why it occurred" [4, p. 171]. Zhao et al. [31] found that relaxed thinking aloud led to the identification of more dialog, functionality, layout, and navigation problems than classic thinking aloud but that the problems unique to relaxed thinking aloud tended to be of low severity. Relatedly, Hertzum et al. [17] categorized user verbalizations from relaxed thinking aloud with respect to their relevance to problem identification. The result of this categorization was 53% (low), 27% (medium), 11% (high), and 9% (other).

Several studies have found that relaxed thinking aloud yielded more user verbalizations of expectations, experiences, and explanations than classic thinking aloud but also that these verbalization categories occurred with some frequency during classic thinking aloud [30, 31]. In Zhao et al. [31] as much as 30% of the verbalizations in classic thinking aloud involved expectations, experiences, or explanations. Formally, these verbalization categories should be specific to relaxed thinking aloud. Their frequent occurrence is surprising and indicates a need for investigating what moderators say to support users in thinking aloud. While several of the studies mentioned above make recommendations about moderator verbalizations, Krahmer and Ummelen [21] is, to our knowledge, the only study that touches upon what moderators actually say during usability tests. Their analysis of moderator verbalizations was restricted to thinking aloud following the speech-communication protocol; they found 72% acknowledgement tokens, 12% suggestions, 11% prompts for clarification, and 6% other moderator verbalizations.

METHOD

To investigate moderator verbalizations, we analyzed twelve test sessions. The sessions were from two usability tests, which differed somewhat in their level of formality but both employed relaxed thinking aloud. We acknowledge that the users' behavior and verbalizations in these usability tests have previously been used in comparisons of moderated and unmoderated usability tests [17, 20]. The data have not previously been analyzed with respect to the two moderators' verbalizations.

Usability Test 1: Music News Site

The first usability test involved gaffa.dk, a Danish news site for music. The main content of the site was news, reviews, feature articles, and an archive of Danish music history since 1983. In addition, the site provided playlists, videos, and photos of contemporary music events. Seven users were recruited for this test, which was conducted by a usability professional from a Danish company specializing in usability testing. The moderator received the test tasks for the test but remained unaware of the specific focus of our study.

Each test session involved a single user. Before they were invited for the test session, the users were screened for their

ability to think aloud; all users passed this screening. At the test session the moderator welcomed the users, explained what was going to happen, and instructed them to think aloud. This instruction consisted of the following statement: 'While you solve the tasks, you are to think aloud. That is, you are to explain what you are in doubt about on the website, what you like, what you dislike, and so forth.' The users received five test tasks, which represented common uses of the website. While the first task was open-ended and partly aimed at ensuring an unintimidating atmosphere, the four other tasks were goal-directed. For example, the second task read 'When does [a named artist] give a concert in Aarhus in the Scandinavian Congress Center?' and was thus about finding a specified fact. In contrast, the last task was a comparison-of-judgement task: 'Find two interesting articles on the Gaffa website and explain which one you like the more, and why.' The tasks were read aloud by the moderator, who also prompted the users for information while they were solving the tasks. After the last task the moderator asked the users a few follow-up questions about their overall experience of the website.

The moderator remained in the room with the users throughout the sessions, which were video recorded. On the basis of the video recordings the verbal content of the sessions was subsequently transcribed for analysis.

Usability Test 2: Truck Rental Site

The second usability test involved uhaul.com, a US website for renting moving trucks. The trucks can be picked up and dropped off at various locations across the US. In addition to truck rental, the users of U-Haul can rent self-storage units, moving boxes, dollies, and the like. Five users were recruited for this test, which was conducted by a usability professional from a US company. The moderator received the test tasks for the test but remained unaware of the specific focus of our study.

Each test session involved a single user. The moderator welcomed the users, explained what was going to happen, and instructed them to think aloud by, for example, saying 'If you can try to think out loud so I know what you are looking for and what your impressions are of the site.' On the basis of a scenario about helping two of their friends to move, the users received seven test tasks, which represented common uses of the website. For example, the first task read 'The couple needs a truck that is suitable for all the furniture and belongings in their 3 room apartment. Please find the total price the couple will have to pay for the truck. Note: They are moving on April 14th from Darlington Rd. in Pittsburgh, PA 15217 to Emerson St. in Denver, CO 80218.' And the fourth task read 'You have a few questions that the U-Haul website hasn't answered. Please find the phone number for the U-Haul pickup location closest to the couple's home on Darlington Rd. in Pittsburgh, PA.' The users had the task descriptions available in writing and were prompted by the moderator while solving the tasks.

The moderator left the room after instructing the user, who was thus alone in the room while solving the tasks. From the adjoining room the moderator could see the user but not vice versa. Communication was via an audio link between the two rooms. After the last task the moderator rejoined the users and asked them a few follow-up questions about their overall experience of the website. The sessions were video recorded and subsequently their verbal content was transcribed for analysis.

Data Analysis

The transcripts reproduced the moderator verbalizations verbatim and divided them into segments. A segment was defined as the stretch of speech from the moderator started speaking to the user started speaking (this definition corresponded to a conversational turn in speech-communication analysis). We refer to a segment as a verbalization. The 12 test sessions comprised 1073 moderator verbalizations. We analyzed these verbalizations by categorizing them according to four classifications: topic, directedness, tense, and valence (see Table 1).

The *topic* classification was devised on the basis of textbook recommendations about how to elicit relaxed and classic thinking aloud [2, 7, 8, 27], supplemented with reading the verbalizations from two test sessions. The *directedness* classification distinguished between open and closed verbalizations. We included this classification because textbooks on usability testing [e.g., 8] recommend the use of open verbalizations to avoid leading or biasing the users. The *tense* classification concerned whether the moderator asked the users about their past, present, or future actions. We included this classification because verbalizations in the past and future tenses clearly invite the user to reflect, that is to relaxed thinking aloud, while verbalizations in the present tense may be open to classic thinking aloud. The final classification concerned the *valence* of the verbalizations and served to distinguish between positive, negative, and neutral verbalizations. We included this classification to be able to analyze how the moderators balanced encouragement and a good atmosphere against detachment and strict compliance with the test tasks.

For each of the four classifications, the categorization of the verbalizations involved four steps. First, a training set of 127 verbalizations was categorized by the two authors independently. The training set consisted of a randomly selected session from the first usability test and a randomly selected session from the second usability test. Each verbalization was assigned either to one of the classification categories or to an 'other' category. Second, all disagreements in the authors' categorizations of the training set were discussed to reach consensus about the categorization of the verbalizations and to create a shared understanding of the classifications. Third, the remaining 946 verbalizations (10 test sessions) were categorized by the two authors independently. Fourth, all disagreements in

Topic classification
Prompt for description: Verbalizations asking the users to say out loud what they are doing and what is happening
Prompt for reflection: Verbalizations asking the users about their assessments, experiences, and explanations, including the reasons for their actions
Affirmation: Verbalizations signaling that the moderator is attending to the user's verbalizations, such as 'Mm hm', 'Okay', and 'Uh-huh'
Task instruction: Verbalizations consisting of reading the task, or part of it, out loud to announce it to the user or remind the user of an aspect of the task
Assistance: Verbalizations providing the user with solicited answers and unsolicited guidance
Directedness classification
Open: Verbalizations that are non-directed, neutral, and open to many issues
Closed: Verbalizations that are directed toward one issue at the expense of others
Tense classification
Past: Verbalizations concerning the user's earlier behavior and experience
Present: Verbalizations concerning the user's current behavior and experience
Future: Verbalizations concerning the user's upcoming behavior and experience
Valence classification
Positive: Verbalizations conveying approval, confirmation, support, and other positive emotions
Negative: Verbalizations conveying disagreement, disapproval, reservation, and other negative emotions
Neutral: Verbalizations conveying neither positive nor negative emotions

Table 1. The four classifications used in categorizing the moderator verbalizations.

the authors' categorizations of these verbalizations were discussed and a consensus was reached.

For the 946 non-training verbalizations Cohen's [5] kappa of the inter-author agreement was .72, .65, .69, and .62 for the topic, directedness, tense, and valence classifications, respectively. These values met the criterion that kappa values above .60 indicate satisfactory reliability [22].

RESULTS

The seven test sessions of the music news site were an average of 25.8 (*SD* = 6.1) minutes long and the five test sessions of the truck rental site an average of 37.3 (*SD* = 8.3) minutes. In the following, we first analyze the verbalizations with respect to their topic, directedness, tense, and valence. For each classification we start by analyzing the part of the sessions during which the users solved the test tasks and then compare with the part of the sessions before the first task and after the last task. Figure 1 summarizes this analysis. After that, we investigate differences between the two tests.

Topic

During the tasks the moderator verbalizations were unevenly distributed across the topic categories (Figure 1, top left). The largest category was *affirmations* (38%, i.e. 258 of 677 verbalizations), which signaled that the moderator attended to the user's verbalizations without assuming speakership. These verbalizations were short, an average of 1.4 words (*SD* = 1.0). Typical affirmations were 'Mm hm', 'Okay', 'Okay, great', 'Yeah', and 'Yes'.

The second largest topic category during the tasks was *task instructions* (32%), which consisted of reading the next task out loud to the user (music news site), informing the user that a slide with the next task was now presented on the screen (truck rental site), repeating part of the task instruction on request, and at the end of the task asking the user to rate the task. For example, the moderator asked 'Okay. So now that you have finished that task how would you rate it on that same scale of very difficult to very easy?' With an average length of 14.3 words (*SD* = 11.2) task instructions were the longest of the topic categories.

Prompts for reflection were the third largest topic category (16%) during the tasks and almost as long as task instructions (*M* = 13.0 words, *SD* = 7.8). This category included a variety of prompts. The moderator for example:

- Asked the users about their impression of the site (e.g., 'Mm hm. Was that… Do you think that was easy or difficult to find?').
- Asked the users about their expectations (e.g., 'So what would you expect there? Where would you expect to see the cost?').
- Asked the users to elaborate verbalizations they made earlier (e.g., 'So is that what you thought when you mentioned they would want to drop it off where they are moving to?').
- Asked the users whether they noticed specific pieces of information (e.g., 'Okay. And did you see any options for choosing which insurance when you were booking the truck?').
- Asked the users to take specific actions to clarify something for the moderator (e.g., 'Why don't you take a look at the shopping cart and show me where you would expect it to show up?').
- Asked the users whether they interpreted the situation in one way or another (e.g., 'So what would you be looking for at this point? Do you think that even though there is no insurance they still might not have to pay all of it, or is something automatically included and you said no to extra insurance?').

- Asked the users to imagine a hypothetical situation (e.g., 'But if you had to find one, would it then be easy?').
- Asked the users to extrapolate from the test to their everyday lives (e.g., 'I don't know how much you use the website but was it something you would make use of if you were?').

Prompts for description (4%) were much rarer than prompts for reflection. They were also shorter (M = 8.8 words, SD = 6.0). During the tasks the moderator would for example prompt for description by asking 'What are you thinking?', 'What are you looking for?', or simply 'Such as?' On occasion, the prompts for description got more elaborate and bordered on prompts for reflection: 'Okay. That is good to know. And at this point, what's been added from those storage units?' This prompt could be interpreted literally and answered descriptively or it could be taken as an opportunity to reflect on whether the user's interaction with the truck rental site had produced the intended result.

The users mostly solved the tasks without help but from time to time the moderator provided *assistance* (5%). Assistance was as often about helping the users correctly interpret the task instructions as it was about helping them navigate the websites. In the test of the truck rental site the moderator would for example assist a user in estimating the size of truck needed by spelling out the task instruction in additional detail: 'And just to clarify, it's a one-bedroom but three-room apartment. So it's living room, kitchen, bedroom, and then a bathroom as well.' On a few occasions users made an error and continued for some time without realizing it. When it became obvious that they had not realized the error the moderator asked them to discontinue their current activity, return to a point at which the error was visible, and verify whether everything was as intended.

In the part of the test sessions before the first task and after the last task the moderator verbalizations were distributed differently onto the topic categories than during the tasks (Figure 1, top right vs top left). Task instructions were the most frequent category (54%) before and after the tasks, and prompts for description and reflection were rare. We note that affirmations (28%) were also frequent during this part of the sessions. For each topic category the average moderator verbalization was longer before and after the tasks than during the tasks. The average verbalization length increased from 8.1 words (SD = 9.3) during the tasks to 14.9 words (SD = 26.1) before and after the tasks. This increase was significant, $F(1, 1071) = 37.61$, $p < .001$.

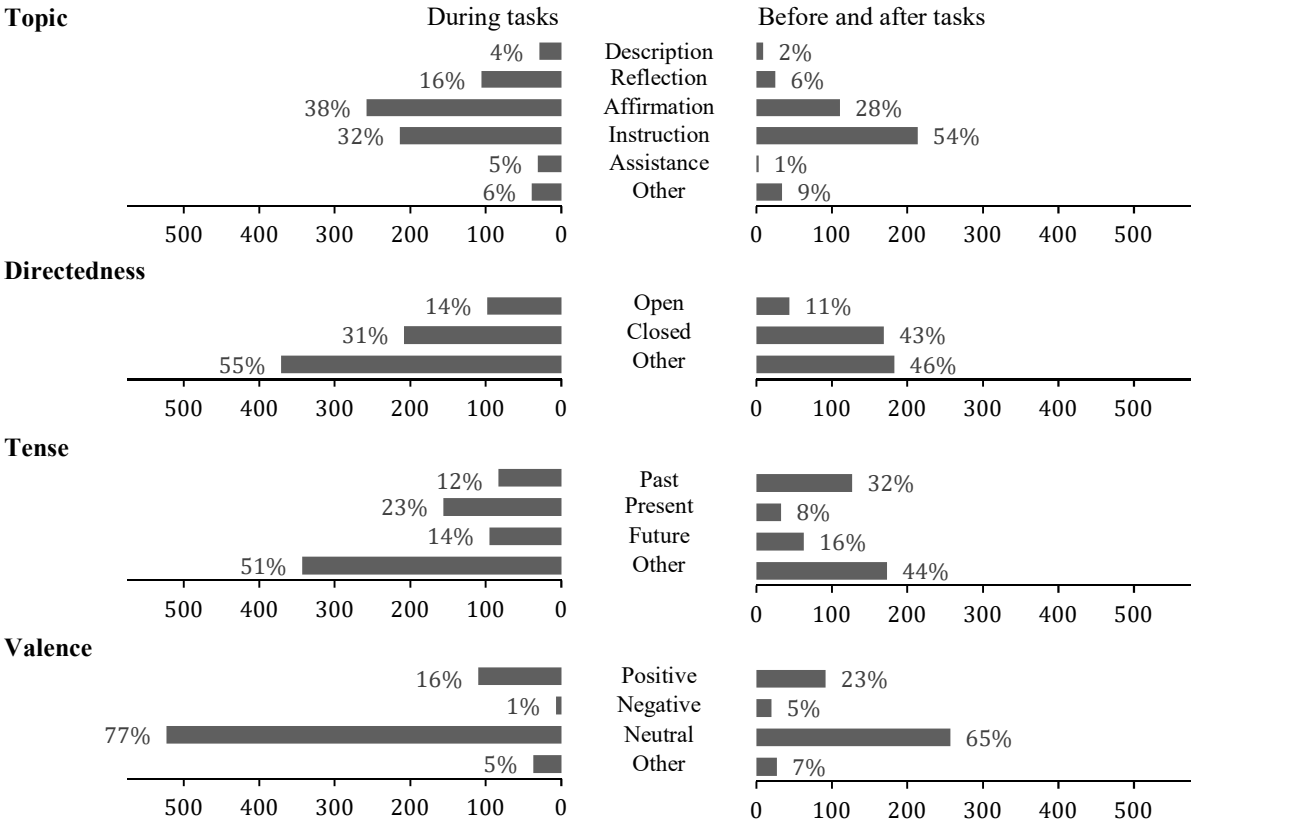

Figure 1. The moderator verbalizations from the 12 test sessions distributed onto the categories of the four classifications, N = 1073 verbalizations. The graphs on the left show the 677 verbalizations made while the users were solving the test tasks. The graphs of the right show the 396 verbalizations made before the users solved the first task and after they solved the last task. The percentages give the distribution across the 677 and 396 verbalizations, respectively.

Directedness

During the tasks the moderator made about half as many open (14%) as closed (31%) verbalizations. The open verbalizations were questions such as 'Okay, so what would you do at this point?' and 'Okay. Have you any… Do you want to comment on this?' These questions did not lead or bias the users but merely invited them to share their thoughts. In contrast, the closed verbalizations included questions such as 'Could you find it?', 'Do you think that is a relevant function?', 'Can you just point with your mouse where you saw the total?', and 'Would you normally continue or is it the point that you would stop at?' These questions suggested a fixed set of response options, often just a binary yes/no. That is, the moderator set the focus and largely reduced the subsequent user verbalization to a confirmation or disconfirmation of the moderator's depiction of the situation. Almost one in three moderator verbalizations were of this closed type, including 26% of the prompts for reflection. In addition to the open and closed verbalizations as much as 55% of the verbalizations made during the tasks were in the 'other' category. The verbalizations in the 'other' category were not questions, for example 70% of them were affirmations (none of which were open or closed).

In the part of the test sessions before the first task and after the last task the proportion of closed verbalizations was even larger than during the tasks (43% vs 31%). For example, the users of the music news site were asked to indicate on a five-point rating scale the extent to which they agreed that it had been frustrating to use the site. The proportion of open verbalizations before and after the tasks was roughly the same as during the tasks (11% vs 14%). However, the open verbalizations were about three times longer before and after the tasks (M = 44.4 words, SD = 57.1) than during the tasks (M = 15.8 words, SD = 12.4). This difference was significant, $F(1, 140) = 22.38$, $p < .001$. The closed verbalizations were also significantly longer before and after the tasks (M = 17.3 words, SD = 17.6) than during them (M = 14.1 words, SD = 8.7), $F(1, 375) = 5.19$, $p < .023$, but the difference was only about three words.

Tense

We classified the moderator verbalizations with respect to tense because asking the users about their past and future experience of the sites required them to attend to information other than the information involved in performing the current stage of the test tasks. That is, these verbalizations invited relaxed thinking aloud. Collectively, verbalizations asking the users about their past (12%) and future (14%) experience of the sites were slightly more frequent during the tasks than verbalizations asking them about their present actions and experience (23%). Verbalizations about the past included 'Which one of them was the more interesting? Or was one of them more appealing to you than the other?' This verbalization not only asked the user to look back but also to compare items attended in the past. The verbalizations about the future included 'Okay. So here is our next task and again I'll just ask you to read it out loud and then give me your expectations of how difficult or easy it will be.' This verbalization explicitly asked the user about her expectations toward a task she had not yet experienced. In contrast, the verbalizations in the present tense for example included 'Okay. And here is our next task', 'Right now you just want to find the price of the truck', and 'Why are you laughing?' While the first of these examples simply instructed the user and the second provided assistance, the last exemplifies that the questions about the present also included prompts for reflection. As it were, the present tense was not a useful indicator of verbalizations that avoided inviting relaxed thinking aloud. The prompts for reflection contained almost as many verbalizations in the present tense as the prompts for description (44% vs 52%).

Before and after the tasks the proportion of verbalizations in the present tense was lower (8% vs 23%) and the proportion of verbalizations in the past tense higher (32% vs 14%) than during the tasks. The increased focus on the past reflected that before the tasks the users were asked about their pre-test knowledge of the site and after the tasks they were asked to reflect back on their experience while solving the tasks. While the verbalizations asking the users about their past and present actions and experiences were a couple of words longer before and after the tasks than during them, the length of the verbalizations in the future tense increased from an average of 21.9 words (SD = 12.0) during the tasks to 45.2 words (SD = 49.5) before and after the tasks. This increase was significant, $F(1, 156) = 19.38$, $p < .001$, and to a considerable extent caused by long task instructions explaining what was going to happen during the test. Throughout the test sessions the 'other' category was dominated by affirmations. Thus, the lower proportion of 'other' verbalizations before and after the tasks as compared to during them (44% vs 51%) reflected the decrease in the number of affirmations.

Valence

The majority of the moderator verbalizations were neutral, that is, neither positively nor negatively valenced. During the tasks 77% of the verbalizations were neutral. These verbalizations for example included affirmations such as 'Mm hm', 'Okay', and 'Yeah'. They also included verbalizations that balanced positive and negative elements, such as 'And then explain what you like and what you do not like'. Before and after the tasks 65% of the verbalizations were neutral; these verbalizations were mainly affirmations and task instructions. However, the non-neutral verbalizations were strongly dominated by positively valenced verbalizations. During the tasks 16% of the verbalizations were positive and merely 1% negative. The positive moderator verbalizations made during the tasks included 'Correct', 'Good', 'Mm hm, okay. That was very well spotted', 'Okay, great', 'Okay, great, thank you. And how confident are you that you found the right answer?', 'Okay. That's good to know', and 'Yes. Wauw.

Do you have any comments about finding it?' As much as 59% of these positive verbalizations were affirmations. The positive verbalizations served to encourage the users by letting them know that they were doing well and providing valuable feedback. In contrast, the negative verbalizations aimed to keep the users on track by repeating task instructions ('No. Just the newest') and to have the users verbalize their dissatisfactions ('And what do you not like about it, including the format and such?'). Before and after the tasks the proportion of negative verbalizations was higher (5%) than during the tasks, but it remained much lower than the 23% positive verbalizations.

Differences Between the Two Tests

Table 2 contrasts the two moderators' verbalizations. The part of the sessions during which the users solved the test tasks was significantly shorter in the test of the music news site than in the test of the truck rental site, $F(1, 10) = 7.63$, $p = .020$. The reason for this difference was that the users of the music news site solved 5 tasks whereas the users of the truck rental site solved 7 tasks. Because the sessions in the two tests were not equally long, Table 2 gives verbalization measures that are independent of session length. The moderator in the test of the music site made significantly more verbalizations per minute than the moderator in the test of the truck rental site, $F(1, 10) = 14.08$, $p = .004$. Relatedly, the moderator in the test of the music news site also spoke significantly more words per minute, $F(1, 10) = 14.15$, $p = .004$. These differences confirmed that though both tests employed relaxed thinking aloud, the test of the truck rental site was somewhat more formal. We found no difference between the two tests in the number of words per verbalization, $F(1, 10) = 3.65$, $p = .085$.

In spite of the differences in the rate of verbalization the distribution of the verbalizations across the categories of the topic classification was similar for the two tests, see Figure 2 (top). The main difference was a smaller proportion of affirmations and a larger proportion of task instructions during the test of the truck rental site.

	Music news site		Truck rental site	
	During tasks	Before/ after tasks	During tasks	Before/ after tasks
Time (in minutes)	20.7 (6.1)	5.1 (0.7)	31.5 (7.2)	5.8 (1.5)
Verbalizations per minute	3.07 (0.53)	7.17 (2.03)	1.65 (0.53)	4.93 (0.77)
Words per minute	22.56 (5.83)	101.00 (15.00)	15.43 (2.86)	80.05 (9.60)
Words per verbalization	7.29 (1.13)	14.59 (2.86)	9.79 (1.89)	16.42 (2.54)

Table 2. The moderator verbalizations in the test of the music news site (7 users) and the truck rental site (5 users). The table gives the mean and, in parenthesis, the standard deviation.

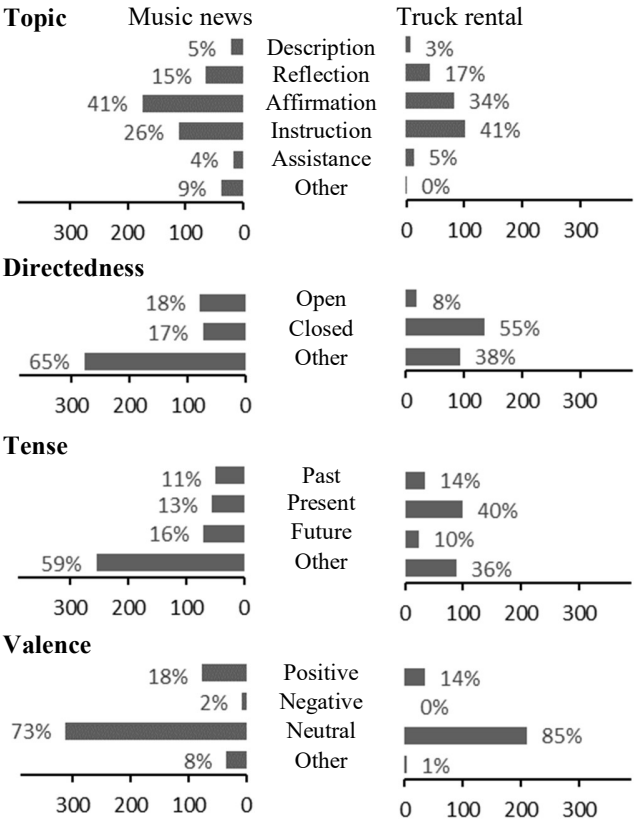

Figure 2. The number of moderator verbalizations during the tasks in the test of the music news site (left) and truck rental site (right). The percentages give the distribution across the 430 (music news) and 247 (truck rental) verbalizations.

In contrast, the verbalizations were distributed quite differently across the categories of the directedness classification. The proportion of closed questions was much larger during the test of the truck rental site (55%) than the test of the music news site (17%). A major contributor to this difference was that the moderator in the test of the truck rental site asked the users the same two questions at the beginning and end of each task. These questions had a rating scale for answering and were, thus, closed. Furthermore, the proportion of verbalizations in the 'other' category of the directedness classification was larger in the test of the music news site (65% vs 38%), partly because there were more affirmations and partly because more of the task instructions were neither open nor closed.

The distribution of the verbalizations across the categories of the tense classification was similar in the two tests, except for a larger proportion of verbalizations in the present tense and a correspondingly smaller proportion of 'other' verbalizations in the test of the truck rental site. The main reason for this shift was more task instructions in the present tense during the test of the truck rental site.

Finally, the distribution of the verbalizations across the categories of the valence classification was similar in the

two tests (Figure 2, bottom). Most verbalizations were neutral but among the non-neutral verbalizations there were many more positive than negative verbalizations.

DISCUSSION

We find that moderators in relaxed thinking aloud tests talk quite a lot. During the tasks the moderator in the test of the music news site made an average of 3.07 verbalizations per minute, corresponding to 23 words per minute. In the test of the truck rental site the numbers were 1.65 verbalizations per minute and 15 words per minute. Before and after the tasks the moderators' rate of verbalization was even higher. As a matter of comparison, Hertzum et al. [17] report that in their moderated test sessions the users spoke an average of 110 words per minute. That is, the moderators in the tests of the music news and truck rental sites spoke at a rate of one word for every 5-7 words spoken by users.

What Do Test Moderators Say?

The largest topic category during the tasks was affirmations ('Mm hm', 'Okay', 'Okay, great' etc.). Still, there was only about half as many affirmations as in Krahmer and Ummelen's [21] study of thinking aloud following the speech-communication protocol. The affirmations had neither directedness, nor tense. They complied with textbook recommendations [2, 7], except that 25% of them were positively valenced rather than neutral. The sizeable number of positive verbalizations (16% of all verbalizations during the tasks) served to create a good atmosphere and encourage the users. While an encouraging atmosphere is important to a productive test, the atmosphere must not become dependent on positive verbalizations. Usability tests are conducted in order to identify problems and possibilities for improvement. Thus, usability tests must be able to accommodate negatively valenced verbalizations, at least from the users. The two moderators in the studied test sessions made very few negative verbalizations.

When the moderators went beyond affirmations and prompted the user for information, they tended to prompt for reflection rather than description. These prompts asked the users for their expectations, impressions of the site, interpretation of the situation, thoughts about hypothetical situations and so forth. That is, the users were asked for information they were not presently attending to solve the test tasks. Alhadreti and Mayhew [1] find that under such relaxed thinking aloud conditions the users experience the moderator as more of a disturbance than during classic thinking aloud. It probably added to the disturbance that the prompts for reflection were substantially longer than the affirmations, which were unlikely to be experienced as a disturbance. The more the users are disturbed in solving the tasks, the more the test session turns into an interview. Such a turn shifts the users' attention from the interaction with the system to the interaction with the moderator, thereby reducing the number of comments caused by concrete experiences from solving the test tasks [16]. Instead we must expect an increase in the number of comments that are made in response to moderator prompts but dissociated from concrete use experiences.

Many of the moderator verbalizations were neither affirmations nor prompts, but task instructions. In the test of the truck rental site the users had the task descriptions available in writing but task instructions were still the largest topic category during the tasks. In the test of the music news site task instructions were the second largest topic category during the tasks. Before and after the tasks more than half of the verbalizations were task instructions. While we expected that task instructions would dominate before the tasks, we are surprised that they are so frequent during the tasks. With frequent instructions during the tasks the moderator and user are, on that basis alone, engaged in interactions that go beyond classic thinking aloud. We speculate that these instructions are necessary because the tasks are new to the users. In contrast, classic thinking aloud was originally devised to study how people, often experts, perform well-learned tasks [10]. Maybe, classic thinking aloud is difficult to transfer to first-time use because task instructions will often be necessary.

Moderators go beyond classic thinking aloud whenever they invite the users to shift their focus of attention away from the information that enters into performing the test tasks. Using the four classifications in the present study this happens when the moderator verbalizations are prompts for reflection, task instructions, assistance, closed, past tense, future tense, positive, or negative. On this basis we find that 65% of the 677 moderator verbalizations during the tasks went beyond classic thinking aloud (before and after the tasks it was 74%). While it is trivial that relaxed thinking aloud tests, such as those in the present study, do not comply with the requirements for classic thinking aloud, it is noteworthy that so many of the moderator verbalizations go beyond classic thinking aloud. It appears that moderators in relaxed thinking aloud tests do not concern themselves with the issue of test reactivity or that they are unaware of the extent to which their verbalizations invite it.

In our analysis we have distinguished verbalizations made during the tasks from verbalizations made before and after the tasks. The rationale for this distinction is that the users thought aloud concurrently with solving the tasks. We find even more task instructions, closed questions, past-tense verbalizations, and positive verbalizations before and after the tasks than during them. The verbalizations were also longer before and after the tasks than during them. If the users think aloud retrospectively rather than concurrently then many of the moderator verbalizations, at least the prompts and affirmations, will move to after the tasks. In comparing two approaches to retrospective thinking aloud Willis and McDonald [28] find that having users think aloud after each task leads to longer task completion times, more errors, and more clicks than postponing thinking aloud until after all tasks have been performed. However, thinking aloud after each task produced more user

verbalizations that helped explain the users' actions, expectations, and experiences. Thus, replacing concurrent with retrospective thinking aloud does not simply dissolve the trade-off between avoiding that thinking aloud influences task performance and ensuring that it produces user verbalizations of benefit to usability testing. This finding makes it even more important to understand moderator verbalizations in usability tests in which the users think aloud concurrently with solving the test tasks.

Implications

Our study has several implications for usability testing. First, the rate of moderator to user verbalizations is high, especially considering that any usability insights emerge from the user verbalizations. Most of the users had few problems thinking out loud so less prompting will not necessarily result in less thinking aloud.

Second, the affirmations could be improved simply by being more conscious about keeping them neutral. 'Mm hm', 'Okay', 'Uh-huh' and the like will go a long way. It appears that little is gained by extending these affirmations with words such as 'good', 'great', and 'wauw'.

Third, the balance between positive and negative moderator verbalizations is skewed. While it is important to maintain a constructive atmosphere during a test, this skewness may bias the users toward also being positive and imperceptibly suppressing some criticism of the tested system.

Fourth, almost one in three moderator verbalizations during the tasks are closed questions. Such verbalizations attract responses that confirm or disconfirm the moderator's depiction of the situation. Open questions would be less leading and invite more informative responses.

Fifth, rather than habitually prompting for reflection moderators may remain alert to the continuous trade-off between test reactivity and usability insights, thereby reserving prompts for issues they see value in knowing and cannot tell from the user's observable behavior.

Sixth, prompting the users for reflections that consist of imagining and extrapolating is virtually dissociated from their use of the system to solve the test tasks. It is more like an interview. Such interview-type prompting should probably be reserved for the period before or after the tasks.

Seventh, it must be assumed that the moderator verbalizations, especially the prompts for reflection, influence the users' actions and experience. It is not clear how, and even whether, usability evaluators can account for these influences in their analysis of the test sessions.

Limitations

The main limitation of this study is that we have only analyzed two moderators' verbalizations. In tests that employ relaxed thinking aloud the moderators have considerable freedom in their interaction with the users. Thus, their verbalizations may reflect their personal styles. We acknowledge the need for additional studies of moderator verbalizations, preferably across clearly defined variants of relaxed thinking aloud. In our study the differences in the content of the moderator verbalizations between the two tests were in large part due to more and different verbalizations about task instructions in the test of the truck rental site. Thus, task instructions and the verbalizations about them appear important in defining variants of relaxed thinking aloud. Another limitation of this study is that we have not linked the content of the moderator verbalizations to the usability problems identified as a result of the tests. The absence of direct links from moderator verbalizations, through user verbalizations and actions, to identified usability problems weakens the implications of our analysis for the practice of usability testing. We can, however, report that both tests identified a large number of usability issues. The test of the music news site identified an average of 14 usability issues for each of its seven users [17] and the test of the truck rental site a total of 134 usability issues across its five users [20].

CONCLUSION

Moderators in usability tests face a trade-off between obtaining usability insights and avoiding test reactivity. The moderators' verbalizations are central to how this trade-off plays out during the test sessions. In this study we have analyzed the moderator verbalizations in sessions that employ relaxed thinking aloud, which accepts some test reactivity to obtain usability insights. We acknowledge that this study has mainly focused on the risks of test reactivity.

We find that the moderators verbalize quite a lot. During the test tasks affirmations are the most common moderator verbalizations, followed by task instructions and prompts for reflection. The prompts for reflection include verbalizations in which the users are asked to imagine situations beyond the test tasks and to extrapolate from the test to their everyday lives, thereby introducing interview elements in the sessions. About one in three verbalizations are closed questions, which lead the users toward a fixed set of response options. In addition, there are many more positive than negative verbalizations, which contributes to a good atmosphere but might bias the users toward also being positive. That said, neutral affirmations are common and the moderators do verbalize at a lower rate during the tasks than before and after the tasks.

We would welcome more research into what usability test moderators say, especially research that tries to link moderator verbalizations to identified usability problems.

ACKNOWLEDGEMENTS

In the interest of full disclosure, we note that at the time of the usability test of gaffa.dk, the second author was an intern in the company that conducted the test. However, she was not involved in running the test sessions. With respect to the usability test of uhaul.com, we gratefully acknowledge that Rolf Molich coordinated the research study, CUE-9, in which this test was conducted. Special thanks are due to the two moderators.

REFERENCES

1. Obead Alhadreti and Pam Mayhew. 2017. To intervene or not to intervene: An investigation of three think-aloud protocols in usablity testing. *Journal of Usability Studies* 12, 3: 111-132.
2. Ted Boren and Judith Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication* 43, 3: 261-278.
3. Anders Bruun, Peter Gull, Lene Hofmeister, and Jan Stage. 2009. Let your users do the testing: A comparison of three remote asynchronous usability testing methods. In *Proceedings of the CHI 2009 Conference on Human Factors in Computing Systems*. ACM Press, 1619-1628.
4. Anders Bruun and Jan Stage. 2015. An empirical study of the effects of three think-aloud protocols on identification of usability problems. In *Proceedings of the INTERACT 2015 Conference on Human-Computer Interaction*. Springer, 159-176.
5. Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1: 37-46.
6. Kent Drummond and Robert Hopper. 1993. Back channels revisited: Acknowledgement tokens and speakership incipiency. *Research on Language and Social Interaction* 26, 2: 157-177.
7. Joseph S. Dumas and Beth Loring. 2008. *Moderating usability tests: Principles & practices for interacting*. Morgan Kaufmann.
8. Joseph S. Dumas and Janice C. Redish. 1999. *A practical guide to usability testing. Revised edition*. Intellect Books.
9. K. Anders Ericsson and Herbert A. Simon. 1980. Verbal reports as data. *Psychological Review* 87, 3: 215-251.
10. K. Anders Ericsson and Herbert A. Simon. 1993. *Protocol analysis: Verbal reports as data. Revised edition*. MIT Press.
11. Mark C. Fox, K. Anders Ericsson, and Ryan Best. 2011. Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin* 137, 2: 316-344.
12. K. J. Gilhooly, E. Fioratou, and N. Henretty. 2010. Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology* 101, 1: 81-93.
13. Elizabeth Goodman, Mike Kuniavsky, and Andrea Moed. 2012. *Observing the user experience: A practitioner's guide to user research. Second edition*. Morgan Kaufmann.
14. Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies* 9, 2-3: 205-217.
15. Rex Hartson and Pardha S. Pyla. 2012. *The UX book: Process and guidelines for ensuring a quality user experience*. Morgan Kaufmann.
16. Morten Hertzum. 2016. A usability test is not an interview. *ACM Interactions* 23, 2: 82-84.
17. Morten Hertzum, Pia Borlund, and Kristina B. Kristoffersen. 2015. What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction* 31, 9: 557-570.
18. Morten Hertzum, Kristin D. Hansen, and Hans H. K. Andersen. 2009. Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology* 28, 2: 165-181.
19. Morten Hertzum and Kristin D. Holmegaard. 2015. Thinking aloud influences perceived time. *Human Factors* 57, 1: 101-109.
20. Morten Hertzum, Rolf Molich, and Niels E. Jacobsen. 2014. What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology* 33, 2: 143-161.
21. Emiel Krahmer and Nicole Ummelen. 2004. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication* 47, 2: 105-117.
22. Jonathan Lazar, Jinjuan H. Feng, and Harry Hochheiser. 2010. *Research methods in human-computer interaction*. Wiley.
23. Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads? Reflections on the think-aloud technique. In *Proceedings of the NordiCHI2002 Conference on Human-Computer Interaction*. ACM Press, 101-110.
24. Mie Nørgaard and Kasper Hornbæk. 2006. What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the DIS2006 Conference on Designing Interactive Systems*. ACM Press, 209-218.
25. Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the CHI2010 Conference on Human Factors in Computing Systems*. ACM Press, 2381-2390.
26. Henry Petroski. 1992. *The evolution of useful things: How everyday artifacts - from forks and pins to paper clips and zippers - came to be as they are*. Vintage Books.

27. Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: How to plan, design, and conduct effective tests. Second edition*. Wiley.
28. Leanne M. Willis and Sharon McDonald. 2016. Retrospective protocols in usability testing: A comparison of post-session RTA versus post-task RTA reports. *Behaviour & Information Technology* 35, 8: 628-643.
29. Timothy D. Wilson and Jonathan W. Schooler. 1991. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology* 60, 2: 181-192.
30. Tingting Zhao and Sharon McDonald. 2010. Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. In *Proceedings of the NordiCHI2010 Conference on Human-Computer Interaction*. ACM Press, 581-590.
31. Tingting Zhao, Sharon McDonald, and Helen M. Edwards. 2014. The impact of two different think-aloud instructions in a usability test: A case of just following orders? *Behaviour & Information Technology* 33, 2: 163-183.